ENCYCLOPÆDIA
BRITANNICA



MACROPÆDIA

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



THE UNIVERSITY OF CHICAGO

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

The New Encyclopædia Britannica

in 30 Volumes

MACROPÆDIA
Volume 8

Knowledge in Depth

FOUNDED 1768
15TH EDITION



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

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

Printed in U.S.A.

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

Geraniales

The Geraniales, or geranium order, is an order of flowering plants comprising 20 families that include approximately 143 genera and about 4,000 species. Half of the families contain 50 species or less, but the two largest families (Malpighiaceae and Oxalidaceae) each have about 900 species. These two large families have predominantly tropical distributions, although the Oxalidaceae has some temperate-region representatives. The families Geraniaceae and Balsaminaceae each have over 500 species with worldwide distributions. The remaining families are mostly tropical, and two (Dirachmaceae and Lepidobotryaceae) are monotypic (*i.e.*, they contain only one genus and one species); the former is confined to the island of Socotra, where it appears to be very rare, and the latter to tropical Africa.

GENERAL FEATURES

Size range and diversity of habitat. With representatives in both the tropics and in temperate regions, the order shows considerable variation in life form, from small annual herbs—nonwoody plants with short life cycles—to trees (*e.g.*, *Lepidobotrys*) of the tropical rain forest that have leathery leaves furnished with the drip tip (a sharp point at the end of the leaf) characteristic of this habitat. The families Zygophyllaceae, Nitrariaceae, and Peganaceae are adapted to dry or saline habitats and contain numerous desert or subdesert species. Certain genera of the family Zygophyllaceae, such as *Zygophyllum* and *Tetradiclis*, show adaptations to the environment in their succulent—thick, fleshy, water-storing—vegetative parts; these are sometimes coupled with an annual habit and abundant seed production. Several *Zygophyllum* species are characteristic of the desert regions of the Near and Middle East and Central Asia, as well as Africa and Australia. *Tetradiclis*, which also grows in the Asian deserts, resembles some of the fleshy species of the family Chenopodiaceae (order Caryophyllales, *q.v.*) encountered in the same regions and can easily be mistaken for them. The genus *Sarcocaulon* (South Africa) is the only species of the family Geraniaceae adapted to desert life, although many species of *Erodium* and *Monsonia* occur in desert regions. The species of *Nitraria*, conversely, are highly salt-tolerant, deep-rooted perennials. Spines or thorns frequently develop in desert plants, presumably as a protection against grazing; two genera of the Zygophyllaceae (*Fagonia* and *Plectrocarpa*) exhibit these features markedly.

Desert adaptations

A succulent form is sometimes found in plants growing in damp ground or, particularly, in shallow water; these species commonly have translucent stems, often much branched and either sprawling or supported by the stronger waterside vegetation among which they grow. The families Limnanthaceae and Balsaminaceae provide good examples of such plants.

Certain plants in disturbed ground exhibit a feature useful in aiding distribution by vegetative means—the production of bulbils (small buds, or bulbs) on the roots, in clusters at the stem base, or in the leaf axils (angles between the leafstalks and the plant stem). Each, if broken off, is capable of growing into a new plant. This feature occurs in several species of *Oxalis*, certain of which (*e.g.*, *O. corymbosa* and *O. latifolia*), although tropical in origin, have now spread widely by this means to become troublesome weeds of cultivated areas even in temperate regions.

Perhaps one of the most striking and obvious anatomical features in the order is that of the very characteristic single-celled plant hairs (trichomes) of the family Malpighiaceae—so widespread in the family that botanists frequently refer to these two-armed hairs as malpighian hairs when describing their presence in other families. They are very diverse in form, sometimes almost without a stalk below the branching point, sometimes in a more or less elongated stalk, with the branch arms spreading or ascending, and from nearly equal to very unequal in length. Sunken, cushion-shaped, or sometimes stalked glands also occur in this family.

Economic importance. The economic importance of the order is not great, with the exception of flax (*Linum usitatissimum*), the source of a valuable fibre that has been used by man since prehistoric times. It is grown in most temperate regions, especially in central and western Europe, not only for its fibre, from which linen is made, but also for its seed, the source of linseed oil.

Timber sources

A few timber trees occur in the order, the most famous being guaiacum wood (or lignum vitae), yielded by *Guaiacum officinale* (family Zygophyllaceae). The wood—very hard, durable, and heavy in weight—is used in shipbuilding and for mallets. The heartwood of *G. sanctum* produces a resin used for making small objects that require weight, hardness, and strength. *Bulnesia arborea* is Maracaibo lignum vitae, and the timber is used as for the *Guaiacum* product. Some species of the family Malpighiaceae are locally employed as timber sources, particularly *Ctenolophon parvifolium*, the durable wood of which is employed for house construction in Malaysia. The timber of *Klainedoxa gabonensis* is tough and hard wearing. *Houmiria balsamifera* is the source of a good timber, easily polished and used for furniture, house framing, and general carpentry. *Sacoglottis gabonensis*, from western Africa, produces an easy-to-work white wood.

Three species of *Balanites* are used locally. *B. aegyptiaca* seeds yield Betu oil, employed in soap manufacture. The wood is hard and used in Africa for clubs, plows, sticks, turnery, and general carpentry. *B. maughamii* yields a clear oil similar to olive oil, and *B. orbicularis* is the source of hanjigoad, a gum resin. *Irvingia gabonensis* (family Ixonanthaceae) is the wild mango (not to be confused with the true mango, *Mangifera indica*), the fruit of which is edible; the fat from the seeds yields dika bread and dika butter, staple foods in West Africa. The tough, hard wood has been used for street paving and in houses. The seeds of *I. oliveri* are the source of cay cay fat, used in candle manufacture.

Bulnesia sarmienti (family Zygophyllaceae), the source of guaiac wood oil, which has a roselike scent, is used in soaps and perfumery. Geranium oil for perfumery is obtained from *Pelargonium radula*, *P. odoratissimum*, *P. capitatum*, and other species of this genus.

A few plants of the order have been used medicinally. The root of *Hugonia mystax*, for example, has been used locally in India as a snakebite remedy and to reduce fever. *Linum catharticum* has laxative properties similar to those of senna (*Cassia*). *Geranium maculatum* is an American species known as alumroot from the styptic and astringent properties—ability to control bleeding and draw together soft tissues—of a liquid extract from the roots. The only species of widespread true medical significance, however, is the coca, *Erythroxylum coca*. From this species the local anesthetic cocaine is prepared. The leaves of *E. coca*, the coca tree, act as a nerve stimulant, and tribal peoples of South America

Use of coca leaves

chew them so that they can perform feats of endurance without fatigue. The components of the leaves act on the gastric nerves, eliminating the sense of hunger and allowing the user to go for a long period without food. As with most stimulants, however, the use of coca is followed by depression. *Banisteriopsis caapi* (family Malpighiaceae) contains the hallucinogenic alkaloid harmine.

The fruits of some *Malpighia* species (*e.g.*, *M. punicifolia*, the Barbados, or West Indian, cherry) are edible either fresh, as preserves, or as flavouring for jellies. The gooseberry-like fruits of *Averrhoa carambola* (family Oxalidaceae) are also edible, and those of *Nitraria* species are sometimes eaten. *Klainedoxa gabonensis* and *Desbordesia glaucescens* have edible seeds; those of the former are eaten fresh or roasted, and those of the latter are used in sauces. The outer layers of the fruit are edible in several *Houmiria*, *Sacoglottis*, and *Vantanea* species; both fruit and seeds contain a fatty oil.

The families Geraniaceae and Tropaeolaceae produce many well-known decorative plants. The Geraniaceae contains not only handsome species of *Geranium* itself but also the cultivated varieties and hybrids (crosses between different species) of the genus *Pelargonium*, to which the well-known pot and bedding geraniums of gardeners belong. The garden nasturtium is *Tropaeolum majus*, and some others of this genus also appear in gardens—notably *T. peregrinum*, the canary creeper. Species of the family Tropaeolaceae contain a hot mustard oil similar to that found in certain members of the mustard family (Cruciferae), a fact that no doubt explains the name "nasturtium" as applied to the garden *Tropaeolum majus*—botanically, the genus *Nasturtium* is the watercress. Some herbaceous *Oxalis* species have been grown as ornamentals, as have one or two trees of the genus *Averrhoa*, also of the Oxalidaceae. The nuisance value of several weedy species of *Oxalis* that spread by means of readily detached bulbils is far greater than the aesthetic value of their relatives, however.

The dye Turkey red is prepared from the seeds of *Peganum harmala* (family Peganaceae), a common plant of arid regions in southwestern and Central Asia.

NATURAL HISTORY

Seed dispersal. Within the Geraniales, the family Geraniaceae is perhaps the most interesting from the point of view of seed dispersal; the genus *Erodium* (storksbill), particularly well represented in desert and subdesert areas, shows adaptation to this environment. The fruit is furnished with a long beak formed from the fused tissues of the styles, the narrow upper parts of the ovary segments, or carpels. The beak segments of all the carpels of each flower remain fused until the fruit is ripe. When the fruit ripens and dries, these segments spring apart and become spirally shaped; each seed-bearing carpel falls to the ground. With rain or increase in humidity, the beak segments absorb water and straighten, driving the carpels into the ground. In some species the inner surface of the beak segments have feathery hairs that aid in wind dispersal. In *Geranium* and some other genera, such as *Pelargonium*, the beak segments separate from the base and curl but remain attached at their tips. This abrupt dehiscence—splitting open—may either catapult the seeds away or leave them hanging by slender threads, from which they are easily dislodged.

Self-planting fruits

Within the family Zygophyllaceae is the genus *Tribulus*, which derives its name from the Latin name of an instrument of warfare—the caltrop, a four-pronged device that served as a horse-crippling spike. The fruit of *Tribulus* is similarly formed; the carpels become hard and are heavily covered with sharp spines and bumps. They are

Drawing by M. Moran based on (Ixonanthaceae, Houmiriaceae, Limnanthaceae) J. Hutchinson, *Families of Flowering Plants*, The Clarendon Press, Oxford; (Zygophyllaceae) drawing in H. Baillon, *Histoire des Plantes*, Hachette; (Erythroxylaceae) reprinted with permission of Macmillan Publishing Co., Inc., from *Taxonomy of Vascular Plants* by G.M.H. Lawrence, Copyright 1951 by the Macmillan Company

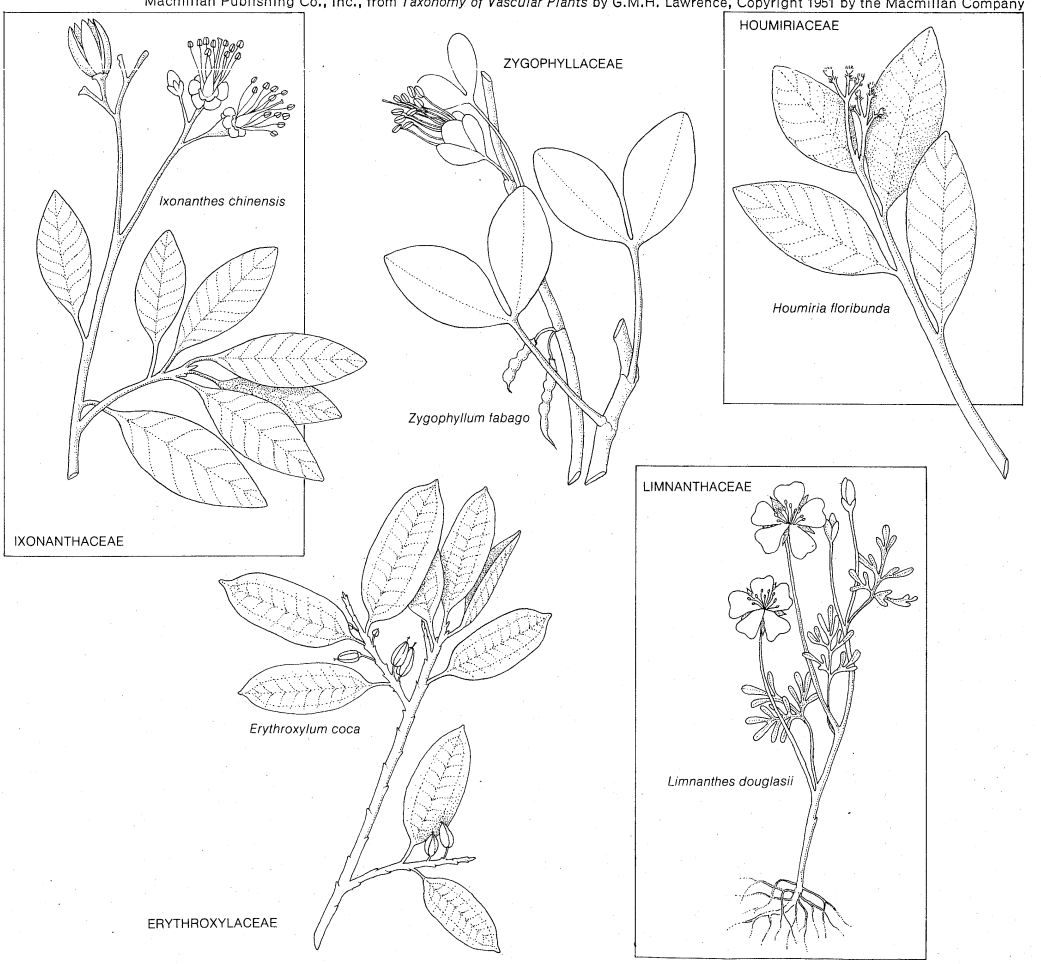


Figure 1: Representative plants of five of the smaller families of the geranium order.

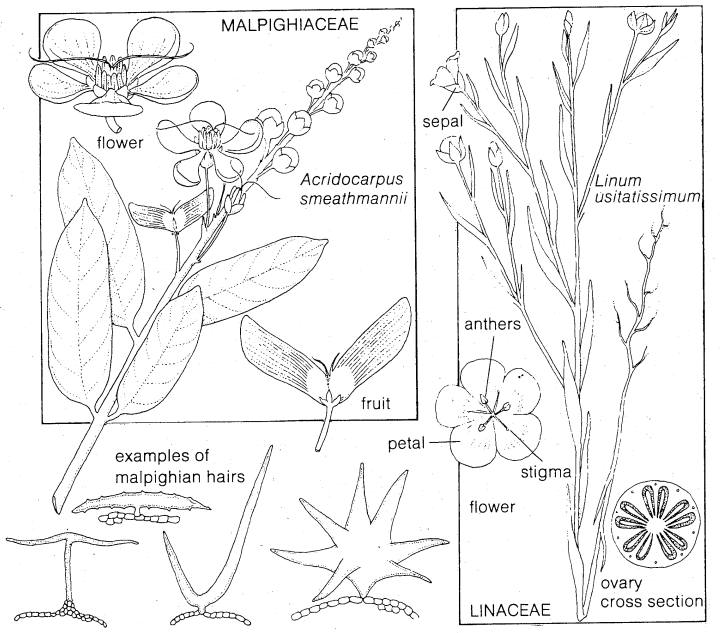


Figure 2: Representative plants of two of the larger families of the geranium order.

From (Malpighiaceae) Crown copyright. Reproduced with permission of the Controller of Her Majesty's Stationery Office and of the Director, Royal Botanic Gardens, Kew, (malpighian hairs) Engler, *Das Pflanzenreich*, (Linaceae) J. Hutchinson, *The Families of Flowering Plants;* The Clarendon Press, Oxford

picked up by grazing animals that disperse them to other parts of the arid regions. A further specialized form of distribution occurs in the genus *Impatiens* (family Balsaminaceae); the fleshy, five-valved capsule opens explosively. The valves remain attached at the base and apex when ripe but split along the joints, their elastic nature drawing base and apex together so that the seeds are shot, often for a considerable distance, through the spaces thus created between the valves. When the fruit is ripe, this explosive rupturing may be set off by a very light brush with the fingers—hence the name of a well-known European species, *Impatiens noli-tangere*, or touch-me-not. A Himalayan species, *I. glandulifera*, has spread widely along waterways in Europe and in North America by this means in a comparatively short space of time. Explosive dehiscence—opening of seed-bearing fruits—also occurs occasionally in the family Oxalidaceae, in this case activated by the elastic separation of the aril, an appendage on the seed, from the seed coat. Wind-distributed samaras, winged one-seeded fruits, as in the maples, occur in the family Malpighiaceae. Otherwise, dehiscent dry capsules, schizocarps, berries, and drupes are prevalent, with distribution by birds in the last two cases. Some (*e.g.*, *Vantanea* in the family Houmiriaceae) are distributed by flowing water.

Pollination. Pollination is almost universally by insects—generally by bees, wasps, and flies, but also by beetles and butterflies. Some North American species of *Impatiens*, however, are pollinated by hummingbirds. In the family Linaceae, floral dimorphism, the existence of two forms of flowers, has long been known. At least 29 species of *Linum* have been shown to exhibit long- and short-styled flowers. (The style is the narrow, upper part of the ovary, bearing the pollen-receiving surface, or stigma.) Fertility is greatest when the long-styled form is fertilized with pollen from the short-styled form and vice versa. The long-styled form is almost self-sterile, but the short-styled form is not; when both kinds of pollen are put on one kind of stigma, only that from the opposite kind of flower is capable of fertilization. Several species of *Linum* are self-fertilizing in the absence of insect visitors; the filaments bend inward, and the style branches spread with age.

Pollination mechanism in the flax family

In *Geranium* both self-sterility and self-fertility are known; it is well established that the more inconspicuous the flower, the more advanced the ability for selfing by movement of the styles and filaments, as in *Linum*. In the allied genus *Erodium*, the flowers may be radially symmetrical (regular or actinomorphic) or strongly bilaterally symmetrical (zygomorphic) in forms of the same species. Apparently, forms with regular flowers tend to be self-fertilizing, and those with zygomorphic flowers to be insect fertilized, the enlarged lower petals serving as a landing platform for visiting insects.

Flowers of the remaining large genus of the Geraniaceae, *Pelargonium*, are odorous at night and are visited by nocturnal insects. Such genera as *Tropaeolum* and *Impatiens* have bee-pollinated flowers with nectar secreted at the end of a spur formed by the sepals. The anthers shed their pollen shortly after the flowers open, the stigma of the same flower remaining immature. Insect visitors, dusted with pollen, transfer it to the mature stigmas of older flowers in which the anthers have fallen, thus ensuring cross-pollination. Insect-pollinated genera often have nectar guides in the form of dark blotches or streaks at the base of the petals. One member of the family Malpighiaceae, *Bunchosia gaudichaudii*, has bee-attracting glands on the outside of the calyx (sepals).

EVOLUTION

Fossil record. Fossils of the order are not of great quantity, and older records, such as those for *Balanites* and *Zygophyllum* from the Tertiary Period (about 65,000,000 to 2,500,000 years ago) of Europe, are considered suspect. The remains have been mostly as fruit or pollen. About 12 or 13 fossil plant groups have been placed in the family Houmiriaceae, and fruits of three different species of *Houmiria* have been found in Tertiary deposits in South America. A fruit of a *Sacoglottis* has been found in Tertiary deposits in Germany. *Wetherellia* fossils occur abundantly as fruits in the London Clay (Eocene Epoch—38,000,000 to 54,000,000 years ago) at the Isle of Sheppey, Kent, as do those of *Decaplatyspermum*. Fossil remains of *Erodium* and *Geranium* have been recorded from Baltic amber, as well as fruiting pedicels (stalks) of *Oxalidites* and a *Linum* fruit. *Linum* seed has been found in the Russian Pliocene Epoch (about 7,000,000 to 2,500,000 years ago), and pollen of *Geranium* in the Miocene Epoch (26,000,000 to 7,000,000 years ago).

Phylogeny. The order Geraniales is considered to have its closest affinity with the order Rutales; indeed, the general relationship is so close that it is difficult to distinguish the two orders. It is also very difficult to plot phylogenetic (evolutionary) relationships between the families included within the order Geraniales. Both the

Geraniales and the Rutales are among the many orders believed to have arisen from the order Saxifragales.

Uncertain relationships of order and families

Within the Geraniales, the families Tropaeolaceae and Balsaminaceae, with their specialized floral structure, are probably the most advanced families. The Balsaminaceae presents certain problems, however. It differs from the Geraniaceae in anatomy and pollen structure, and the disposition of the ovules is similar to that found in the Sapindaceae (order Sapindales, *q.v.*); on this basis, some authorities have included it in that family. It has also been placed near the family Vochysiaceae in the order Polygalales—an opinion given some weight by the similarity in development of the embryo in *Impatiens* and *Polygala*. Overall resemblances with the remaining families in the order Geraniales, however, outweigh these contradictory features.

The tropical family Hugoniaceae is considered to be the most primitive in the order; from it the family Linaceae probably arose. The families Ixonanthaceae, Houmiriaceae, and Erythroxylaceae may also be derivatives. The family Malpighiaceae, however, is close to the Erythroxylaceae and Nitrariaceae, and through the latter, a transitional family, to the Zygophyllaceae.

Two families are of particular interest for the form of their pollen grains. Those of the Limnanthaceae are unique and serve to emphasize its isolated position. They appear to be asymmetrically bicolpate (*i.e.*, they have two unequal germination pores or furrows). In the family Balsaminaceae the grains are very strongly flattened, with three or four germination furrows and a reticulate (netted, ridged) surface pattern. This strong compression clearly differentiates them from the Geraniaceae, with which they are most commonly allied, and all other families of the order. The closest resemblance is with *Jollydora* in the Connaraceae (order Connarales).

The order Geraniales is as diverse in anatomy as in pollen structure, particularly the family Linaceae and the small families that have been separated from it. Diversities within this group are so complex in pattern as to render difficult any phylogenetic interpretation based on anatomy. Similarly, anatomy provides no clues on the separation of the family Oxalidaceae or the Biebersteiniaceae from the Geraniaceae, all of which have the same basic vascular structure—the system of internal water- and food-conducting tubes, the vessels—and the same characteristic ring of thick-walled cells (sclerenchyma cells) in the pericycle, a layer of cells surrounding the central vascular cylinder in roots. This ring, however, is absent in other families that have been closely linked with the Geraniaceae, notably the Tropaeolaceae, Balsaminaceae, and Limnanthaceae. The family Balsaminaceae also differs from the Geraniaceae in the presence of raphide (needle-like crystals) sacs in the leaf and stem, which are sometimes visible as transparent dots. In addition, the vascular bundles in the leafstalks (petioles) are arranged in an arc, not a full ring.

The woods of the Zygophyllaceae are very specialized, showing this family to be a natural group. The family Balanitaceae differs from it in having high, wide wood rays; nevertheless, the general anatomy suggests affinity with the Zygophyllaceae rather than the family Simaroubaceae. The anatomy of the family Erythroxylaceae differs considerably from that of the Linaceae, with which it was formerly believed to be united. The wood structure in Malpighiaceae, more highly specialized than in the Linaceae, Houmiriaceae, and Erythroxylaceae, differs also from that of the Zygophyllaceae.

Thus, relationships within the order follow a confusing pattern because so many anomalous members are included, and no clear phylogenetic conclusions are possible.

CLASSIFICATION

Annotated classification. The geranium order is very heterogeneous, consisting of a number of families united by apparent general relationship rather than by a readily defined common group of characters. It is thus difficult to define by features that do not break down at some point or other.

ORDER GERANIALES

Leaves commonly alternate. Flowers hypogynous (*i.e.*, sepals, petals, and stamens arise at the base of the ovary), usually bisexual and radially symmetrical (actinomorphic). Sepals overlapped (imbricate) or rarely merely touching at the edges (valvate), free or more or less fused. Petals generally free, contorted (twisted), or imbricate, rarely valvate or convolute (rolled up, lengthwise). Anthers (pollen sacs) open by longitudinal slits. Nectar-producing disk or disk glands frequently present. Ovary of fused carpels with ovule attachment along central axis near base or apex of ovary chambers, ovules mostly solitary to few in each locule (chamber). Seeds mostly without endosperm (starch nutrient tissue for developing embryo).

Family Hugoniaceae

Trees, shrubs, or woody lianas with hooked prickles on the branchlets or flower stalks. Leaves alternate, simple, with entire or toothed margins, with deciduous (falling) stipules. Flowers actinomorphic, bisexual, in terminal or axillary (the upper angle between a leafstalk and stem) cymes (flower clusters that mature from the top downward) or panicles (many-branched flower clusters); more rarely in axillary clusters of spikes, or solitary, often yellow. Sepals and petals 5, free. Sepals generally imbricate, petals contorted. Stamens fused into a short tube at the base, almost invariably 10, except in the genus *Indorouchera* (5–7 stamens). Stamen filaments frequently with nectar-producing glands at the base. Ovary 3- to 5-locular, each locule with 2 ovules. Styles free or slightly fused at the base, with capitate—headed, with enlarged ends—stigmas. Fruit a slightly fleshy or berrylike drupe, the stones woody or bony, seed usually with scanty endosperm. Six genera and about 60 species distributed in the tropics of both hemispheres.

Family Linaceae (flax family)

Entirely shrubs, subshrubs, and herbs—no trees. Leaves alternate (rarely opposite), simple (entire or toothed). Stipules commonly absent or represented by stipular glands; if present, then falling early or, rarely, persisting and becoming conspicuous (in *Anisadenia*). Flowers actinomorphic, bisexual, disposed in cymes, racemes (spikelike flower clusters that mature from the bottom upward), or corymbs (flat-topped flower clusters), or very rarely solitary and lateral. Sepals and petals sometimes 4, usually 5; sepals sometimes fused at their bases, imbricate; petals free and contorted. Stamens equal in number to the petals, but between each pair of stamens a minute to conspicuous and bristle-like staminode (sterile stamen) frequently occurs. Stamen filaments more or less expanded at the base and fused into a short tube that usually bears small nectar-producing glands. Ovary occasionally 2- but usually 3- to 5-locular, each locule with 2 ovules. Styles free or shortly fused at the base, with capitate or more rarely linear stigmas. Fruit a capsule, rarely breaking up into 1-seeded segments twice as numerous as the styles (*Reinwardtia*) or not opening (indehiscent; *Anisadenia*). Endosperm usually scanty. About 6 genera and 240 species with nearly worldwide distribution.

Family Ixonanthaceae

Trees or shrubs with simple, alternate leaves that can be entire, toothed, or crenate. Stipules usually present, sometimes small and early deciduous; sometimes conspicuous, very long and convolute, leaving a conspicuous ring-shaped scar on the twig when they fall. Inflorescence an axillary or terminal raceme, cyme, or panicle. Flowers small, bisexual or rarely with male and female flowers on separate plants (dioecious), actinomorphic. Sepals and petals 5 or sometimes 4; sepals free or fused at the base, imbricate; petals free, imbricate, sometimes persistent and becoming firm in texture. Stamens 5 to 20, the filaments slender, free, inserted on or below the conspicuous annular or cup-shaped nectar-producing disk; anthers short and small. Ovary normally 4- or 5-locular, more rarely with only 2 locules, with a single simple or shortly divided style; ovules 1 or 2 in each locule. Fruit variable in form from large, fleshy (and edible) drupes, to winged samara, and 2- to 5-locular septicidal capsule. Arils (fleshy appendages on the seeds) present and variable, from vestigial to large organs completely enveloping the seed. Eight genera and about 45 species, exclusively tropical in distribution.

Family Houmiriaceae

Trees or shrubs with simple, entire- or lobe-margined, opposite leaves. Stipules either absent or small and early deciduous. Inflorescences axillary or rarely terminal, cymose or paniculate. Flowers hermaphroditic and actinomorphic. Sepals 5, imbricate, fleshy at the base and persistent, shortly fused at the base or united almost throughout their length into 5-toothed cup. Petals 5, contorted or imbricate, deciduous and membranous or persistent and of thick texture. Stamens variable in number, mostly 10 to 30 in 1 or 2 series, sometimes very numerous (50–180) and then in several rows. Stamen filaments slender or thick, more or less fused below;

staminodes sometimes also present. Nectar-producing disk present, annular and often toothed, or divided into separate scales. Ovary typically 5-locular but sometimes 4-, 6-, or 7-locular, with a single simple style. Ovules 1 to 3 in each locule, pendulous. Fruit a drupe, the outer layers slightly to distinctly fleshy, the inner layer hard and woody and often with resin-bearing cavities, usually with only 1 or 2 seeds developing; endosperm of seeds fleshy and oily. Eight genera and about 50 species distributed in tropical regions of the New World and 1 species in tropical West Africa.

Family Erythroxylaceae

Trees and shrubs with simple, entire, alternate leaves often showing 2 persistent longitudinal folds (leaves opposite in the genus *Aneulophus*, however). Flowers bisexual, actinomorphic, small, disposed in axillary fascicles, or solitary. Sepals 5, fused into bell-shaped calyx with valvate or imbricate teeth. Petals 5, free, deciduous, convolute or imbricate, frequently with a tonguelike appendage within. Stamens 10 in 2 whorls, more or less fused into a tube below, persistent. Ovary 2- or 3-locular, usually only 1 locule fertile. Ovules 1 or 2 pendulous. Styles free or fused, with oblique stigmas. Fruit drupaceous. Seeds usually with fleshy endosperm. Four genera and more than 200 species distributed in the tropics of both hemispheres, mostly in Africa, however; only the genus *Erythroxylum* is found in America.

Family Lepidobotryaceae

Small tree with alternate, single-bladed leaves with stipules, the single leaflet subtended by stipels—appendages similar to stipules—which, like the stipules, are deciduous. Petioles (leafstalks) and petiolules (the petiole, or stalk, of a leaflet in a compound leaf) jointed. Flowers actinomorphic, small, dioecious, the male flowers in short, sessile, catkin-like axillary inflorescences; female flowers in fasciculoid racemes. Sepals 5, imbricate, shortly fused at the base. Petals 5, free. Male flowers with 10 stamens on the margin of a flesh disk and a rudimentary ovary. Female flowers similar but with sterile stamens and a 3-chambered ovary with 2 collateral ovules in each locule. Styles 3, fused at the base. Fruit a single-seeded capsule. Seeds partly covered by a fleshy aril. One genus and species (*Lepidobotrys staudtii*), native to tropical Africa.

Family Malpighiaceae

Small trees, shrubs, or very frequently woody lianas. Malpighian hairs—characteristic 1-celled, 2-branched plant hairs—present. Leaves normally opposite and entire-margined (though occasionally alternate or in whorls of 3 or with wavy, toothed, or lobed margins), and frequently dotted with glands. Stipules usually present and deciduous, rarely absent, sometimes large, conspicuous, fused together, and persistent. Inflorescence in terminal or axillary raceme. Flowers usually large, bisexual or rarely polygamous (with male, female, and bisexual flowers on the same plant), commonly yellow or red (sometimes white or blue), obliquely zygomorphic, some cleistogamous (closed, self-pollinated) flowers often occurring. Sepals 5, imbricate, free or slightly fused at the base, some or all frequently with a large sessile (stalkless) or stalked gland on the lower outer surface. Petals also 5, free, imbricate, usually distinctly clawed and often toothed or fimbriate (ragged or fringed) along the margin. Stamens 10, in 2 whorls, usually fused into a ring at the base, all fertile or frequently with some reduced to staminodes; disk inconspicuous. Ovary asymmetrically disposed, carpels usually 3 (rarely 2 or 4), free or more or less fused, each with 1 ovule. Styles equal in number to the carpels, free or rarely fused, or very rarely only 1 carpel with a style. Fruit typically a schizocarp breaking up into 3 frequently winged mericarps, but sometimes nutlike or drupaceous. Seeds without endosperm. Sixty genera and 800 to 900 species distributed exclusively in tropical regions of the world, especially in South America.

Family Nitrariaceae

Stiff shrubs, often armed with spines or rigid leafstalks. Leaves alternate or clustered, thick and fleshy, simple, entire or with a few teeth at the apex, and furnished with small stipules. Flowers small, yellowish or white, actinomorphic, bisexual, in short lax cymes in the axils of the bracts. Sepals considerably fused below, imbricate, persistent. Petals 5, free and valvate, concave. Stamens 10 to 15 with 5 opposite the sepals and the remainder opposite the petals singly or in pairs. Stamen filaments subulate (awl-shaped), without appendages; disk absent. Ovary of 3 carpels, rarely of 6, gradually narrowing above into 3 decurrent stigmas; each carpel with a single pendulous ovule. Fruit berrylike, with a fleshy exocarp and a thin, bony, grooved endocarp. Seeds without endosperm. One genus (*Nitraria*) with about 10 species occurring in saline deserts, mostly from the Sahara to Central Asia and Siberia, with 1 species in Australia.

Family Zygophyllaceae

Mostly woody perennials; if annual they are sometimes very succulent (some *Zygophyllum* species); rarely trees (*Guaiacum*). Branches often jointed at the nodes in the more fleshy species. Leaves usually (but not invariably) opposite, mostly pinnately compound (with small leaflets on both sides of a central axis), sometimes simple or with 2 leaflets (these also jointed to the leafstalks in fleshy *Zygophyllum* species. Stipules present, often persistent, sometimes in the form of spines. Flowers in cymes or solitary and terminal, actinomorphic or rarely zygomorphic, bisexual. Sepals 5 (rarely 3 or 4), free or sometimes fused at the base, imbricate or rarely valvate. Petals equal in number to the sepals (or rarely none), free, usually imbricate or contorted. Stamens in 1 or 2 whorls, each whorl containing the same number as the sepals and petals. Stamen filaments frequently furnished with straplike appendages at the base. Nectar-producing disk usually present between the stamens and the ovary. Ovary usually with 4 or 5 carpels, sometimes as few as 2 or up to 12, each carpel with 1 or many pendulous ovules. Ovary narrowed above into a single unparted or shortly divided style. Fruit a capsule, frequently angled or even broadly winged, or rarely dividing into single-seeded mericarps. Seeds with or without endosperm. About 23 genera and 225 species distributed almost entirely in arid or semi-arid areas of the tropics and subtropics of the world.

Family Balanitaceae

Similar to the Zygophyllaceae in most respects, but leaves totally without stipules; spines in leaf axils, not modified stipules. Disk thick. Fruit drupaceous with a very thick, bony, 5 angled endocarp but only a thin fleshy mesocarp. One genus (*Balanites*) with 25 species in tropical Africa and Asia.

Family Peganaceae

Perennial herbs or subshrubs. Leaves alternate, deeply divided into narrow segments, with stipules. Flowers solitary, leaf opposed, actinomorphic, and bisexual. Sepals 4 or 5 valvate, sometimes pinnately divided. Petals 4 or 5, free, imbricate. Stamens 12–15, in alternating rows, the filaments thicker below but without appendages, inserted on an annular or angular nectar-producing disk. Ovary 2-, 3-, or occasionally 4-locular, each locule with numerous ovules disposed along the central axis. Style simple, 2- or 3-keeled above with the stigmatal surfaces on the keels. Fruit a many-seeded capsule (*Peganum*) or a 2-loculed berry (*Malacocarpus*). Seeds with a fleshy endosperm. Two genera with about 6 species distributed in dry places from the Mediterranean area to Mongolia and in the southern U.S. and Mexico.

Family Oxalidaceae

Mostly herbs (sometimes fleshy), sometimes shrubs, rarely trees. Fleshy axillary or underground tubers or bulbils frequently produced. Leaves alternate, palmately or pinnately divided (often trifoliolate—3 leaflets—in *Oxalis*), rarely simple by suppression of all but 1 leaflet; stipules absent. Flowers actinomorphic, bisexual, solitary or in umbels, more rarely in cymes or racemes, minute, nonpetalled, cleistogamous flowers sometimes present. Sepals usually fused below and imbricate. Petals 5, free or only shortly fused below and contorted. Stamens 10, in 2 alternating rows, basally fused, 5 of them sometimes reduced to antherless staminodes. Ovary 5-locular, each locule with 1 or more axile ovules. Styles 5, distinct; stigmas capitate or shortly divided. Fruit a capsule. Five genera with more than 900 species distributed predominantly in the tropics, particularly in southern Africa, South America, and Mexico.

Family Geraniaceae (geranium family)

Plants chiefly herbaceous but sometimes suffrutescent (woody at the base only) or shrubby, the leaves alternate or opposite, mostly lobed or palmately or pinnately divided, and generally with stipules. Flowers large and showy to small, radially or bilaterally symmetrical, bisexual, solitary to umbellate. Sepals 4 or 5, free or fused to the midpoint, imbricate with valvate tips, a series of bracts resembling sepals (epicalyx) sometimes present. Petals five, rarely 4 or none, free, imbricate or contorted, usually with alternating nectar-producing glands. Stamens mostly 10, usually fused at the base, in 2 rows (but the row opposite the petals sometimes reduced to staminodes), or more rarely 15 and fused in bundles of 3. Ovary of 3 to 5 carpels with 1 or 2 ovules in each locule (many in *Balbisia*). Style simple or very short, with 5 stigmas. Fruit of 3 to 5 separating mericarps, each commonly bearing a water-absorbing awn of hardened tissue ("beak"), rarely a capsule. Seeds without or with scanty endosperm. Eight genera and about 800 species with worldwide distribution, especially in temperate regions.

Family Vivianiaceae

Woody herbs or subshrubs. Leaves opposite, simple, entire to toothed, lacking stipules. Flowers radially symmetrical, bisexual, in loose clusters or laxly cymose in the upper leaf axils. Calyx (sepals) tubular or bell-like with 4 or 5 "teeth." Petals 4 or 5, free, contorted, alternating with nectar-producing glands. Stamens sometimes 8, usually 10, free. Ovary 2- or

3-locular with 2 ovules in each locule. Styles 2 or 3, fused or almost free. Fruit a 3-lobed, 3-valved capsule. Seeds with a fleshy embryo. One genus (*Viviania*) with 30 species, confined to Brazil and Chile.

Family Biebersteiniaceae

Perennial herbs with multicellular glandular hairs. Leaves pinnate or pinnatipartite, alternate, with stipules. Flowers radially symmetrical, bisexual; disposed in terminal racemes or panicles. Sepals 5, free, imbricate. Petals 5, free, frequently with "teeth" at the tip, and alternating with nectar-producing glands. Stamens 10, filaments shortly fused into a ring at the base. Ovary of 5 carpels, deeply lobed, set on a short stalk. Styles 5, arising from the base of the lobes and fused into a capitate stigma. Ovules solitary in each locule, pendulous. Fruit of 5 indehiscent, 1-seeded mericarps. Seeds with scanty endosperm. One genus (*Biebersteinia*) with 5 species occurring in arid places from southeastern Europe to Central Asia.

Family Dirachmaceae

Shrubs with long and short shoots bearing alternate, simple, serrate leaves and persistent stipules. Flowers solitary in the leaf axils of the long shoots, bisexual, and subtended by an epicalyx of 4 small bracts. Sepals 8, free, valvate. Petals 8, contorted. Stamens 8, opposite the petals, with large anthers. Ovary of 8 carpels, deeply lobed, each locule containing 1 ascending ovule. Style solitary with 8 linear stigmas. Fruit a capsule, with 8 ventrally opening sections, woolly within. Seeds with scanty endosperm. One genus and species (*Dirachma socotrana*) occurring on the island of Socotra.

Family Tropaeolaceae

Succulent herbs climbing by prehensile leafstalks. Leaves alternate, entire (often shield-shaped) or variously lobed or palmate, with stipules (stipules, however, sometimes rudimentary). Flowers bilaterally symmetrical, bisexual, often showy in shades of red or yellow, solitary in the leaf axils. Sepals 5, fused below, the dorsal sepal modified and produced into a long spur, imbricate or valvate. Petals 5, the upper 2 often different from the lower 3 (which may be absent), clawed, inserted on the calyx, imbricate. Stamens 8 in 2 rows of 4, curved downward, with slender filaments. Ovary 3-locular, 1 pendulous ovule in each locule. Fruit of 3 finally hardening indehiscent carpels separating from a central axis. Seeds without endosperm. Two genera, *Magallana* (2 species) and *Tropaeolum* (about 90 species), confined to Central and South America from Mexico to Chile.

Family Balsaminaceae (balsam family)

Weak herbs with juicy, frequently translucent stems and alternate or opposite simple leaves. Stipules indicated by small glands. Flowers solitary or in umbel-like clusters, strongly bilaterally symmetrical, frequently showy in shades of pinks, purple, yellow, and orange. Sepals 5 (but 2 frequently reduced or aborted), the lowest prolonged into a spur, coloured. Petals 5, corolla 2-lipped, the upper petal large and ascending, the laterals fused in pairs. Stamens 5, with short, broad filaments fused at the tip; anthers fused around the ovary. Ovary 5-locular with numerous ovules in each locule. Styles short, or the 1 to 5 stigmas almost sessile (set directly on the ovary). Fruit a succulent capsule opening elastically by action of fleshy arils attached to seeds, rarely a berrylike capsule (*Hydrocera*). Seeds with scanty endosperm or none. Two genera with more than 450 species mostly in tropical Asia and Africa but some in Europe and North America.

Family Limnanthaceae

Succulent annual marsh plants. Leaves alternate, much divided, without stipules. Flowers solitary on long stalks in the leaf axils, radially symmetrical, bisexual. Sepals and petals 3 to 5, the former valvate and the latter contorted. Stamens 6 or 10 in 2 rows, the outer row alternate with the petals, frequently with basal glands. Carpels 3 or 5, almost free, with a single common enlarged gynobasic style and 1 ovule in each locule. Fruit of free indehiscent carpels. Seed without endosperm. Two genera and 11 species confined to North America.

Family Hypseocharitaceae

Stemless perennial herbs. Leaves in a rosette, pinnately divided, without stipules. Flowers bisexual, radially symmetrical, either sessile (no stalks) or with short stalks in few-flowered cymes. Sepals and petals 5, the former imbricate, the latter contorted. Stamens 15 in a single series, with persistent slender filaments. Ovary 5-locular, lobed, with a single style and numerous axile ovules in each locule. Fruit an irregularly loculicidal capsule, opening late. Seeds numerous, lacking endosperm. One genus (*Hypseocharis*) with 8 species occurring in the Andes.

Critical appraisal. Above the generic level, it seems unlikely that orthodox taxonomic methods based on external morphology will shed much more light on classification within the order or between the Geraniales and allied orders—particularly the Rutales, Sapindales, and Polygalales. In addition, many conclusions based on other plant disciplines, such as palynology and phytochemistry, are still founded on inadequate sampling of the families concerned, and more complete information may shed the necessary light, particularly if species showing primitive characters are first selected. At the generic level and below, taxonomic revisions of such large and troublesome genera as *Oxalis* and *Impatiens* are still much needed.

BIBLIOGRAPHY. J. HUTCHINSON, "Malpighiales," in *The Genera of Flowering Plants*, vol. 2, pp. 568–621 (1967); E.F. WARBURG, "Taxonomy and Relationship in the Geraniales in the Light of Their Cytology," *New Phytol.*, 37:130–159, 189–210 (1938), an interesting and useful paper dealing with the families Linaceae, Zygophyllaceae, Geraniaceae, Limnanthaceae, Oxalidaceae, Tropaeolaceae, and Balsaminaceae; O. WARBURG and K. REICHE, "Balsaminaceae," in A. ENGLER and K. PRANTL (eds.), *Die natürlichen Pflanzenfamilien*, 3: 383–392 (1897), a brief conspectus—more recent literature is extensive but scattered; O.E. SCHULZ, "Erythroxylaceae," in A. ENGLER (ed.), *Das Pflanzenreich IV*, 134:1–176 (1907), a very competent monograph, still the only general account of the family; R. KNUTH, "Geraniaceae," and "Oxalidaceae," *ibid.*, vol. 129–130 (1912, 1930) not of outstanding taxonomic merit, but the only general treatments of these families available; J. CUATRECASAS, "A Taxonomic Revision of the Humiriaceae," *Contr. U.S. Natn. Herb.*, 35:25–214 (1961), a very fine revision including an account of the relationships and evolution of the family, with an invaluable bibliography; H. WINKLER, "Linaceae," in *Die natürlichen Pflanzenfamilien*, 2nd ed., 19a:82–130 (1931), a now rather dated but still useful account, including in subordinate rank the Houmiriaceae, Ixonanthaceae, and Hugoniaceae; J. LEONARD, *"Lepidobotrys* Engl., type d'une famille nouvelle de Spermatophytes: les Lepidobotryaceae," *Bull. Jard. Bot. État. Brux.*, 20:31–40 (1950), the case for the separation of this genus into a new family; C.T. MASON, "A Systematic Study of the Genus *Limnanthes* R. Br.," *Univ. Calif. Publs. Bot.*, 25:455–506 (1952), a sound taxonomic revision of the genus, containing most of the Limnanthaceae, with information on cytology and experimental genetics; F. NIEDENZU, "Malpighiaceae," in *Pflanzenreich IV*, 141:1–870 (1928), of average taxonomic merit, marred by the fact that little or no use was made of the great British or American herbaria in its preparation but still quite invaluable; F. BUCHENAU, "Tropaeolaceae," *ibid.*, 131:1–36 (1902), competent, but of limited value since the number of species recognized has now almost doubled; A. ENGLER, "Zygophyllaceae," *ibid.*, 2nd ed., 19a:144–184 (1931), a brief conspectus of the family.

(C.C.T.)

German Democratic Republic

The most developed and prosperous of the Communist countries of eastern Europe and one of the major industrial nations of the world, the German Democratic Republic (often referred to as the GDR and called East Germany in the West) lies in a vital strategic region. The nation was established on October 7, 1949, as one of the two Germanys that emerged from the ruins of the Nazi period, a circumstance that molded much of the character of the young state during the years of the Cold War, as it attempted to construct a new and emphatically Communist society. By the 1970s the republic had attained a strong position economically and politically; its location in the continental heartland gives it a significant place—symbolized by the annual trade fair at Leipzig—in trade between eastern and western and between northern and southern Europe.

The country also derives trade advantages from the Baltic Sea coastline, its northern boundary. On the east the Oder and Neisse rivers form a long and important frontier with Poland; this boundary, part of the post-World War II Potsdam agreement, was confirmed in the Treaty of Zgorzelec of 1950. On the south the boundary with Czechoslovakia corresponds to an earlier boundary of 1918, renewed by treaty in 1945. On the west another long and major boundary separates the republic from the Federal Republic of Germany; it is based on the lines of demarcation, agreed upon at the Yalta Conference of 1945, separating the then Soviet occupation zone of Germany from the zones occupied by the Western allies, on whose territory the German Federal Republic subse-

Boundaries

quently emerged. Situated within the republic's territory, the 185 square miles of West Berlin form a unique political unit based on the three sectors of the city originally held by the Western powers; long a major potential source of conflict during the Cold War era, the western part of the city faded somewhat in international significance following a settlement by the major powers, announced in the summer of 1971. The republic's frontiers have thus had more than usual significance in determining the course of internal events.

The German Democratic Republic has an area of 41,768 square miles (108,178 square kilometres) and by the mid-1970s was the home of nearly 17,000,000 people. East Berlin ("Democratic Berlin") is the capital.

The following article covers various aspects of the land, people, and society of the contemporary republic. For related historical information, see GERMANS, ANCIENT; and GERMANY, HISTORY OF. See also BERLIN; LEIPZIG; ELBE RIVER; and ODER RIVER.

THE LAND

Topography. The German Democratic Republic lies across two major European topographic zones: a northern lowland belt, which narrows from east to west, and the smaller but rugged strip of the Mittelgebirge, or Mid-German Highlands, which borders it to the south. This basic division has had great influence on many aspects of the contemporary nation.

The northern lowlands. The northern lowlands of the republic consist mainly of wide, flat, and rolling plains, with the Elbe and the Oder river systems cutting only slightly into the terrain. Only the east–west ridges of the Nördlicher and Südlicher Landrücken rise to some 330 feet (100 metres) above the level plains.

The rocks underlying the plains are young Pleistocene sediments (less than 2,500,000 years old), thickening northward. Underlying older rocks include lignite (brown-coal) deposits in the south, but they lie too deep for easy exploitation. Exceptions include the important limestone quarries northeast of Berlin and the ancient rock-salt and potash deposits, which can be mined because of upthrusts associated with later mountain-building movements.

Effects of the Ice Age

The surface of the plains was shaped largely by Ice Age forces; a line from Wilhelm-Pieck-Stadt Guben, on the Polish frontier, to Brandenburg, west of Berlin, separates a region molded by southward-moving Scandinavian glaciers, which were formed during the Weichsel (or Vistula) Glacial Stage (the last Ice Age advance), from the more southerly belt, affected only by the earlier Saale and Elsterian Glacial stages. The northern portion of the lowlands exhibits the classic features of glacial molding—a lumpy, lake-strewn relief with terminal moraine ridges marking the limit of glacier-carried debris—while the older southern surfaces have been levelled, and their lakes have become dry land.

The plains, which were covered by clay and studded with ice-scoured boulders, thus generating loamy soils, are important agriculturally, but the terminal moraines tend to be heaped with boulders and sand, are much less fertile, and are often forest clad. Trees also cover the dry, sandy valley regions, but prehistoric meltwater valleys are marked by a natural grassland cover.

The southern edge of the lowlands is marked by a loess belt of windblown deposits laid down during the Weichsel Glacial Stage. Given favourable conditions, a rich, black soil developed on this rolling countryside, with less fertile podzol soils developing in the areas of higher rainfall. The whole belt, which stretches from the Harz Mountain foothills as far as Magdeburg (the Magdeburger Börde) in the west and the area of Lusatia, or Lausitz (Lausitzer Gefilde) in the east, has supported agriculture based on wheat, barley, and sugar beets.

The uplands of the south and southwest. South of the loess belt rises the contrasting topography of the Mittelgebirge region. In addition to the Harz mountains and the Erzgebirge (Ore Mountains) and the mountains of the Oberlausitz (Upper Lusatia) and the Thüringer Wald (Thuringian Forest), the region contains the rolling hills of the Thuringian and Erzgebirge basins and, near Dresden, the lowlands of the Elbe Valley. The core of the whole region is hard, crystalline rock associated with the ancient Hercynian mountain-building period (some 300,000,000 years ago), when molten granitic intrusions formed ore deposits. Toward the end of the Paleozoic Era (about 250,000,000 years ago), this eroded mountain range was filling its associated structural depressions with debris, and it was here that the region's valuable hard-coal beds—such as those of the Erzgebirge basin—were formed. Sediments of limestone, clays, and sands were packed into gigantic hardened rock masses during the ensuing Mesozoic Era, and the vast upthrusts associated with the Alpine mountain-building process consolidated them into the ranges seen today, although much of the Mesozoic cover has been worn away.

The highlands are thus a mosaic of many-layered structural blocks, some lifted and some sunk deeper during the Tertiary upheavals. Climatic and soil conditions have stimulated the growth of extensive forests—now largely spruce, as a result of 19th-century planting—above the 2,300-foot mark, whereas the rain-favoured higher meadows are increasingly utilized for cattle raising. The high rainfall is retained by an extensive dam system and is then piped to the industrial cities of the foothills and basins. Bismuth, cobalt, uranium, silver, zinc, lead, and tin ores are still mined in the mountains to which they have given their name, and, although industrialization has left its harsh mark on the regional landscape, the more rural areas retain a considerable scenic charm, which has fostered tourism.

Climate. The climate of the German Democratic Republic reflects the transitional climate of central Europe, which is determined by constantly interchanging air masses of maritime and continental origin. The exposed Baltic coastline enhances the maritime component.

The lowlands. In the lowland belt, precipitation diminishes eastward as the plains open toward the Eurasian interior, and the average temperatures for the warmest and coldest months become more extreme. Schwerin, in the northwest, for example, has an annual rainfall of 24.7 inches (627 millimetres), a mean annual maximum of 63.5° F (17.5° C), and a mean January minimum of 31.8° F (−0.1° C), while for Frankfurt an der Oder the corresponding figures are 21.3 inches (540 millimetres), 65.7° F (18.7° C), and 30.2° F (−1.0° C).

The uplands. To the south the mountains have a wetter and cooler climate, with westward-facing slopes receiving the highest rainfall from maritime air masses. At a station on the Brocken, a mountain in the Harz near the western frontier, annual precipitation reaches 58.4 inches (1,483 millimetres) at an altitude of 3,747 feet (1,142 metres). The sheltered lee slopes and the basins, by contrast, being in a rain shadow, are actually arid. The basins and low-lying tracts also tend to be warmer and hence are especially favourable to agriculture; fruit and grapes for wine, for example, are grown in the Elbe Valley region.

The human imprint. *Traditional regions and modern planning.* The human imprint on the republic's landscape also reflects a basic division, this time between a highly industrialized south and a predominantly rural north. Until 1945 metropolitan Berlin was the only industrial cluster north of a line from Wilhelm-Pieck-Stadt Guben to Magdeburg.

The rural north and the industrialized south

The precious metals of the southern mountains supported a dense population from late medieval times until silver mining declined with the discovery of New World deposits. The people turned to the development of other crafts, trades, and, notably, the textile and metalworking industries. The elaborate transportation network and the lignite- and salt-mining industry, developed in the later 19th century, diversified and enlarged the regional economy to encompass chemicals, heavy machinery, and power plants. In the north, meanwhile, the lack of mineral wealth, poor transport and marketing conditions, a sparse population, and, notably in Mecklenburg and Brandenburg provinces, the presence of a powerful class of landowning Junkers all impeded industrial development.

MAP INDEX

Political subdivisions

Cottbus.....51·45n 14·00e
Dresden.....51·10n 14·00e
East Berlin.....52·30n 13·25e
Erfurt.....51·10n 10·45e
Frankfurt.....52·30n 14·00e
Gera.....50·45n 11·45e
Halle.....51·30n 11·45e
Karl-Marx-Stadt .50·45n 12·45e
Leipzig.....51·15n 12·45e
Magdeburg.....52·15n 11·30e
Neubrandenburg.....53·30n 13·15e
Potsdam.....52·30n 12·45e
Rostock.....54·15n 12·30e
Schwerin.....53·30n 11·30e
Suhl.....50·40n 10·30e

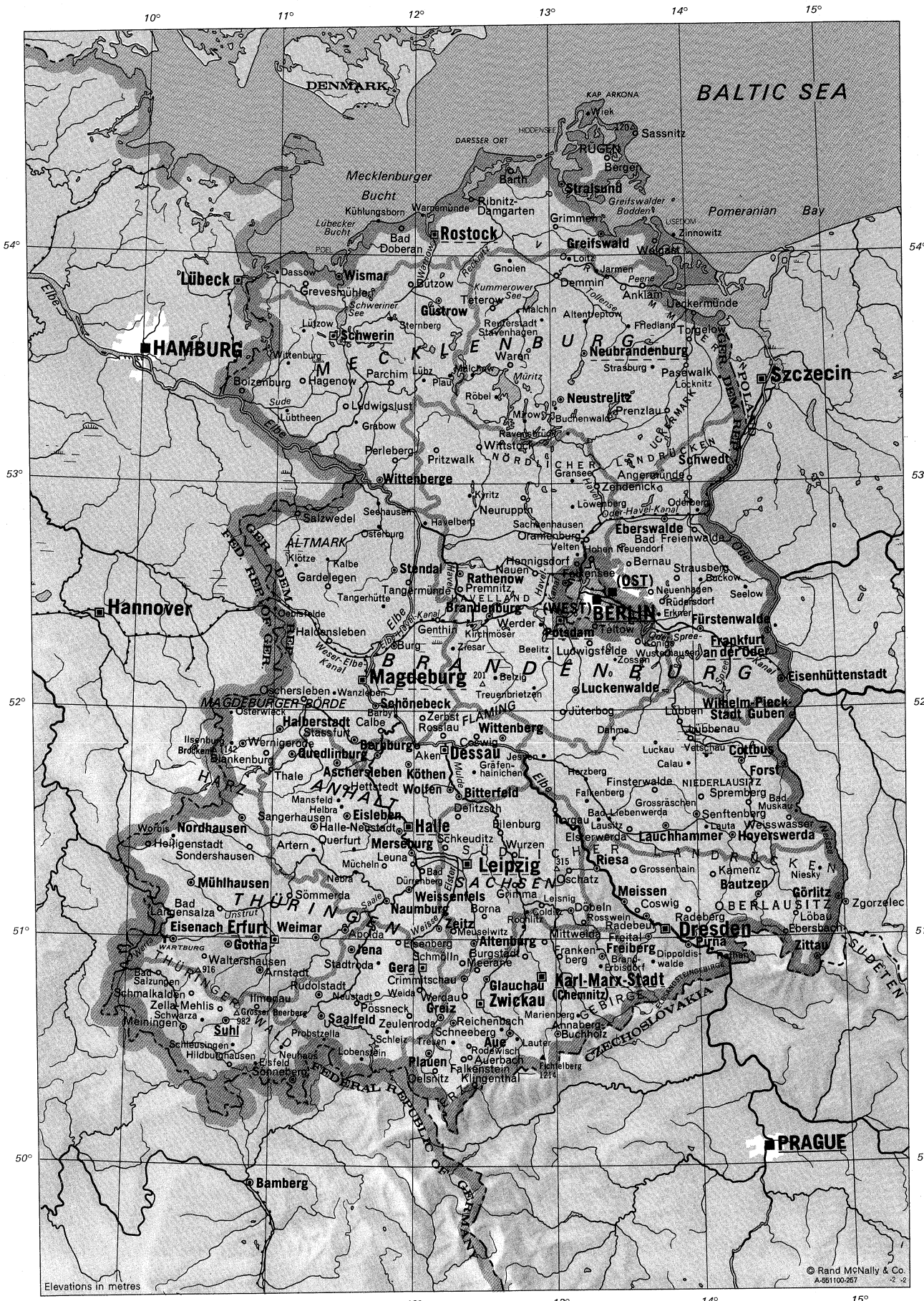
The names of the political subdivisions do not appear on the map because they are the same as the names of their capital cities.

Cities and towns

Aken.....51·51n 12·02e
Altenburg.....50·59n 12·26e
Altentreptow.....53·42n 13·14e
Angermünde.....53·01n 14·00e
Anklam.....53·51n 13·41e
Annaberg-Buchholz.....50·35n 13·00e
Apolda.....51·01n 11·31e
Arnstadt.....50·50n 10·57e
Artern.....51·22n 11·17e
Aschersleben.....51·45n 11·27e
Aue.....50·35n 12·42e
Auerbach.....50·30n 12·23e
Bad Blankenburg.....50·41n 11·16e
Bad Doberan.....54·06n 11·53e
Bad Dürrenberg.51·18n 12·04e
Bad Freienwalde.52·47n 14·01e
Bad Langensalza.....51·06n 10·38e
Bad Liebenwerda.....51·31n 13·23e
Bad Muskau.....51·32n 14·43e
Bad Salzungen..50·48n 10·13e
Barby.....51·58n 11·53e
Barth.....54·22n 12·43e
Bautzen.....51·11n 14·26e
Beelitz.....52·14n 12·58e
Belzig.....52·08n 12·35e
Bergen.....54·25n 13·26e
Berlin.....52·30n 13·25e
Bernau.....52·40n 13·35e
Bernburg.....51·48n 11·44e
Bitterfeld.....51·37n 12·20e
Blankenburg.....51·48n 10·58e
Boizenburg.....53·22n 10·43e
Borna.....51·07n 12·30e
Brandenburg.....52·24n 12·32e
Brand-Erbisdorf .50·52n 13·19e
Buchenwald.....53·21n 13·04e
Buckow.....52·34n 14·04e
Burg.....52·16n 11·51e
Burgstädt.....50·55n 12·49e
Bützow.....53·50n 11·59e
Calau.....51·45n 13·56e
Calbe.....51·54n 11·46e
Chemnitz see Karl-Marx-Stadt
Colditz.....51·07n 12·48e
Coswig.....51·07n 13·34e
Coswig.....51·53n 12·26e
Cottbus.....51·45n 14·19e
Crimmitschau.....50·49n 12·23e
Dahme.....51·52n 13·25e
Dassow.....53·50n 10·59e
Delitzsch.....51·31n 12·20e
Demmin.....53·54n 13·02e
Dessau.....51·50n 12·14e
Dippoldiswalde..50·54n 13·40e
Döbeln.....51·07n 13·07e
Dresden.....51·03n 13·44e
Ebersbach.....51·00n 14·35e
Eberswalde.....52·50n 13·49e
Eilenburg.....51·27n 12·37e
Eisenach.....50·59n 10·19e
Eisenberg.....50·58n 11·53e
Eisenhüttenstadt.52·10n 14·39e
Eisfeld.....50·26n 10·54e
Eisleben.....51·31n 11·32e
Elsterwerda.....51·28n 13·31e
Erfurt.....50·58n 11·01e
Erkner.....52·25n 13·45e
Falkenberg.....51·35n 13·14e
Falkensee.....52·33n 13·04e
Falkenstein.....50·29n 12·22e
Finsterwalde.....51·38n 13·42e
Forst.....51·44n 14·39e
Frankenberg.....50·54n 13·01e
Frankfurt an der Oder.....52·20n 14·33e
Freiberg.....50·54n 13·20e
Freital.....51·00n 13·39e
Friedland.....53·40n 13·33e
Fürstenwalde.....52·21n 14·04e
Gardelegen.....52·31n 11·23e
Genthin.....52·24n 12·09e
Gera.....50·52n 12·04e
Glauchau.....50·49n 12·32e
Gnoien.....53·58n 12·42e
Görlitz.....51·09n 14·59e
Gotha.....50·57n 10·41e
Grabow.....53·16n 11·34e
Gräfenhainichen.51·44n 12·27e
Gransee.....53·00n 13·09e
Greifswald.....54·05n 13·23e
Greiz.....50·39n 12·12e
Grevesmühlen...53·51n 11·10e
Grimma.....51·14n 12·43e
Grimmen.....54·07n 13·02e
Grossenhain.....51·17n 13·31e
Grossräschen....51·35n 14·00e
Güstrow.....53·48n 12·10e
Hagenow.....53·26n 11·11e
Halberstadt.....51·54n 11·02e
Haldensleben....52·18n 11·26e
Halle.....51·29n 11·58e
Halle-Neustadt...51·28n 11·56e
Havelberg.....52·50n 12·04e
Heiligenstadt.....51·23n 10·09e
Helbra.....51·33n 11·29e
Hennigsdorf.....52·38n 13·12e
Herzberg.....51·41n 13·14e
Hettstedt.....51·38n 11·30e
Hildburghausen .50·25n 10·44e
Hohen Neuendorf.....52·40n 13·16e
Hoyerswerda....51·26n 14·14e
Ilmenau.....50·41n 10·55e
Ilsenburg.....51·52n 10·41e
Jarmen.....53·55n 13·20e
Jena.....50·56n 11·35e
Jessen.....51·47n 12·58e
Jüterbog.....51·59n 13·04e
Kalbe.....52·40n 11·25e
Kamenz.....51·16n 14·06e
Karl-Marx-Stadt (Chemnitz).....50·50n 12·55e
Kirchmöser.....52·22n 12·25e
Klingenthal.....50·21n 12·28e
Klötze.....52·38n 11·10e
Königs Wusterhausen..52·18n 13·37e
Köthen.....51·45n 11·58e
Kühlungsborn...54·09n 11·43e
Kyritz.....52·56n 12·23e
Lauchhammer...51·30n 13·47e
Lausitz.....51·31n 13·21e
Lauter.....50·33n 12·44e
Lauta.....51·27n 14·04e
Leipzig.....51·19n 12·20e
Leisnig.....51·09n 12·56e
Leuna.....51·19n 12·01e
Löbau.....51·05n 14·40e
Lobenstein.....50·26n 11·38e
Löcknitz.....53·27n 14·12e
Loitz.....53·58n 13·07e
Löwenberg.....52·54n 13·08e
Lübben.....51·56n 13·53e
Lübbenau.....51·52n 13·57e
Lübtheen.....53·18n 11·04e
Lübz.....53·27n 12·01e
Luckau.....51·51n 13·43e
Luckenwalde....52·05n 13·10e
Ludwigsfelde....52·17n 13·16e
Ludwigslust.....53·19n 11·30e
Lützow.....53·40n 11·11e
Magdeburg.....52·07n 11·38e
Malchin.....53·44n 12·46e
Malchow.....53·28n 12·25e
Mansfeld.....51·35n 11·28e
Marienberg.....50·39n 13·10e
Meerane.....50·51n 12·28e
Meiningen.....50·34n 10·25e
Meissen.....51·10n 13·28e
Merseburg.....51·21n 11·59e
Meuselwitz.....51·02n 12·17e
Mirow.....53·16n 12·49e
Mittweida.....50·59n 12·59e
Mücheln.....51·18n 11·48e
Mühlhausen.....51·12n 10·27e
Nauen.....52·36n 12·52e
Naumburg.....51·09n 11·48e
Nebra.....51·17n 11·34e
Neubrandenburg.....53·33n 13·15e
Neuenhagen.....52·32n 13·41e
Neuhaus.....50·30n 11·08e
Neuruppin.....52·55n 12·48e
Neustadt.....50·44n 11·44e
Neustrelitz.....53·21n 13·04e
Niesky.....51·17n 14·49e
Nordhausen.....51·30n 10·47e
Oderberg.....52·52n 14·02e
Oebisfelde.....52·25n 10·59e
Oelsnitz.....50·24n 12·10e
Oranienburg.....52·45n 13·14e
Oschatz.....51·17n 13·07e
Oschersleben....52·01n 11·13e
Osterburg.....52·47n 11·44e
Osterwieck.....51·58n 10·42e
Parchim.....53·25n 11·51e
Pasewalk.....53·30n 14·00e
Perleberg.....53·04n 11·51e
Pirna.....50·58n 13·56e
Plau.....53·27n 12·16e
Plauen.....50·30n 12·08e
Pössneck.....50·42n 11·37e
Potsdam.....52·24n 13·04e
Premnitz.....52·32n 12·19e
Prenzlau.....53·19n 13·52e
Pritzwalk.....53·09n 12·10e
Probstzella.....50·32n 11·22e
Quedlinburg.....51·48n 11·09e
Querfurt.....51·23n 11·36e
Radeberg.....51·07n 13·55e
Radebeul.....51·06n 13·40e
Rathen.....50·59n 14·05e
Rathenow.....52·36n 12·20e
Ravensbrück.....53·12n 13·09e
Reichenbach....50·37n 12·18e
Reuterstadt Stavenhagen....53·42n 12·53e
Ribnitz-Damgarten.....54·15n 12·28e
Riesa.....51·18n 13·17e
Röbel.....53·23n 12·35e
Rochlitz.....51·03n 12·47e
Rodewisch.....50·32n 12·24e
Rosslau.....51·53n 12·14e
Rosswein.....51·03n 13·10e
Rostock.....54·05n 12·07e
Rüdersdorf.....52·29n 13·47e
Rudolstadt.....50·43n 11·20e
Saalfeld.....50·39n 11·22e
Sachsenhausen..52·47n 13·14e
Salzwedel.....52·51n 11·09e
Sangerhausen...51·28n 11·17e
Sassnitz.....54·31n 13·38e
Schkeuditz.....51·24n 12·13e
Schleiz.....50·34n 11·49e
Schleusingen....50·31n 10·45e
Schmalkalden...50·43n 10·26e
Schmölln.....50·53n 12·20e
Schneeberg.....50·36n 12·38e
Schönebeck.....52·01n 11·44e
Schwarza.....50·38n 10·32e
Schwedt.....53·03n 14·17e
Schwerin.....53·38n 11·25e
Seehausen.....52·53n 11·45e
Seelow.....52·32n 14·23e
Senftenberg.....51·31n 14·00e
Sömmerda.....51·10n 11·07e
Sondershausen..51·22n 10·52e
Sonneberg.....50·22n 11·10e
Spremberg.....51·34n 14·22e
Stadtroda.....50·51n 11·44e
Stalinstadt, see Eisenhüttenstadt
Stassfurt.....51·51n 11·34e
Stendal.....52·36n 11·51e
Sternberg.....53·43n 11·49e
Stralsund.....54·19n 13·05e
Strasburg.....53·30n 13·44e
Strausberg.....52·35n 13·53e
Suhl.....50·37n 10·41e
Tangerhütte.....52·26n 11·48e
Tangermünde...52·32n 11·58e
Teltow.....52·23n 13·16e
Teterow.....53·46n 12·34e
Thale.....51·45n 11·02e
Torgau.....51·34n 13·00e
Torgelow.....53·37n 14·00e
Treuen.....50·32n 12·18e
Treuenbrietzen..52·06n 12·52e
Ueckermünde....53·44n 14·03e
Velten.....52·41n 13·10e
Vetschau.....51·47n 14·04e
Waltershausen..50·53n 10·33e
Wanzleben.....52·03n 11·26e
Waren.....53·31n 12·40e
Warnemünde....54·10n 12·04e
Weida.....50·45n 12·04e
Weimar.....50·59n 11·19e
Weissenfels.....51·12n 11·58e
Weisswasser.....51·30n 14·38e
Werdau.....50·44n 12·22e
Werder.....52·23n 12·56e
Wernigerode.....51·50n 10·47e
Wiek.....54·37n 13·17e
Wilhelm-Pieck-Stadt Guben....51·57n 14·43e
Wismar.....53·53n 11·28e
Wittenberg.....51·52n 12·39e
Wittenberge.....53·00n 11·44e
Wittenburg.....53·31n 11·04e
Wittstock.....53·10n 12·29e
Wolfen.....51·40n 12·16e
Wolgast.....54·03n 13·46e
Worbis.....51·25n 10·21e
Wurzen.....51·22n 12·44e
Zehdenick.....52·59n 13·20e
Zeitz.....51·03n 12·08e
Zella Mehlis.....50·39n 10·39e
Zerbst.....51·58n 12·04e
Zeulenroda.....50·39n 11·58e
Ziesar.....52·16n 12·17e
Zinnowitz.....54·04n 13·55e
Zittau.....50·54n 14·47e
Zossen.....52·13n 13·27e
Zwickau.....50·44n 12·29e

Physical features and points of interest

Altmark, *historic region*.....52·40n 11·20e
Anhalt, *historic region*.....51·30n 11·30e
Arkona, Kap, *cape*.....54·41n 13·26e
Baltic Sea.....54·40n 14·45e
Brandenburg, *historic region*...52·00n 13·30e
Brocken, *mountain*.....51·48n 10·37e
Darsser Ort, *cape*.54·29n 12·31e
Elbe, *river*.....53·22n 10·31e
Elbe-Havel-Kanal, *canal*....52·24n 12·23e
Elbsandsteingebirge, *mountains*.....50·50n 14·20e
Erzgebirge, *mountains*.....50·30n 13·15e
Fichtelberg, *mountain*.....50·26n 12·57e
Fläming, *physical region*.....52·00n 12·30e
Greifswalder Bodden, *bay*....54·15n 13·35e
Grosser Beerberg, *mountain*.....50·37n 10·44e
Harz, *mountains*..51·45n 10·35e
Havel, *river*.....52·53n 11·58e
Havel-Kanal, *canal*.....52·36n 12·55e
Havelland, *physical region*..52·25n 12·45e
Hiddensee, *island*.....54·33n 13·07e
Kummerower See, *lake*.....53·49n 12·52e
Lübecker Bucht, *bay*.....54·00n 10·55e
Magdeburger Börde, *plain*....52·00n 11·30e
Mecklenburg, *historic region*...53·30n 13·00e
Mecklenburger Bucht, *bay*.....54·20n 11·40e
Mulde, *river*.....51·10n 12·48e
Müritz, *lake*.....53·25n 12·43e
Neisse, *river*.....52·04n 14·46e
Niederlausitz, *physical region*..51·45n 14·30e
Nördlicher Landrücken, *physical region*..53·10n 13·30e
Oberlausitz, *physical region*..51·10n 14·20e
Oder, *river*.....53·10n 14·22e
Oder-Havel-Kanal, *canal*....52·52n 14·02e
Oder-Spree-Kanal, *canal*....52·23n 13·41e
Peene, *river*.....54·09n 13·46e
Poel, *island*.....54·00n 11·26e
Pomeranian Bay.54·00n 14·15e
Recknitz, *river*...54·14n 12·28e
Rügen, *island*....54·25n 13·24e
Saale, *river*.....51·57n 11·55e
Sachsen, *historic region*.....51·00n 13·00e
Saxony, see Sachsen
Schweriner See, *lake*.....53·45n 11·28e
Spree, *river*.....52·32n 13·13e
Sude, *river*.....53·22n 10·45e
Südlicher Landrücken, *physical region*..51·20n 14·00e
Thüringen, *historic region*...51·00n 11·00e
Thüringer Wald, *mountains*.....50·30n 11·00e
Tollense, *river*...53·54n 13·02e
Uckermark, *physical region*.53·10n 13·35e
Unstrut, *river*....51·10n 11·48e
Usedom, *island*..54·00n 14·00e
Vorpommern, *physical region*..53·40n 13·45e
Warnow, *river*....54·06n 12·09e
Wartburg, *castle*..50·58n 10·18e
Weisse Elster, *river*.....51·26n 11·57e
Werra, *river*.....51·20n 9·56e
Weser-Elbe-Kanal, *canal*....52·14n 11·42e


BALTIC SEA
DENMARK
Mecklenburger Bucht
Pomeranian Bay
Rostock
Lübeck
HAMBURG
Schwerin
Wismar
Güstrow
Neubrandenburg
Szczecin
Neustrelitz
Wittenberge
Hannover
Magdeburg
Stendal
Rathenow
Brandenburg
BERLIN
(OST)
(WEST)
Potsdam
Eberswalde
Frankfurt an der Oder
Eisenhüttenstadt
Luckenwalde
Wilhelm-Pieck-Stadt Guben
Cottbus
Forst
Halberstadt
Quedlinburg
Aschersleben
Dessau
Wittenberg
Bitterfeld
Halle
Leipzig
Nordhausen
Mühlhausen
Erfurt
Weimar
Gotha
Jena
Gera
Zwickau
Karl-Marx-Stadt (Chemnitz)
Dresden
Meissen
Riesa
Bautzen
Görlitz
Zittau
Suhl
Saalfeld
Plauen
Aue
Freiberg
Pirna
PRAGUE
Bamberg
MECKLENBURG
BRANDENBURG
ANHALT
THÜRINGEN
SACHSEN
SÜDLICHER LANDRÜCKEN
OBERLAUSITZ
NIEDERLAUSITZ
CZECHOSLOVAKIA
FEDERAL REPUBLIC OF GERMANY
Elevations in metres
© Rand McNally & Co.
A-551100-257
GERMAN DEMOCRATIC REPUBLIC
Size of symbol indicates relative size of town
0 25 50 100 150 km
0 25 50 100 mi

A major goal of overall Socialist planning has been the development of the national territory as a whole. The southern industrial clusters, centred on, among others, the cities of Dresden, Karl-Marx-Stadt, Halle, and Leipzig, were restructured to diminish the disproportions created by the dismemberment of the Third Reich. With similar goals in mind, the formerly backward northern regions developed such new industries as, for example, shipbuilding, supplied by an industrial hinterland, in the port cities of Rostock, Wismar, and Stralsund. Special efforts were made to modernize agriculture and to develop a coordinated, industrialized food-producing and food-processing system of communes—model, state-owned farms and cooperatives and food-processing plants associated with them—in coordination with an extensive restructuring of the transportation system.

Urban settlement. The GDR considers communities of more than 2,000 persons urban; by this definition, almost three-quarters of the population is urban. Yet the distribution of city dwellers is very uneven; the largest concentration is in a zone beginning in the west, between the Harz and Magdeburg, reaching southeast through Dessau, Halle, and Leipzig, and then penetrating into the Erzgebirge valleys through Karl-Marx-Stadt and Zwickau, continuing eastward to Freiberg and Dresden. To the west the city series Gera–Jena–Weimar–Erfurt–Gotha–Eisenach forms a link to the fringe of the Thüringer Wald. In the north a cluster of urban centres has formed around Berlin, but, apart from that, generally in this region only local and provincial administrative centres are significant, which is directly opposite to the pattern in the industrialized south. Roughly one-third of the country's population still inhabits the great metropolitan complexes of Berlin, Halle, Leipzig, Karl-Marx-Stadt, Zwickau, and Dresden, but this pattern is changing with the accelerating need for new housing outside the old city centres.

Rural settlement. National plans call for the restructuring of villages. Many village communities were established in the 13th century in connection with a feudal eastward expansion policy. They tend to dissolve into smaller settlements with fewer than 100 inhabitants, notably in the Rostock, Schwerin, Neubrandenburg, Frankfurt, and Cottbus districts. The creation of large communal societies and cooperative units has been a catalyst in transforming rural settlement: such villages have multistoried housing blocks and a careful zonation into work areas, schools, and rural shopping, cultural, and service centres. Barns, sheds, haylofts, storerooms, and industrial meat-processing centres are set apart. The process has done away with privately owned farm property and even individual houses.

Changes in village communities

THE PEOPLE

Ethnic, linguistic, and religious divisions. The population of the German Democratic Republic, declining since World War II, dropped below the 17,000,000 mark in 1973 and by the end of 1975 was estimated to have fallen to 16,820,000. The overwhelming majority is made up of Germans who speak the German language. Regional dialects, notably the Plattdeutsch, or Low German, of the northern Mecklenburg region and the Saxon of Saxony in the south, continue to have significance in everyday life. In addition, many thousands (according to Western estimates, 35,000 to 70,000) of Sorbs (or Wends), descendants of early Slavic settlers, live in Lusatia, in Cottbus and Dresden districts. Other Slavs who had settled early in the Elbe and Saale basins and even farther westward retreated or were pushed out or Germanized during the medieval German colonization eastward. The Sorbs have preserved their individuality, and the region is officially bilingual, there are Sorbian schools and theatres, and the Domowina (the official Sorb homeland organization) protects the Sorbian heritage.

In religious affiliation, Protestants outnumber Catholics eight to one; adherence to religious groups is declining, however, especially among the young, although religious affiliations are ostensibly protected by the constitution.

Demographic structure. The German Democratic Republic is virtually unique among the major countries of Europe in that its population in 1976 was less than it was in 1947, when a peak was reached in the chaotic conditions following the collapse of the Third Reich. It declined steeply until 1961 and since then has declined at a slower rate. Distribution, as noted above, is very uneven, with a concentration in the southern industrial belt and in the north around Berlin. Urbanization is greatest in the Karl-Marx-Stadt and Leipzig districts and least in rural, northern Neubrandenburg.

German Democratic Republic, Area and Population

	area		population†	
	sq mi	sq km	1971 census	1975 estimate
Districts (*Bezirke*)				
Berlin, capital city	156	403	1,086,000	1,094,000
Cottbus	3,190	8,262	863,000	872,000
Dresden	2,602	6,738	1,877,000	1,845,000
Erfurt	2,837	7,348	1,256,000	1,247,000
Frankfurt	2,774	7,185	681,000	689,000
Gera	1,546	4,004	739,000	739,000
Halle	3,386	8,771	1,925,000	1,890,000
Karl-Marx-Stadt	2,320	6,009	2,047,000	1,994,000
Leipzig	1,917	4,966	1,491,000	1,458,000
Magdeburg	4,450	11,525	1,320,000	1,298,000
Neubrandenburg	4,167	10,793	638,000	629,000
Potsdam	4,854	12,572	1,133,000	1,125,000
Rostock	2,731	7,074	859,000	868,000
Schwerin	3,348	8,672	598,000	592,000
Suhl	1,489	3,856	553,000	550,000
Total GDR	41,768*	108,178	17,068,000*	16,891,000*

*Figures do not add to total given because of rounding. †De jure.
Source: Official government figures.

The legacy of World War II. As a result of World War II and its aftermath, and also because of emigration, the demographic structure of the republic has been markedly distorted. Such factors as the scarcity of young people, the large surplus of women in middle and older age groups, and, indeed, the general concentration in the older age groups continue to exert an influence. Birth rates and death rates emphasize the significance of these factors. Live births in the mid-1970s were only a low 11 per 1,000 population, and this in spite of the extremely low degree of infant mortality. Similarly, in a population with a predominance of elderly persons the death rate is high (about 14 per 1,000) in spite of a steadily declining mortality rate among younger people and an average life expectancy of 69 years for men and 74 years for women. The highest live-birth figures are found in the rural areas of the north, with a decline southward as urbanization increases.

No other European country has such a high proportion of pensioners—almost 20 percent in 1974, compared with 11 percent for the same areas in 1939. The proportion of the population in the labour force declined from 68 to 50 percent over the same period. Questions concerning automation, the rationalization and mechanization of the national economy, and the problems and potential of the social-welfare resources of the country become much clearer when seen against such a demographic background.

Another peculiarity is the surplus of women. In 1946 there were 135 women for every 100 men as a result of World War II and its aftermath. This disproportionality subsequently improved with the return of prisoners of war and with the death of women who strongly predominate in the older age groups; by 1974 the ratio had fallen to 116 to 100. This female surplus tends to increase with the size of communities; the trend toward equalization of the sex ratio is strongest in places with such male-oriented industries as mining and power plants, as in Halle-Neustadt and the Cottbus district.

The surplus of women

Contemporary trends in the labour force. By the mid-1970s more than 8,309,000 persons were gainfully employed. The increase of more than 500,000 since the early 1950s resulted largely from the entry of women into the labour force; the number of male workers showed a decrease. The fact that almost half the labour force is fe-

male has important consequences, notably in the provision of children's nurseries, kindergartens, play centres, all-day (rather than the usual 8 AM to noon) schools, and so on. A growing number of women are moving into leading professional positions, particularly in Berlin and in the industrial south, where both supply of and demand for trained personnel are highest. The working force is also augmented (to about 7 percent of the total) by persons who have qualified for retirement pensions but prefer to continue working.

Declining agricultural employment

More than a third of the working force is now employed in industry, and the proportion is steadily rising. A similar increase is taking place in commerce and service industries, while agriculture takes a declining eighth of the total, still a high figure compared with other countries. The steadily increasing need for workers, especially in industry, has caused planners to consider further automation and rationalization plans. In the geographic distribution of the labour force, the basic north–south contrast again emerges: agricultural workers in 1974 accounted for 29 percent and industrial workers for 17 percent of the total in the Neubrandenburg district, but in the district of Karl-Marx-Stadt the percentages were 6 and 51, respectively.

Internal migrations also take place in the republic, though to a declining extent; former movements trended from smaller rural settlements to larger towns, especially where these were the focus of new, planned industrial expansion. Men and young persons (18–30 years old) are prominent in this internal migration, causing social and economic changes in both the departure and reception areas. Centres of attraction are Berlin and the Cottbus and Frankfurt regions (both focal points for planned new growth). Karl-Marx-Stadt and Halle, because of a declining mining industry, and rural Neubrandenburg and Magdeburg are the main losers. Before the sealing of the national boundaries (especially around West Berlin) in 1961, a substantial migration, including much trained manpower, to West Germany caused severe problems to the national economy.

The outlook. It is of some significance that the low point of available working people seems likely to occur by the mid-1970s. The disproportionality is expected to diminish through the end of the decade and beyond, with the proportion of older people who are pensioners, to take a key gauge, falling from some 20 percent in 1970 to an estimated 14 percent by 2000. The final working out of the asymmetry introduced by World War II should strengthen immensely the economic and social development of the country. (H.-J.K./R.E.H.M.)

THE NATIONAL ECONOMY

The German Democratic Republic is eastern Europe's most developed and prosperous country. Its national income per person has been estimated by the World Bank to be considerably higher than that in any other European Communist country, with the exception of Czechoslovakia. While its agriculture accounts for a fairly small part of total east European farm production, in industry it occupies a very prominent position. It is an important producer of machinery, especially of advanced equipment, much of which is sold to other eastern European countries. Compared with other countries of the region, however, its overall economic progress has been relatively slow; during 1970–74, for example, its net material product grew by only about 26 percent, a rate that was exceeded in every east European country except Czechoslovakia.

International standing of the country's industry

The extent and distribution of resources. *Mineral resources.* The republic is relatively poor in mineral resources: only lignite, potash, and rock salt are found and produced in significant quantities, although other minerals, such as hard coal, iron ore, copper, bismuth, cobalt, and uranium are mined on a small scale. A small quantity of natural gas is produced, but there are no oil resources to speak of. By 1975, annual output of lignite (brown coal) amounted to about 270,000,000 tons, making the country the world's largest producer. Output grew rapidly up to 1964, but in subsequent years it showed little change. Much of the country's lignite comes from two main fields: Niederlausitz (northeast of Dresden) and the Halle–Leipzig region. There are a number of smaller producing areas, of which that lying near Magdeburg and Stassfurt is the most important. The lignite is easy to extract but has a high water content and a low calorific value; its main use is in the form of briquettes for thermal power stations and as domestic fuel, though large quantities are also used by the chemical industry.

Coal, potash, and iron ore

Local supplies of hard coal are not adequate to meet domestic requirements. Production, which is based on two small fields lying, respectively, between Karl-Marx-Stadt and Zwickau and between Freital and Dresden, has been declining, and by 1975 it amounted to scarcely 600,000 tons annually. Hard-coal imports (including coke), mainly from the Soviet Union, Poland, and Czechoslovakia, amounted to about 15 times as much. Potash and salt, which are found in large quantities on the eastern edge of the Harz Mountains, near Halle, play a significant part in the country's chemical industry and are important export products. Iron ore is mined in Thuringia and near Ilsenburg in the Harz, but reserves are small, the ore is of poor quality, and production had virtually ceased by 1975. Copper is found at Mansfeld in Thuringia, and uranium is produced in the Elster Valley and in the Erzgebirge.

Other resources. More than 60 percent of the country's total land area is agricultural land. Most of it is arable, only about an eighth being classified as meadows and pastures. The quality of the soil varies considerably from area to area, but, with the exception of that in the Thuringian basin and between Leipzig, Halle, and Magdeburg, it is not very productive. More than a quarter of the country's surface is wooded; some 80 percent of the forest area consists of coniferous trees, notably Scotch pine, spruce (especially in the Harz Mountains), and fir. Deciduous forests are made up of beech, oak, and birch. Extensive reforestation has taken place since the country was established. Hydroelectric potential is limited; small amounts are generated in the Erzgebirge and the Sudeten region.

Sources of national income. *Agriculture.* In the mid-1970s GDR agriculture, which is fairly advanced when compared with that of its Communist neighbours, contributed about 11 percent of the country's net material product and occupied about 11 percent of the labour force. There is one 15-horsepower tractor for every 76 acres (31 hectares) of arable land, a figure surpassed only by Czechoslovakia in eastern Europe, and the consumption of chemical fertilizers—at more than 220 pounds per acre (260 kilograms per hectare)—is one of the highest among Socialist and a number of European capitalist countries. As a result, yields are relatively high, but the country is still dependent on imports for a large part of its food.

Farming is organized on the basis of nearly 5,800 cooperatives and a number of state farms. The latter occupy about 7 percent of the socialized land area, which itself takes up all but 5 percent of the total agricultural area. The principal crops include potatoes, followed by sugar beets, barley, wheat, rye, oats, and oilseeds. Livestock raising is also important. Western observers feel that one of the main problems in agriculture has been the relatively slow and uneven growth in output, resulting from the lack of adequate investment funds and the lack of incentive thought to be associated with collectivization. From time to time there have been large shortages of certain commodities, necessitating large-scale imports. The position was particularly difficult in 1969, when, partly as a result of unfavourable weather, the cereal crop was some 12 percent lower than in the previous year, while the output of oil-bearing crops and sugar beets was, respectively, 28 and 26 percent lower than the average of the previous five years.

Industry. The German Democratic Republic is one of the most heavily industrialized Communist countries. Industry contributed more than 67 percent of the total net material product by the mid-1970s—more than in any

other east European country except Czechoslovakia—and engaged more than 38 percent of the total labour force. Industrial development was very rapid, the index of industrial production advancing more than 170 percent during the 1950s, 70 percent during the '60s, and 28 percent in 1970–74. Much of this progress resulted from improvements in labour productivity, which, per person, rose almost 210 percent between 1955 and 1974.

Industrial policy

The main early feature of industrial policy was the disproportionate attention given to heavy industry and a preoccupation with quantitative rather than qualitative progress. By the 1960s it was realized that these and other policy shortcomings were beginning to have an adverse effect on the rate of growth. A more rational economic and industrial policy, involving a limited decentralization of decision making and responsibility and a more equitable distribution of resources among the various sectors of industry, was initiated. The share of industrial investment devoted to energy, fuel, and metallurgy fell by half, to about a quarter of the total, during the early to late 1960s, but the share of light industry and the textile and food industries was expanded from 12 percent to 18 percent. Investment in the electronics industry more than doubled, to about 10 percent; engineering saw a rapid growth, but the share of chemicals remained constant.

The policy of the early and mid-1970s was to improve quality, make greater use of advanced technology, and raise productivity. The main emphasis is placed on fuel and power, engineering, chemicals, and electronics; specialization, especially in the field of sophisticated electronic and complex engineering products, is actively encouraged.

Chemicals, machinery, and electric and electronic products have been among the fastest growing branches of industry. Output in engineering (including vehicles), for example, grew by 90 percent in the 1960s and by the mid-1970s was the largest single branch of industry, accounting for nearly a quarter of total output. Food has a 17 percent share, closely followed by chemicals, light industry, and the rapidly expanding electronics sector. Other important branches include metallurgy and the relatively slowly advancing textile industry.

The output of electricity has been growing rapidly, expanding to more than 84,500,000,000 kilowatt-hours by the mid-1970s, but has failed to keep up with the expansion of demand. Most electricity is generated from lignite and coal. Oil, which is pumped from Kuybyshev in the Soviet Union to Schwedt and Leuna by means of the Comecon (the Communist bloc economic organization) Friendship (Russian Druzhba) pipeline, covers only a very small part of the country's total energy requirements. A similarly small amount (less than 5 percent) of electricity is derived from hydroelectric power stations and from the country's small atomic power plant, which opened in 1966.

The country is eastern Europe's principal producer of plastics and synthetic fibres; other important industrial products are pig iron, crude steel, rolled steel, chemical fertilizers, sulfuric acid, paper, and various passenger and freight vehicles.

Foreign trade. In per capita terms the country is eastern Europe's second largest foreign trader after Czechoslovakia, and foreign trade was about 75 percent higher during 1971–75 than during the previous five years. As it has no significant nontrade sources of foreign revenue, it has had to balance its trade (and sometimes even earn a small surplus) by limiting the growth of its imports to match its exports. In practice, however, exports have often exceeded imports by a fair margin. In contrast to the trend seen in other eastern European countries, the importance of the Communist world in the country's foreign trade has not declined significantly; by the mid-1970s the Communist bloc's share of the total was about 64 percent, only slightly less than in 1960, and among the highest in east Europe. Trade with the capitalist countries, which has grown faster than total trade, constituted about 31 percent of the total, and the remainder (about 5 percent) was made up of trade with the developing world.

Trading partners

The country's largest trading partner, namely, the Soviet Union, accounted for about a third of both exports and imports.

The GDR's principal non-Communist customer and supplier was the Federal Republic of Germany, which accounted for about 8 percent of both exports and imports. Since the mid-1960s the balance of trade between the two Germanys has been in favour of West Germany.

The scarcity of official statistics makes it difficult for Western observers to establish detailed commodity composition of the republic's foreign trade, but it is known that in the mid-1970s one-third of all imports consisted of raw materials (notably hard coal, coke, and crude oil), while another third was made up of the products of the metalworking industries. Products of the food, light, and related industries constituted more than 20 percent of the total, with agricultural produce (including wheat and oilseeds) accounting for roughly one-tenth. On the export side, more than 50 percent was derived from sales of machinery and metalworking products; the remainder was shared, about equally, by basic raw materials and light industrial and related goods. Sales to the republic by the advanced western European countries included machinery and foodstuffs (accounting jointly for just over half the total), manufactured products, and crude materials and chemicals. Exports to those countries included manufactured products and machinery (about half the total), chemicals, and food.

Management of the economy. *The role of the government.* As in other Communist countries, the government plays an all-important role in the economy. Agriculture is in the hands of the collectives and the state farms, while industry and most trading and financial enterprises are owned and controlled by the state. Reforms begun in 1972 and completed by 1975 virtually eliminated private enterprise, although about 14 percent of retail trade was still handled outside the Socialist sector.

Economic plans and the more important production targets are laid down, after discussion at all levels, by the government, as are the detailed economic policies necessary to fulfill these plans. The result, Western observers conclude, is a highly centralized and bureaucratic system of economic management, in which market forces are not allowed to operate to any significant extent. Western observers would, nevertheless, agree with the republic's own economists that, as a result of a gradual process of economic reform embarked on in 1963, the current system of controlling the economy is more flexible and sophisticated than it was at the start of the 1960s. The aim of these reforms was to eliminate some of the worst features of rigid central planning and detailed state supervision without seriously weakening the central control of the economy. As a result, economic plans were less detailed and binding production targets were laid down in respect to fewer goods; although prices were still determined by the state, price-fixing methods had been greatly improved.

Centralized economic management

At the same time, steps have been taken to give at least a limited role to such economic forces and instruments as demand, profit, taxation, and interest. Production targets for individual enterprises are now expressed in terms of profit and sales value rather than in terms of physical volume, and enterprises have been given some independence in deciding such matters as the type of goods produced, incentive payments, and the allocation of investment resources.

Direct control over the larger enterprises is exercised by some 80 industrial associations (Vereinigungen volkseigener Betriebe) that check the performance and coordinate the plans of individual enterprises and are responsible to the appropriate ministry. Smaller firms report to regional economic organizations *(Bezirkswirt-schaftsräte)*.

Trade unions. The country's trade unions are organized on the basis of industries and are affiliated with the national group known as the Freier Deutscher Gewerkschaftsbund (FDGB; Free German Trade Union Association), which claims a membership of some 8,000,000. They are not active in social and economic affairs in the

same way as their Western counterparts; as in other Socialist countries, they take part in discussions concerning, and then assist in the implementation of, national planning policies.

They may, in this respect, be considered an important part of the overall state apparatus, with responsibility for such matters as labour discipline.

TRANSPORTATION

The basic geographical division in the republic again may be seen in its transportation network, with a dense pattern of railways and highways in the southern industrialized regions contrasting with a thinner network in the agricultural north. Requirements of internal and foreign trade have strengthened this pattern, and topographic conditions have also played a part; both the rugged topography of the Mittelgebirge and the northern fens, swamps, and lakes have made route construction expensive.

Coastal conditions have strongly influenced the flow of shipping and harbour construction; fogs, low winter temperatures, snow, and the freezing of inland waterways are also important factors affecting costs, performance, and the accident rate.

The railway system. The state railway system, the Deutsche Reichsbahn (DR), still partly suffers from extensive war damage and reparations dismantling, but it is being modernized and rationalized. Uneconomical secondary lines, whose traffic is handled more efficiently by road vehicles, have been closed. Some short, local narrow-gauge lines remain. Added emphasis has been given to fast, modern intercity traffic. Railway tracks are most dense in the southern industrial agglomerations, and important lines connect Berlin to the major regional centres, which are themselves interlinked. Main freight lines in Saxony have been electrified.

A noteworthy feature is the Seddin marshalling yard south of Potsdam, a vast control centre handling international traffic at the very heart of the whole central European railway system.

Roads and highways. Highway traffic is playing an increasing role in transport. Most vehicles are state-owned, but a few private or semiprivate companies survive. Private motorcar ownership is low by western European standards. The road network is dense—thinner in the rural north but very dense and congested in the industrial south. The arterial system dates from the mostly long, straight 19th-century highways, the *Chausseen*. The *Autobahnen* (expressways) system, dating from Nazi times and influenced in design by military considerations, is only partially suited to contemporary traffic needs in the republic. Starting from the incomplete Berliner Ring, expressways radiate to the northeast, east, southeast, southwest, and west. In the south a cross-connection joins Dresden, Karl-Marx-Stadt, Gera, and Erfurt. The Leipzig–Dresden link was completed in the 1970s, the Berlin–Rostock expressway and remaining Berliner Ring are under construction, and plans exist for a Halle–Magdeburg connection.

The *Autobahnen*

Inland waterways. The lowland portions of the republic offer favourable conditions for a part natural, part artificial waterway system. About 80 percent of the total network is made up of regulated or canalized river reaches, while wholly artificial canals make good use of the east–west courses of ancient riverbeds. Main natural thoroughfares are the Elbe and its tributaries, the Saale (navigable up to Halle–Trotha), the Havel (navigable up to Zehdenick), and the Oder. The Elbe–Havel, Havel, Oder–Havel, Hohensaaten–Friedrichsthaler, and Oder–Spree canals and, farther west, a portion of the important Mittelland Kanal as far as the locks at Magdeburg–Rothensee form the main artificial links connecting the rivers; the east–west routes carry the most traffic. As the waterway system does not extend to the industrialized south, insofar as larger vessels are concerned, it carries only a small portion of freight; tourist traffic is of some significance, especially along the Elbe and on the scenic Baltic coast.

In spite of unfavourable natural conditions, the Baltic coast also has a number of commercial shipping port facilities, which are capable of handling all of the republic's overseas trade. In addition to the traditional port cities of Wismar, Rostock, and Stralsund, new deepwater harbour facilities on the Breitling at Rostock–Petersdorf serve as a gateway to the world.

Air traffic. Regular, scheduled air traffic was instituted in 1956 between more than half a dozen internal airfields and to 36 destinations in Europe, Asia, and Africa. The Berlin-Schönefeld Central Airport (Zentralflughafen Berlin-Schönefeld) dominates the system. Interflug is the small domestic airline and the Soviet airline Aeroflot the other major operator.

Pipelines. The important Schwedt refinery is the terminus of the Friendship (Druzhba) pipeline that brings crude oil from Russia. Schwedt also has links to the oil port at Rostock and to the Leuna chemical works near Leipzig, while further links for crude oil and products are planned.

ADMINISTRATION AND SOCIAL CONDITIONS

The structure of government. *The constitutional framework.* Under the terms of a constitution ratified in April 1968, the German Democratic Republic is a Socialist country. According to the prevailing interpretation of Marxist–Leninist theory, all classes of the population belong to an alliance known as the National Front of the German Democratic Republic (Nationale Front der Deutsche Demokratische Republik), an organization that embraces all political parties and mass organizations. This, it is claimed, represents an organizational expression of the principle that each citizen carries a responsibility for the direction of the entire community. Political power is exercised through the People's Delegations (Volksvertretungen), which form the basis of all political, economic, social, and cultural agencies in the republic, according to the principle embodied in the slogan "Work together, plan together, govern together" ("*Arbeite mit, plane mit, regiere mit*!"). Elections to the delegations are open to those who have attained their 18th birthday on the day of the election. Citizens over 21 can participate in elections to the national People's Legislature (Volkskammer). As in other Communist countries, much guided public discussion of basic questions of policy, as well as public examination of the candidates offered, precedes the actual election. The names of all candidates appear on a unified list approved by the ruling SED (see below).

The Volkskammer is the supreme agency of political power, resolving the main principles under which social and political life are to be carried on, both for individuals and for the various agencies and bodies concerned. It consists of 500 delegates elected for four-year terms.

The Council of State (Staatsrat) carries on the function of the legislative body between sessions. The Ministerial Council (Ministerrat), executive of the Staatsrat, carries out administrative and policy decisions that include planned development of the economy. Each council member is responsible for the management of the region assigned to him.

The regional People's Delegations function as elective agencies of the central government, and each produces its own councils and commissions to carry out national policies.

The dominant spearhead of this political movement is the Sozialistische Einheitspartei Deutschlands (SED), a product of the merger in 1946 of the Kommunistische Partei Deutschlands (KPD) and the Sozialdemokratische Partei Deutschlands (SPD). Other political parties tolerated are the Liberal-Demokratische Partei Deutschlands (LDPD, established 1945), the Christlich-Demokratische Union (CDU, established 1945), the Demokratische Bauernpartei Deutschlands (DBD, established 1948), and the Nationaldemokratische Partei Deutschlands (NDPD, also set up in 1948). Mass trade-union, youth, and women's organizations belong to the Demokratischer Block der Nationalen Front.

Political parties

International associations of the republic include adherence to the military Warsaw Pact and to Comecon, the economic organization linking east Europe; economic

ties are particularly close with the Soviet Union. More broadly, the GDR maintains economic and cultural links with many nations and is a member of more than 200 international organizations. The German Democratic Republic was accepted as a member of the United Nations in September 1973, following the signing of a mutual recognition treaty with the Federal Republic of Germany in December 1972.

The legal system. The highest judicial body is the Supreme Court (Oberstes Gericht), and its members (elected for four years), as well as those of the local and regional courts, are selected carefully for their understanding of the Socialist character of the social system. Prosecution for violations of law is handled by a body of public prosecutors, headed by an attorney general.

The armed forces. In 1956 the National People's Army (Nationale Volksarmee) replaced the garrison units of the People's Police (Volkspolizei, or VP), formed in 1952. It cooperates closely with the Soviet armed forces (which have some 300,000 troops stationed in the republic) and those of other Communist countries. There are also armed combat groups and the VP. These forces (equipped with Soviet weapons), along with Soviet troops and Czechoslovak forces periodically stationed in the republic, guard an important western sector of the border of the Communist bloc. Compulsory military service for men has existed since 1962.

Education. The school system of the German Democratic Republic is uniform and carefully structured. Its nucleus is the polytechnic public school (*Oberschule*), which consists of 10 grades and offers a general education. There is also an extensive preschool system of day nurseries and kindergartens. After the 10-year public school education, studies may be continued, first in vocational schools, then, after gaining the skilled workman's certificate known as the *Facharbeiterbrief*, in trade and technical schools and schools for engineering. Apprentices may earn their high school diplomas (*Abitur*) in a special course offered in the vocational school itself, at evening high schools, or at factory schools or then may take a mature student's high school diploma. The *Abitur* enables the student to continue studies at a *Hochschule* or university. Polytechnic high schools offering a general education also confer the *Abitur*. Other specialized institutions include factory schools, village academies, radio and television academies, and bodies specializing in agriculture and health care.

The *Abitur*

The public colleges (*Volkshochschulen*) and factory training schools (*Betriebsoberschulen*), together with other special schools, complete a complex educational system generally acknowledged as having a high level of proficiency though heavily imprinted by Marxist dogmas. Advancement is based on ability; there are no tuition fees.

The well-organized system of public libraries is also part of the educational apparatus; it is focussed on the Deutsche Staatsbibliothek (a significant centre of science collections) in Berlin and the Deutsche Bücherei, the national library based in Leipzig. The latter includes a collection of almost all the literary output in German since 1913. There is an elaborate system of regional, university, and specialist libraries, in addition to some 12,600 public library branches. There are also many valuable museums.

Social conditions and services. The republic possesses an extensive health-care system, administered by a Cabinet office. Financial support and free medical assistance, medicine, and nursing care are available in case of sickness or accident. The scheme, which is available to all citizens, is funded in part by a uniform and obligatory national insurance system, the management of which is part of the national budget responsibility. There is, in addition to the health-care system, a comprehensive preventive medical program and care, including a pension system, for the aged, who form a significant element in the population. Certain occupational groups—including miners, postal and railroad workers, and professionals—receive higher pensions. There are about 170 sanatoriums.

The republic's housing policy allows each citizen and his family the right to adequate quarters insofar as it is consistent with local conditions and the political and economic situation. The concentration on reconstruction, especially the setting up of heavy industry, in the country's early years impeded a successful solution of the gigantic housing problem, although the changed national priorities were alleviating the situation by the 1970s. Within the family women enjoy full equality; families with many children, unwed mothers, and one-parent families receive special attention.

Of the total national income of the republic, about 23 percent is put by for capital accumulation, and the remainder is used for consumption. About 7 percent goes to national educational funds, the same amount to sciences, art, and culture generally, and about 20 percent for the social insurance plan. Even though private home ownership still plays a considerable role in the country, it is slowly declining as a result of state and cooperative housing developments. Rents vary regionally, with Berlin rents running 50 percent higher than those in localities of fewer than 100,000 people.

Changing consumer conditions are reflected in figures that show that between 1955 and 1974 automobile ownership (per 100 households) rose from 0.2 to almost 24, motorcycles from 11 to 20, radios from 77 to 96, television sets from 1.2 to 80, and refrigerators and washing machines from 0.5 to 70.

CULTURAL LIFE AND INSTITUTIONS

Much of the cultural life of the German Democratic Republic has been conditioned by three basic and inescapable factors: the rich German cultural heritage; the dismemberment of Germany and the Cold War era; and the introduction of a new cultural environment as one of the fundamental precepts of the Socialist society. The state now plays a prominent role in all cultural activities, including physical culture, sports, and tourism, all of which are regarded as components of Socialist culture.

The fine arts. Among the organizations founded with the new aim of creating a realistic art closely linked to the people was the Association of Artists (Verband Bildender Künstler) in 1950, while a decree on culture (Kulturverordnung) of the same year endeavoured to assure the material well-being of individual artists. Commissions for the creation of monumental works of art, notably at the former concentration camps of Buchenwald, Ravensbrück, and Sachsenhausen, influenced the development of the plastic arts, and ideological "guidance" has been constantly exercised.

Association of Artists

In the early postwar years, architecture also reflected Socialist ideological influence modelled on Soviet designs; several major showpieces were carried out regardless of cost, but sometimes exaggerated decorations and eclectic designs led to tasteless results, as, for example, the overall appearance of the Karl-Marx-Allee in Berlin. In an attempt to remedy the situation, principles of urban construction closer to contemporary Western forms were laid down to serve as the foundation for ensuing architectural design.

These principles perhaps found best expression in such entirely new planned cities as Eisenhüttenstadt and Hoyerswerda, as well as in the reconstructed parts of some of the older, historic cities such as Berlin, Dresden, and Leipzig.

Literature. Literature plays an important part in the cultural life of the republic. The publishing industry, centred in Berlin and Leipzig, produces many works stemming from the rich classical German tradition and also the works of a good number of modern writers. Representatives of the Communist tradition who are in the first rank internationally include Bertolt Brecht, who died in 1956 and whose austere poetry, sardonic humour, and biting social commentary left their mark on readers in many lands; the social-documentary novels of Anna Seghers, perhaps best illustrated by *Das siebte Kreuz* (1939; *The Seventh Cross*, 1942); and, possibly more significantly, Arnold Zweig, whose *Der Streit um den Sergeanten Grischa* (*The Case of Sergeant Grischa*, 1927) was a seminal work of the pre-Nazi era. The Bit-

terfeld Conference of 1959 formulated the role of the writer in the evolving Communist society and generated the movement known as the Movement of Writing Workers (Bewegung Schreibender Arbeiter) to encourage latent talent in the portrayal of "Socialist Realism" while at the same time allowing closer supervision and guidance by the authorities.

The theatre. All the theatres in the republic are owned by the state, a circumstance that, among other things, ensures an adequate subsidy. There are some 110 permanent theatres, almost 1,000 cultural centres, and almost 200 open-air seasonal theatres, including the Bergtheater Thale and the Felsenbühne in Rathen. The Deutsches Theater in Berlin opened as early as September 1945, being the first German theatre to perform following the Nazi collapse. Other historic theatres, including the Deutsche Staatsoper, the Komische Oper, and the Volksbühne—all in the capital—and the Schauspielhaus and the Oper in Leipzig, as well as the old Nationaltheater in Weimar, were rebuilt. Theatres have forged a strong functioning link between performers, audiences, and social organizations concerned with cultural affairs, and they perform a vital role in disseminating the ideological message.

The Berliner Ensemble

Berlin dominates theatrical life, and is the location of the Berliner Ensemble, long associated with Bertolt Brecht and probably the best known group internationally. Its often austere realism and disciplined professionalism have had a strong influence on contemporary theatre in many countries. Children's and youth theatres in Berlin, Halle, Leipzig, and Dresden are another new feature; their performances include not only productions of traditional favourites, such as fairy tales, but also productions that attempt to meet the problems of modern youth. Theatre, judging by the number of people attending performances, is highly popular.

Music. Concerts, too, are extremely well attended; the republic has more than 80 major orchestras of one kind or another, with the Leipziger Gewandhausorchester and the Staatskapelle in both Berlin and Dresden holding leading positions. The Thomaner Choir in Leipzig and the Kreuzchor of Dresden have a long and renowned tradition in the field of church music. There are also many amateur groups.

Sports and athletics. Sports occupy an extraordinarily large place in GDR life, under the aegis of the constitution, which invokes sport as essential to "development of a Socialist personality." Physical education and swimming are compulsory subjects in the schools, and around 5 percent of the labour force works at least part-time as coaches or sports officials. Members of state-supported sports clubs live on the club premises, and younger athletes even go to school on the club grounds, their classes being arranged around their training schedules. The German College of Physical Culture (Deutsche Hochschule für Körperkultur, founded 1950) in Leipzig has a renowned medical school that trains sports doctors. Women are trained as the men are, and women athletes, in consequence, have developed remarkably. Training emphasis is very much on Olympic sports, and its success was demonstrated at the summer Olympic Games in 1976, when the GDR took second place in gold medals won—40, surpassed only by the Soviet Union with 47; its women swimmers won 10 of 11 individual titles and broke eight world records.

The mass media. The state-owned film-making enterprise, DEFA (from Deutsche Film-Aktiengesellschaft), is located in Potsdam–Babelsberg and produces a wide range of cartoons, documentaries, and feature films, many of which are exported. Motion pictures continue to be a popular form of entertainment, especially in the rural communities. The highly developed radio network Deutscher Demokratischer Rundfunk is also government-owned and plays an important part in national life, while Radio Berlin International broadcasts on German subjects to many countries, including those of the developing world.

The Deutscher Fernsehfunk television network transmits through 10 channels, in two main program formats, from studios located in Berlin. It is a member of Intervision—the system linking east European countries—and, through this connection, of the west European Eurovision network.

Newspapers in the republic function as propagators of a Socialist viewpoint. There are some 40 dailies as well as many weekly, regional, and specialist publications. *Neues Deutschland*, the official newspaper of the Sozialistische Einheitspartei Deutschlands, is probably the best known internationally. (H.-J.K./R.E.H.M.)

THE OUTLOOK

The focal point of the future development envisaged for the German Democratic Republic is the economy. Up to the early 1960s, in the wake of the havoc wrought by World War II and in the light of the problems following the establishment of two German states, the main aim of economic policy was to ensure rapid growth in certain branches of industry, even at the expense of other industrial areas and sectors of the economy. More recently, steps have been taken to allocate investment funds more equitably, both within industry itself and among the various sectors of the economy. In the industrial field, the aim is to improve the technical level of production and move toward a new production pattern by concentrating on more sophisticated products, especially in engineering and electronics. It is felt that it is in these fields that the country has the greatest competitive advantage, since most other east European countries are not in the position to supply the required technological input. The German Democratic Republic is an enthusiastic member of Comecon, mainly because that body's specialization schemes have, on the whole, been helpful to the government's objective of rapid development in the more advanced industrial products.

Competitive advantages

The outlook for the economy is promising. The remainder of the 1970s is expected to see more economic reforms, which should lead to a further improvement in the system of economic management. This should help to make some progress in solving the country's principal problems, which Western observers interpret as the wasteful use of investment resources and the irrational price structures. This, in turn, is likely to have a beneficial effect on productivity, growth, and the advance of exports and should ensure a satisfactory trade balance, despite the need for large-scale food imports.

The Berlin Accord of 1971 seemed to reflect a changing climate of European affairs and some easing of regional tensions. The question of German unification, however, seems certain to occupy the republic and its citizens for some time to come. (E.I.U./Ed.)

BIBLIOGRAPHY. Comprehensive monographs concerning the physical and human geography of the German Democratic Republic are few in number in any language, but there is a wealth of periodical literature, especially in the magazines *Petermanns geographische Mitteilungen, Geographische Berichte*, and *Zeitschrift für den Erdkundeunterricht* (all published in the German Democratic Republic). The *Verfassung der Deutschen Demokratischen Republik*, 2nd ed., 2 vol. (1969), reflects the social conditions of the German Democratic Republic in official form. Political, demographic, economic, scientific, and cultural developments are reflected in great detail in the *Statistisches Jahrbuch der DDR* (annual). For economic geography, *Ökonomische Geographie der Deutschen Demokratischen Republik* (1969), a collective work, ed. by HORST KOHL, and *Die Bezirke der Deutschen Demokratischen Republik*, also ed. by HORST KOHL (1974), are standard reference sources. See also H.J. KRAMM, *Beiträge zur ökonomischen Geographie der Deutschen Demokratischen Republik*, 2 vol. (1962). English-language studies include ARTHUR M. HANHARDT, *The German Democratic Republic* (1968); and PETER C. LUDZ, *The German Democratic Republic from the Sixties to the Seventies: A Sociopolitical Analysis* (1970); a standard reference work is PETER C. LUDZ *et al.* (eds.), *DDR Handbuch* (1975).

(H.-J.K./R.E.H.M.)

Germanic Languages

The Germanic languages, a branch of the Indo-European language family, include a number of extinct languages as well as the earlier and present forms of German,

Netherlandic, Afrikaans, English, Frisian, the Scandinavian languages, Yiddish, and their many dialects.

In numbers of native speakers, English, with 320,000,000, clearly ranks second among the languages of the world (after Chinese); German, with 120,000,000, probably ranks sixth (after Hindi-Urdu, Spanish, and Russian). To these figures may be added those for persons with another native language who have learned one of the Germanic languages for commercial, scientific, literary, or other purposes. Though not even approximate statistics are available, English is unquestionably the world's most widely used second language.

Table 1 presents information on each of the modern standard Germanic languages.

Table 1: Modern Standard Germanic Languages

	where spoken	native speakers (1970 estimate)	use as a 2nd language
English	Great Britain, Ireland, United States, Canada, Australia, New Zealand, Republic of South Africa	320,000,000	extreme
German	Germany, Austria, Switzerland (part)	120,000,000	extensive
Netherlandic (Dutch-Flemish)	The Netherlands, Belgium (part)	19,000,000	moderate
Swedish	Sweden, Finland (part)	9,000,000	slight
Danish	Denmark	5,000,000	slight
Afrikaans	Republic of South Africa (part)	5,000,000	slight
Norwegian	Norway	4,000,000	slight
Yiddish	various countries	3,000,000	slight
Frisian	The Netherlands, Germany	311,000	—
Icelandic	Iceland	205,000	—
Faeroese	Faeroe Islands	35,000	—

The earliest historical evidence for Germanic is provided by isolated words and names recorded by Latin authors beginning in the 1st century BC. From *c.* AD 200 there are Scandinavian inscriptions carved in the 24-letter runic alphabet. The earliest extensive Germanic text is the (incomplete) Gothic Bible, translated *c.* AD 350 by the Visigothic bishop Ulfilas (Wulfila), and written in a 27-letter alphabet of the translator's own design. Although the Gothic alphabet was hardly used outside of this Bible translation, later versions of the runic alphabet were used sparingly in England, Germany, and particularly Scandinavia—in the latter area down to early modern times. All extensive later Germanic texts, however, use adaptations of the Latin alphabet.

The names and approximate dates of the earliest recorded Germanic languages are recorded in Table 2.

Table 2: Earliest Recorded Germanic Languages

	dates* (AD)
Early Runic	200–600
Gothic	350
Old English (Anglo-Saxon)	700–1050
Old High German	750–1050
Old Saxon (Old Low German)	850–1050
Old Norwegian	1150–1450
Old Icelandic	1150–1500†
Middle Netherlandic	1170–1500†
Old Danish	1250–1500†
Old Swedish	1250–1500†
Old Frisian	1300–1500†

*Indicates approximate range of dates.
†Cutoff date for beginnings of modern Germanic languages.

The Germanic languages are related in the sense that they can be shown to be different historical developments of a single earlier parent language. Although for some language families there are written records of the parent language (*e.g.*, for the Romance languages, which are variant developments of Latin), in the case of Germanic no written records of the parent language exist. Much of its structure, however, can be deduced by the comparative method of reconstruction (a reconstructed language is called a protolanguage; reconstructed forms are marked with an asterisk). For example, a comparison of Runic *-gastiR*, Gothic *gasts*, Old Norse *gestr*, Old English *giest*, Old Frisian *iest*, and Old Saxon and Old High German *gast* "guest" leads to the reconstruction of Proto-Germanic **gastiz*. Similarly, a comparison of Runic *horna*, Gothic *haurn*, Old Norse, Old English, Old Frisian, Old Saxon, and Old High German *horn* "horn" leads scholars to reconstruct the Proto-Germanic form **hornan*. Reconstruction of parent language

Such reconstructions are, in part, merely formulas of relationship. Thus the Proto-Germanic **o* of **hornan* in this position gave *au* in Gothic and *o* in the other languages. In other positions (*e.g.*, when followed by a nasal sound plus a consonant) **o* gave *u* in all the languages: Proto-Germanic **dumbaz*, Gothic *dumbs*, Old Norse *dumbr*, Old English, Old Frisian, and Old Saxon *dumb*, Old High German *tumb* "dumb." What may be deduced is that this vowel sounded more like *u* in some environments, but like *o* in others; it may be written as **u~o*, indicating that it varied between these two pronunciations.

The above example shows that such reconstructions are more than mere formulas of relationship; they also give some indication of how Proto-Germanic actually sounded. Occasionally scholars are fortunate enough to find external confirmation of these deductions. For example, on the basis of Old English *cyning*, Old Saxon and Old High German *kuning* "king," the Proto-Germanic **kuningaz* can be reconstructed; this would seem to be confirmed by Finnish *kuningas* "king," which must have been borrowed from Germanic at a very early date.

By pushing the comparative method still farther back, it can be shown that Germanic is related to a number of other languages, notably Celtic, Italic, Greek, Baltic, Slavic, Iranian, and Indo-Aryan (Indic). All of these language groups are subsequent developments of a still earlier parent language for which there are, again, no written records but which can be reconstructed as Proto-Indo-European (see also INDO-EUROPEAN LANGUAGES).

This article is divided into the following sections:

I. Characteristics of the Germanic languages
  Phonology
  Grammar
  Branches of Germanic
II. East Germanic
  History
  Characteristics
III. West Germanic
  English
  Frisian
  Netherlandic (Dutch-Flemish)
  Afrikaans
  German
  Yiddish
IV. North Germanic (the Scandinavian languages)
  History
  Characteristics

I. Characteristics of the Germanic languages

The special characteristics of the Germanic languages that distinguish them from other Indo-European languages result from numerous changes, both phonological and grammatical.

PHONOLOGY

Consonants. Proto-Indo-European had 12 stop consonants: *p, t, k, k*^{w}; *b, d, g, g*^{w}; *bh, dh, gh, g*^{w}*h;* and one sibilant, *s*. (Stops are produced with momentary complete stoppage of the breathstream at some point in the vocal tract.) By a change known as the Germanic consonant shift (or Grimm's law, after the German scholar Jacob Grimm, who was one of the first to describe it), the 12 stops changed in Germanic to voiceless fricatives, voiceless stops, and voiced fricatives, as illustrated in Table 3. A few examples: (1) Proto-Indo-European *p, t, k,* and *k*^{w}, as in Latin *piscis, tenuis, centum,* and *quod*, became Proto-Germanic *f, þ, x,* and *x*^{w}, as in English "fish," "thin," "hund(red)," and "what." Proto-Germanic *x* and *x*^{w} early became *h* and *h*^{w} in some positions, giving the Grimm's law

Table 3: Sound Changes in the Germanic Consonant Shift

Proto-Indo-European voiceless stops	p	t	k	k^w
Proto-Germanic voiceless fricatives	f	þ	x	x^w
Proto-Indo-European voiced stops	b	d	g	g^w
Proto-Germanic voiceless stops	p	t	k	k^w
Proto-Indo-European voiced aspirated stops	bh	dh	gh	g^wh
Proto-Germanic voiced fricatives	ƀ	ð	g	g^w

alternations of *h*∼*x* and h^w∼x^w. (2) Proto-Indo-European *d* and *g*, as in Latin *decem* and *genus*, became Proto-Germanic *t* and *k*, as in English "ten" and "kin." (3) Proto-Indo-European *bh*, *dh*, and *gh*, as in Sanskrit *bhū-*, *dhā-*, and *(g)hā-*, became Proto-Germanic *ƀ*, *ð*, and *g*, which later changed to the stops *b*, *d*, and *g* in some positions (*e.g.*, English "be," "do," and "go"), giving *b*∼*ƀ*, *d*∼*ð*, and *g*∼*g*. Proto-Indo-European *s*, as in Latin *sedeō*, was unchanged; Proto-Germanic kept *s*, as in English "sit."

In addition to these general changes, there were two special ones. (1) Proto-Indo-European *p*, *t*, and *k* remained voiceless stops when preceded by *s* or another stop; thus, Proto-Indo-European *sp*, *st*, *sk*, *pt*, and *kt* gave Proto-Germanic *sp*, *st*, *sk*, *ft*, and *xt*, respectively. For example, Proto-Indo-European *sp* and *st*, as in Latin *spuō* and *hostis*, remained *sp* and *st* in Proto-Germanic, as in English "spew" and "guest"; Proto-Indo-European *pt* and *kt*, as in Latin *captus* and *octō*, became Proto-Germanic *ft* and *xt*, respectively, in Old English *hæft* "captured" and *eahta* "eight." (By still another change, Proto-Indo-European *tt* gave Proto-Germanic *ss; e.g.*, Sanskrit *sattá-*, Old English *sess* "seat.") (2) By a change known as Verner's law (named for the Danish scholar Karl Verner, who first described it), early Germanic voiceless *f*, *þ*, *x*, x^w, and *s* (from Proto-Indo-European *p*, *t*, *k*, k^w, and *s*) were voiced to *ƀ*, *ð*, *g*, g^w, and *z*, respectively, when they followed an unaccented syllable, and the first four of these thereby merged with the already existing *ƀ*, *ð*, *g*, and g^w (from Proto-Indo-European *bh*, *dh*, *gh*, and g^wh). Thus, Proto-Indo-European **bhrā́tēr* became Proto-Germanic **brōþēr* (with *þ* after an accented syllable) and Old English *brōþor* "brother"; but by Verner's law Proto-Indo-European **mātḗr* became Proto-Germanic **mōðēr* (with *ð* after an unaccented syllable) and Old English *mōdor* "mother." (The *th* of modern English "mother" is the result of a subsequent change.)

Verner's law

These changes gave the following Proto-Germanic system of consonants: voiceless stops and fricatives, *p*, *f*, *t*, *þ*, *k*, *h*∼*x*, k^w, h^w∼x^w; voiced stops and fricatives, *b*∼*ƀ*, *d*∼*ð*, *g*∼*g*,(g^w∼g^w); sibilants, *s*, *z;* nasals, *m*, *n;* liquids, *l*, *r;* and semivowels, *w*, *j*(*y*). The sound alternation of g^w∼g^w is parenthesized because it early became either *g*∼*g* or *w*. The sounds k^w and h^w∼x^w occurred as such more or less clearly only in Gothic; elsewhere they became the sequences *kw* and *hw*∼*xw*, or the labial element ^{w} was lost. All remaining consonants except *z* occurred between vowels both singly and doubly (*e.g.*, *-p-* and *-pp-*, *-t-* and *-tt-*).

Vowels. In addition to the above consonants (12 stops and the sibilant *s*), Proto-Indo-European also had vowels and resonants. The vowel of any given root was not necessarily fixed but varied in an alternation called ablaut. Thus, the root that means "sit" was alternately **sed-*, *sod-*, **sd-*, **sēd-*, and **sōd-* (English "sit" is from **sed-*, "sat" from **sod-*, and "seat" from **sēd-*); and the root that means "do" was **dhē*, **dhō-*, and **dhə-* (English "deed" is from **dhē-*, and "do" is from **dhō-*). Other Proto-Indo-European vowels were *a*, *ā*, *ī*, and *ū*. The Proto-Indo-European resonants, which functioned as vowels in some positions and as consonants in others, were *i*, *u*, *m*, *n*, *l*, and *r*. Thus, **bhṛtó-* (Sanskrit *bhṛtá-* "borne") had syllabic *ṛ* (*i.e.*, *ṛ* functioning as a vowel), but **bhéreti* (Sanskrit *bhárati* "he bears") had nonsyllabic *r* (*i.e.*, *r* functioning as a consonant).

Ablaut

This Proto-Indo-European system of vowels contrasting with resonants was reshaped in Germanic by a number of changes. Syllabic *i*, *u*, *ṃ*, *ṇ*, *l̥*, and *ṛ* became in Proto-Germanic the vowels *i* and *u* and the sequences *um*, *un*, *ul*, and *ur*, respectively; nonsyllabic *m*, *n*, *l*, and *r* developed into the nasals and liquids *m*, *n*, *l*, and *r*, respectively; nonsyllabic *i* and *u* before vowels resulted in the semivowels *j* (also symbolized as *y*) and *w*, though after vowels they continued to form diphthongs (*ei*, *ai*, *oi; eu*, *au*, *ou*). The Proto-Indo-European vowels and diphthongs then changed into Proto-Germanic sounds as follows:



In this diagram the lines between two sounds indicate that the Proto-Indo-European sound developed into the corresponding Proto-Germanic sound; for example, Proto-Indo-European *i* became either *i* or *e*, and Proto-Indo-European *ə*, *a*, and *o* coalesced in Proto-Germanic as *a*. These changes gave the following vowels for early Proto-Germanic: short vowels, *i*, *e*, *a*, *u*∼*o;* long vowels, *ī*, $ē^1$, *ū*, *ō;* diphthongs, *ai*, *au*, *iu*∼*eo*. The vowel *ē* is noted here as $ē^1$ because Proto-Germanic had (or developed) a second $ē^2$ of uncertain origin. In Gothic the two *ē*'s merged. Elsewhere $ē^2$ remained a midvowel, but $ē^1$ was lowered; thus, for example, $ē^2$ in Old Saxon *hēr* "here" but lowered $ē^1$ in Old Saxon *dād* "deed." In addition to the above oral vowels, Proto-Germanic also had three nasalized vowels: long *ĩ*, *ã*, and *ũ*, which arose when, in the sequences *inx*, *anx*, and *unx*, the *n* was lost with nasalization and lengthening of the preceding vowel.

Accent. Proto-Indo-European had a variable pitch accent that could fall on any syllable of a word (*e.g.*, on the first syllable in **bhrā́tēr* "brother" but on the last syllable in **mātḗr* "mother." This was replaced in Germanic by a fixed stress accent that always fell on the first syllable: **brṓþēr*, **mṓðēr*. One effect of this strong initial stress seems to have been the progressive weakening and loss of unstressed final syllables; *e.g.*, Proto-Indo-European *sodéionom*, Proto-Germanic **sátjanan*, Old English *settan*, Middle English *sette(n)*, and modern English (to) "set." Strong initial stress is also reflected in the basic unit of old Germanic poetry, which consisted of two half lines, each with one of a small number of stress patterns, linked by the alliteration of stressed initial consonants or vowels (*e.g.*, from *Beowulf: Béo-wùlf wæs brḗme / blǣd wīde spràng* "Beowulf was famous; his renown went far").

Strong initial stress

GRAMMAR

Declensions. Proto-Germanic kept the Proto-Indo-European system of three genders (masculine, feminine, neuter) and three numbers (singular, dual, plural), though the dual was becoming obsolete. It reduced the Proto-Indo-European system of eight cases to six: nominative, accusative, dative, genitive, instrumental, and vocative, though the last two were becoming obsolete. In the adjective declensions there were two innovations: (1) To the Proto-Indo-European vowel types (*o-*, *ā-*, *i-*, and *u*-stems), it added some pronominal endings to give the Germanic "strong" adjective declension. (2) It extended the Proto-Indo-European *n*-stem endings to all adjectives to give the Germanic "weak" adjective declension. Contrast, in modern German, strong *gutes Bier* "good beer" with weak *das gute Bier* "the good beer."

Conjugations. The Proto-Indo-European verb seems to have had five moods (indicative, imperative, subjunctive, injunctive, and optative), two voices (active and mediopassive), three persons (1st, 2nd, and 3rd), three numbers (singular, dual, and plural), and several verbal nouns (infinitives) and adjectives (participles). In Germanic these were reduced to indicative, imperative, and subjunctive moods; a full active voice plus passive found only in Gothic; three persons; full singular and plural forms and dual forms found only in Gothic; and one infinitive (present) and two participles (present and past). The Proto-Indo-European tense-aspect system (present, imperfect, aorist, perfect) was reshaped to a single tense contrast between present and past. The past showed two innovations: (1) In the "strong" verb Germanic transformed Proto-Indo-European ablaut into a specific tense marker (*e.g.*, Proto-Indo-European **bher-*, **bhor-*,

Strong and weak verbs

**bhēr-*, **bhṛ-* in Old English *beran* "bear," past singular *bær*, past plural *bāron*, past participle *boren*). (2) In the "weak" verb Germanic developed a new type of past and past participle (*e.g.*, Old English *fyllan* "fill," past *fylde*, participle *gefylled*). Weak verbs fell into three classes depending on the syllable following the root (*e.g.*, Old High German *full-e-n* [from **full-ja-n*] "fill," *mahh-ō-n* "make," *sag-ē-n* "say"). Gothic also had a fourth class: *full-nō-da* "it became full."

Many Proto-Germanic strong verbs showed a consonant alternation between *f* and *ƀ*, *þ* and *ð*, *x* and *g*, and *s* and *z* that was the result, through Verner's law, of the alternating position of the Proto-Indo-European accent. The forms in Table 4 illustrate changes resulting from Verner's law. In this particular word, English has generalized the *s* (now *z*): "freeze," "froze," "frozen." German has generalized the *z* (now *r*): *frieren, fror, gefroren.* And Netherlandic still shows the alternation: *vriezen, vroor, gevroren.* English has kept the alternation in only one verb: singular "was," plural "were." Traces of it still survive, however, in a few now isolated forms: "seethe" (Proto-Germanic *þ*) and its old past participle "sodden" (Proto-Germanic *ð*); "lose" (Proto-Germanic *s*) and its old past participle "(for)lorn" (Proto-Germanic *z*).

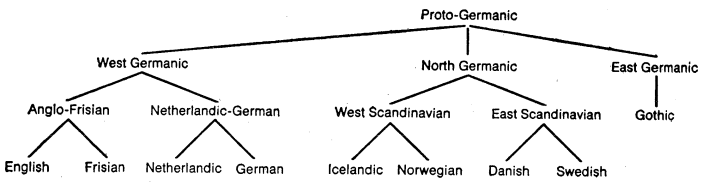
Table 4: Illustration of Verner's Law

Proto-Indo-European	Proto-Germanic	Old English	English translation
*préusonom	*freosanan	frēosan	"(to) freeze"
*próuse	*fraus	frēas	"(it) froze"
*prusṇt	*fruzun	fruron	"(they) froze"
*prusénos	*frozenaz	froren	"frozen"

*Unattested, reconstructed form.

BRANCHES OF GERMANIC

Like every language spoken over a considerable geographic area, Proto-Germanic presumably consisted of a number of geographical varieties or dialects, which, in the course of time, developed in different ways to give the different early and modern Germanic languages. Late-19th-century scholars used a family tree diagram to show this splitting into dialects and the relationships among the dialects:



Though there is much truth in such a diagram, it overemphasizes the notion of "splits" into separate "branches" and obscures the fact that the transition from one dialect to another may be gradual rather than abrupt. Modern Netherlandic and German, for example, constitute a single speech area at the level of local dialects; they have "split" only in the sense that they have developed different standard languages.

Dialect similarities

Mid-20th-century scholars, using the findings of archaeology and the methods of geographical linguistics, attempted to correct the distortions of this family-tree model by noting also the linguistic features shared by two or more dialect areas. Archaeological evidence suggests that a relatively uniform Germanic people at *c.* 750 BC were located in southern Scandinavia and along the North Sea and Baltic coasts from The Netherlands to the Vistula. Five hundred years later (*c.* 250 BC) they had spread south, and five general groups are distinguishable: North Germanic in southern Scandinavia, excluding Jutland; North Sea Germanic, along the North Sea and in Jutland; Rhine-Weser Germanic, along the middle Rhine and Weser; Elbe Germanic, along the middle Elbe; and East Germanic, between the middle Oder and the Vistula.

By *c.* AD 250 the division was much the same, though the Elbe group had spread southward to the Danube, and the East Germanic group moved southeast into the Carpathians and beyond. Then, toward the end of the 4th century, began the great Germanic tribal migrations. North Germanic speakers migrated into Jutland, approximately to the modern Danish–German language border; part of the North Sea group crossed the North Sea and conquered much of England; the Elbe group (the later Alamanni, Bavarians, and Langobardi) spread still farther south into part of Switzerland and into Austria and northern Italy; and the East Germanic group left the Oder-Vistula area to begin their many wanderings.

This five-way division of the Germanic peoples is based on archaeological evidence, but it agrees well with deductions that can be made from the earliest linguistic evidence. Five linguistic groups are indeed distinguishable, though they are linked into sets of two, three, or four through shared linguistic innovations.

Linguistic groups

1. North Germanic, North Sea Germanic, Rhine-Weser Germanic, and Elbe Germanic share the change of *z* to *r; e.g.*, Proto-Germanic **maiz-* "more," Gothic *maiza*, contrasting with Old Norse *meire*, Old English and Old Frisian *māra*, Old Saxon *mêro*, and Old High German *mēro*. In addition, they also show *i*-umlaut, as in the raising of *a* to *e* before *j* (pronounced as the *y* in "year"); *e.g.*, Proto-Germanic **satjanan* "set," Gothic *satjan* in contrast to Old Norse *setia*, Old English *settan*, Old Frisian *setta*, Old Saxon *settian*, and Old High German *setzen*. In certain strong verbs they share a new past tense form with $\bar{e}^2$ and without reduplication (the repetition of a part of a word)—*e.g.*, Proto-Germanic **le-lōt* "let," Gothic *lailot* in contrast to Old Norse, Old English, Old Frisian, and Old Saxon *lēt* and Old High German *liez*.

2. North Germanic, North Sea Germanic, and Rhine-Weser Germanic partly share the loss of nasal sounds before voiceless fricative sounds. As noted above, *n* was lost in Proto-Germanic before *x*. North Sea Germanic shows loss of nasals before the remaining fricatives *f*, *þ*, and *s;* thus, the nasals in Proto-Germanic **fimf* "five," **munþ* "mouth," and **ũns* "us" are preserved in Gothic *fimf*, *munþs*, and *uns*, as well as in Old High German *fimf*, *mund*, and *uns*, but are lost in Old English *fīf*, *mūþ*, and *ūs* and Old Frisian *fīf*, *mūth*, and *ūs*. Rhine-Weser Germanic shows this same loss only sporadically; Old Saxon has *fīf*, *mūð*, and *ūs*, without the nasals, but also has *andar* (from Proto-Germanic **anþar-* "other"), with the nasal consonant, beside *āðar* and *ōðar*, without it. Old Norse, which is North Germanic, shows regular loss of a nasal sound only before *s* (*e.g.*, *oss* "us").

3. North Sea Germanic, Rhine-Weser Germanic, and Elbe Germanic (usually grouped together as West Germanic) share the change of *ð* to *d* in all positions (*e.g.*, Proto-Germanic **blōð-* "blood," and Gothic and Old Norse *blōð-* in contrast to Old English, Old Frisian, and Old Saxon *blōd* and Old High German *bluot*), the loss of *-z* after unstressed vowels (*e.g.*, Proto-Germanic **dagaz* "day," Gothic *dags*, and Old Norse *dagr* in contrast to Old English *dæg*, Old Frisian *dei*, Old Saxon *dag*, and Old High German *tag*), and a 2nd-person-singular past formation in strong verbs different from that of East Germanic and North Germanic (*e.g.*, Proto-Germanic **gaft* "[thou] gavest" occurs in Gothic and Old Norse *gaft*, but Proto-Germanic **gē¹þi* appears in Old English *gēafe*, Old Saxon *gabi*, and Old High German *gābi;* Old Frisian has a new analogical form, *iefst*).

In addition, they share the doubling of most consonants in certain positions, especially before *j* (the *y* sound); *e.g.*, Proto-Germanic **satjanan* "set" appears in Gothic as *satjan* and in Old Norse as *setia* but in Old English as *settan*, in Old Frisian as *setta*, in Old Saxon as *settian*, and in Old High German as *setzen*. North Germanic also shows doubling of *g* and *k* before *j* (*y*); *e.g.*, Proto-Germanic **lagjanan* "lay" becomes Gothic *lagjan* but Old Norse *leggia*, Old English *lecgan*, Old Frisian *ledza*, Old Saxon *leggian*, and Old High German *lecken*.

4. North Sea and Rhine-Weser Germanic share a single ending for the 1st, 2nd, and 3rd persons plural of verbs (North Germanic, Elbe Germanic, and East Germanic

show two or three different endings), loss of the Proto-Germanic reflexive pronoun **sik*, and loss of *-z* in pronouns (*e.g.*, Proto-Germanic **wīz* or **wiz* "we," which appears in Gothic as *weis*, in Old Norse as *vēr*, and in Old High German as *wir*, appears in Old English as *wē* and in Old Frisian and Old Saxon as *wī*).

5. North Germanic, Elbe Germanic, and East Germanic share the addition of the Proto-Germanic nominative-accusative neuter singular pronominal ending **-at* in the strong adjective declension (*e.g.*, Proto-Germanic **hailan* "whole," Old English and Old Frisian *hāl*, and Old Saxon *hêl* in contrast to Gothic *heilata*, Old Norse *heilt*, and Old High German *heilaz*).

6. Elbe Germanic and East Germanic share the pronoun reconstructed for Proto-Germanic as **iz* "he": Gothic *is*, Old High German *ir* or *er*, instead of Proto-Germanic **h-* occurring in Old Norse *hann* and in Old English, Old Frisian, and Old Saxon *hē*.

7. North Germanic and East Germanic share the change of *jj* (pronounced as English *yy*) and *ww* to a long stop plus a semivowel (*e.g.*, from Proto-Germanic **twajj-* "of two" and *triww-* "true"—as in Old High German *zweiio* and *triuwi*—to Old Norse *tueggia* and *tryggr*, Gothic *twaddje* and *triggws*).

II. East Germanic

The East Germanic languages, all of which have long been extinct, developed from the dialects of the East Germanic group mentioned above; they were spoken by Germanic tribes located between the middle Oder and the Vistula.

HISTORY

According to historical tradition, at least some of the Germanic tribes migrated to the mouth of the Vistula from Scandinavia. Little is known of Gepidic, Rugian, and Burgundian; some knowledge of Vandalic, Visigothic, and, especially, Ostrogothic is provided by the names recorded in Greek and Latin writings. The only East Germanic language on which there is extensive information is the Gothic—more specifically, Visigothic—that was spoken along the western shore of the Black Sea around the middle of the 4th century AD. Its special importance lies in the fact that, except for a few scattered runic inscriptions, it is by far the oldest Germanic language preserved.

Gothic language

Knowledge of Gothic is derived primarily from the remains of a Bible translation made for the Visigoths living along the lower Danube by a Visigothic bishop of the Arian church named Ulfilas, who lived during the 4th century. The surviving manuscripts of this translation, which are not originals but later copies thought to have been written in northern Italy during the period of Ostrogothic rule (493–554), include considerable portions of the New Testament and parts of Nehemiah from the Old Testament. Although most of them are palimpsests, manuscripts in which earlier erased writings are found, a handsome exception is the famous Codex Argenteus, written in silver and gold letters on purple parchment and containing (in 188 leaves remaining from an original 330 or 336) portions of the four gospels. Closely related to these biblical manuscripts are eight leaves containing fragments of a commentary (called the

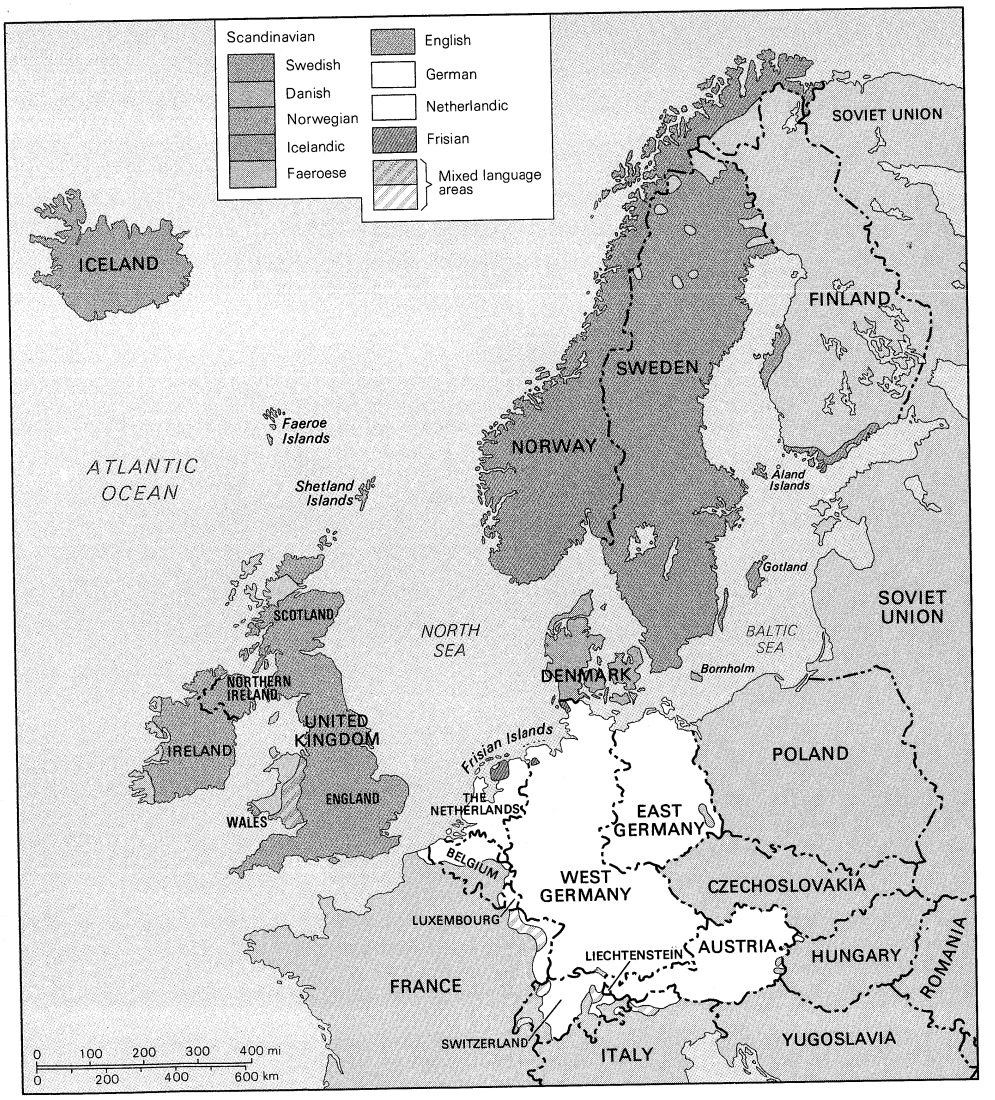


Figure 1: Distribution of the Germanic languages in Europe.

Skeireins in Gothic) on the Gospel According to St. John. Minor nonbiblical texts include a fragment of a calendar, two deeds containing some Gothic sentences, and a 10th-century Salzburg manuscript that gives the Gothic alphabet, a few Gothic words with Latin transliteration, and some phonetic remarks with illustrative examples.

In the 4th and 5th centuries, Gothic (Visigothic and Ostrogothic) must have spread, along with the conquering Goths, at least thinly over much of southern Europe; but there is no evidence for its survival in Italy after the fall of the Ostrogothic kingdom, and in Spain it is doubtful whether the Visigoths retained their language until the Arab conquest. In the 9th century the German monk Walahfrid Strabo mentions that Gothic was still being used in some churches near the lower Danube. After that time Gothic seems to have survived only among the Goths of the Crimea, who were last mentioned in the middle of the 16th century by a Flemish diplomat named De Busbecq, who, while on a mission to Constantinople in 1560–62, collected a number of words and phrases showing that their language was still essentially a form of Gothic.

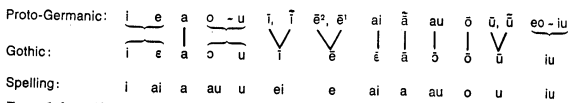
CHARACTERISTICS

Gothic alphabet

The Gothic alphabet, said to have been created by Ulfilas, contained 27 symbols, two of which functioned only as numbers, while the remaining 25 were used as both numbers and letters. The shape, numerical value, and ordering of the symbols show clearly that the alphabet was based primarily on that of Greek, though a few symbols seem to have been adapted from the Latin alphabet.

Phonology. *Consonants.* The Gothic consonant system seems to have been largely identical with that assumed above for Proto-Germanic: *p, t, k, k*^{w} (this last sound was probably much like the *qu* in "queen"); *f, þ, h, h*^{w} (this last sound was probably pronounced much like the *wh* in "white"); *b, d, g; s, z; m, n; l, r; w, j.* The nasal *n* was presumably velar before the velar consonants *k, q,* and *g;* in these positions it was usually written (as in Greek) as *g* or *gg.* Examples of this spelling include *dragk* "drank," *igqis* "you two," and *briggan* "bring," although *n* was occasionally used as in Latin (*e.g., þank* "thanks," *inqis* "you two," and *bringiþ* "bring ye").

Vowels. The Gothic alphabet contained the five simple vowel symbols, *i, e, a, o,* and *u,* from which four compound symbols, *ei, ai, au,* and *iu,* were also made; in addition, *w* was used to transliterate Greek υ and οι (both of which were pronounced as umlauted *u* [ü] in 4th-century Greek). The generally accepted development of the Proto-Germanic vowels in Gothic can be diagrammed as follows:



In this diagram straight lines indicate that the Proto-Germanic sound developed into the Gothic sound below. Brackets in the Proto-Germanic line indicate that the two linked sounds coalesced into one; brackets in the Gothic line indicate two variants of the same sound that are in different phonetic environments. Proto-Germanic *i* and *e* apparently first merged as a single vowel and then became Gothic *i* in most positions, but became *ai* before *h, h*^{w}, and *r.* Similarly, Proto-Germanic **o~u* became Gothic *u* in most positions, but *au* before *h, h*^{w}, and *r.*

Special characteristics. Gothic shows a number of archaic features that had been almost or entirely lost by the time the other Germanic languages began to appear in writing; among these are a passive voice and one type of past tense formed with reduplication, a dual number in the 1st and 2nd persons of its verbs and pronouns, and a special vocative case in one noun class. At the same time, Gothic also shows changes from Proto-Germanic, among which are the shortening of most long vowels in final unstressed syllables and the loss of most short vowels (*e.g.,* Proto-Germanic **erþō* "earth" became Gothic *airþa,* Proto-Germanic **stainaz* "stone" became Gothic *stains*). Finally, voiced fricatives that occurred or came to occur at the end of a word have been unvoiced (*e.g.,* nominative **hlaiƀaz,* accusative **hlaiƀan* "bread, loaf" changed to *hlaifs* and *hlaif,* respectively [but dative *hlaiba*]).

III. West Germanic

The West Germanic languages are those that developed from the North Sea, Rhine-Weser, and Elbe groups mentioned above. Out of the many local West Germanic dialects the following six modern standard languages have arisen: English, Frisian, Netherlandic (Dutch-Flemish), Afrikaans, German, and Yiddish.

ENGLISH

English and Frisian changes

English and Frisian are descended from North Sea Germanic. The most striking changes that distinguish them from the other Germanic languages are the loss of nasal sounds before the Proto-Germanic voiceless fricatives *f, þ,* and *s* (contrast the following pairs of words, in which English loses the nasal but German preserves it: before *f*—"soft"/*sanft;* before *þ*—"other"/*ander;* before *s*—"us"/*uns,* "goose"/*Gans*); palatalization of Proto-Germanic *k* before front vowels and *j,* giving modern English *ch* (English/German pairs: "chin"/*Kinn,* "birch" [Old English *birce*]/*Birke*); and palatalization of Proto-Germanic *g* before front vowels, giving modern English *y* (English/German pairs include "yield"/*gelten,* "yester-[day]"/*gestern,* "yard" [Old English *geard*]/*Garten;* this palatalized *g* merged with the *j* [*y* sound] from Proto-Germanic *j:* "year"/*Jahr*).

Other changes include palatalization of *gg* before *j* to Old English *cg* (Proto-Germanic **brugjō,* pre-Old English **bruggju,* Old English *brycg* "bridge"; contrast the unpalatalized *ck* from *gg* of German *Brücke* "bridge"); fronting and raising of *ā* from Proto-Germanic *ē*^{1} (English/German pairs include "deed"/*Tat,* "seed"/*Saat,* "sleep"/*schlafen,* "meal"/*Mahl*); and backing and raising of nasalized *ã,* from Proto-Germanic *ā* and from Proto-Germanic *a* before nasal plus *f, þ,* and *s* (English/German pairs include "brought"/*brachte,* "thought"/*dachte,* "other"/*ander,* and "goose"/*Gans*).

For further information on English, see the article ENGLISH LANGUAGE.

FRISIAN

A thousand years or so ago, Frisian was apparently spoken throughout a North Sea coastal area extending from the modern Netherlands province of Noord-Holland (North Holland) on up to modern German Schleswig and the adjacent offshore islands. During the following centuries, the Frisian of much of this area was gradually replaced by local Netherlandic and Low German dialects, so that Modern Frisian is now spoken in only three remaining areas: (1) West Frisian, in the Netherlandic province of Friesland, including the island of Schiermonnikoog and two-thirds of the island of Terschelling (altogether some 300,000 speakers); (2) East Frisian, in the German Saterland (some 1,000 speakers; this area was apparently settled in the 12th or 13th centuries from the former East Frisian area to the north); and (3) North Frisian, along the west coast of German Schleswig and on the offshore islands of Sylt, Föhr, Amrum, the Halligen, and Helgoland (altogether some 10,000 speakers).

History. The earliest manuscripts written in Frisian date from the end of the 13th century, though the legal documents that they contain were probably first composed, in part, as early as the 11th century. This stage of the language, until about 1575, is known as Old Frisian. The last written document of this period dates from 1573, after which Frisian was hardly used at all as a written language for some three centuries, though it continued to be spoken.

From the start Old Frisian shows all of the features that distinguish English and Frisian from the other Germanic languages. These include loss of the nasal sound before Proto-Germanic *f, þ,* and *s* (*e.g.,* Proto-Germanic

**fimf*, **munþ-*, and **uns* became Old Frisian *fīf* "five," *mūth* "mouth," and *ūs* "us"), palatalization of Proto-Germanic *k* before front vowels and *j* (*e.g.*, Proto-Germanic **kinn-* and **lē¹kj-* became Old Frisian *tzin* "chin" and *lētza* "physician" [*cf.* English archaic "leech"]), and palatalization of Proto-Germanic *g* before front vowels (*e.g.*, Proto-Germanic **geldan-* became Old Frisian *ielda* "yield"). This merged with the *j* from Proto-Germanic *j*, as in Proto-Germanic **jē¹r-* or Old Frisian *iēr* "year." In addition, Old Frisian shows palatalization of *gg* from Proto-Germanic *g* before *j* (*e.g.*, Proto-Germanic **lagjan-*, with doubling **laggjan*, became Old Frisian *ledza* "to lay"); fronting and raising of *ā* from Proto-Germanic *ē¹*, as in Proto-Germanic **dē¹ð-*, lowered to **dād*, and raised again to Old Frisian *dēd* "deed"; and backing and raising of nasalized *ã* from Proto-Germanic *ã* and Proto-Germanic *a* before nasal plus *f*, *þ*, *s*, as in Proto-Germanic **brãxt-*, **anþar-*, and **gans-*, which became Old Frisian *brocht* "brought," *ōther* "other," and *gōs* "goose."

Around the beginning of the 19th century it appeared that the age-old replacement of Frisian by Netherlandic and German would continue unabated and that the language would soon become extinct. But with 19th-century Romanticism a new interest in local life arose, and societies were formed for the preservation of the Frisian language and culture. Very slowly, the aims of this "Frisian movement" came to be realized, especially in the Netherlands province of Friesland, where in 1937 Frisian was accepted as an optional course in elementary schools; a Frisian Academy was founded in 1938; and in 1943 the first Frisian translation of the Bible was published. Later, in 1955, Frisian was approved as the language of instruction in the first two years of elementary school (though only about one-fourth of all schools use it in this way), and in 1956 the use of Frisian in courts of law was approved.

Status of Frisian

Despite this gradual re-emergence of Frisian, Netherlandic still functions as the primary standard language of Friesland. Nearly all school instruction is given in Netherlandic; all daily newspapers are printed in Netherlandic (though they contain occasional articles in Frisian); and all television broadcasts and nearly all radio broadcasts are in Netherlandic. There is a small and enthusiastic Frisian literary movement, but its works are not widely read. Furthermore, though Frisian continues to be widely used as the language of everyday oral communication, it is increasingly a "Netherlandic" Frisian, with numerous borrowings from standard Netherlandic.

The status of Frisian in the East and North Frisian areas of Germany is far more tenuous. There German performs all the functions of a standard language, and Frisian serves only as yet another local dialect, comparable to the many surrounding local dialects of Low German. No standard North Frisian or East Frisian exists.

Characteristics. The following remarks refer to the more or less standard West Frisian that is developing in the province of Friesland.

Consonants. Frisian has the following system of consonants, given here in the usual spellings: stops, *p*, *b*, *t*, *d*, *k*, *g*; fricatives, *f*, *v*, *s*, *z*, *ch*, *g*; nasals, *m*, *n*, *ng*; liquids, *l*, *r*; and glides, *w*, *h*, *j*. Examples (given here in part to show the close relationship between Frisian and English) include *p*, *t*, and *k* (unaspirated) in *peal* "pole," *twa* "two," and *kat* "cat"; *b*, *d*, and the stop symbolized by the letter *g* in *boi* "boy," *dei* "day," and *goed* "good"; *f*, *s*, and *ch* in *fiif* "five," *seis* "six," and *acht* "eight"; *v*, *z*, and the fricative symbolized by the letter *g* in *tolve* "twelve," *tûzen* "thousand," and *wegen* "ways"; *m*, *n*, and *ng* in *miel* "meal," *need* "need," and *ring* "ring"; *l* and *r* in *laem* "lamb" and *reep* "rope"; *w*, *h*, and *j* in *wy* "we," *hy* "he," and *jo* "you." Word-finally, voiced *b*, *d*, *z*, and *g* are generally unvoiced to *p*, *t*, *s*, and *ch*.

Vowels. Frisian has the following system of stressed vowels and diphthongs. The symbols given in Table 5 refer to the actual sounds rather than to Frisian spellings, which are often irregular. Frisian also has an unstressed vowel *ə* (pronounced as the *a* in English "sofa"), which occurs only in unstressed syllables.

Table 5: The Vowel System of Frisian

short vowels	long vowels	rising diphthongs	falling diphthongs
i ü u	ī ǖ ū		ie üö uo
e ö o	ē ō ō	öi	eɛ öə oa
ɛ ɔ	ɛ̄ ɔ̄	ei ou	
a	ā	ai	

Dialects. The Frisian dialects of The Netherlands province of Friesland are, with three exceptions, relatively uniform, though it is customary to make a distinction between Wouden Frisian in the east, Klei Frisian in the west (the variety on which standard Frisian is largely based), and Southwest Corner Frisian in the southwest. The three exceptions are the island dialect of East and West Terschelling and the dialects of the city of Hindeloopen and of the island of Schiermonnikoog. These latter two differ so greatly that they are not intelligible to other speakers of West Frisian and are both dying out. Quite different from any of these is the so-called City Frisian (Stedfrysk, or Stedsk) spoken in the cities of Leeuwarden, Franeker, Harlingen, Bolsward, Sneek, Staveren, and Dokkum. Despite the name, this is not Frisian at all but a variety of Netherlandic strongly influenced by Frisian. Similar in nature are the dialects of Heerenveen and Kollum, of the middle section of the island of Terschelling, and of Het Bildt (a coastal area northwest of Leeuwarden, diked in and settled by Hollanders during the 16th century).

Dialects of Friesland

East Frisian survives today only in the German Saterland, consisting of the three parishes of Ramsloh, Strücklingen, and Scharrel, each with a slightly different dialect. The area to the north is called East Frisia (German Ostfriesland), and the local dialect East Frisian (German Ostfriesisch), although it is really not Frisian but the local variety of Low German.

Though North Frisian is spoken in only a small geographical area by only some 10,000 persons, it exists in an extraordinary number of local dialects, some of which are mutually unintelligible. Because of this, it would be almost impossible to develop a single standard North Frisian that could be used throughout this area. North Frisian dialects are customarily divided into Insular North Frisian (Sylt, Föhr-Amrum, Helgoland) and Continental North Frisian (the Halligen Islands and the coast of Schleswig), the latter in seven main varieties and further subvarieties. Because this whole area bordered until recently on Danish, it was extensively influenced by the neighbouring Danish dialects. In more recent times it has been heavily influenced by German, both standard German and the neighbouring Low German dialects. Today all speakers of North Frisian are probably bilingual or trilingual; all of them learn Frisian at home and standard German in school, and many also learn dialectal Low German.

NETHERLANDIC (DUTCH–FLEMISH)

Netherlandic is the national language of The Netherlands and one of the two national languages (beside French) of Belgium. Popular English usage applies the term Dutch to the Netherlandic of Holland and the term Flemish to the Netherlandic of Belgium, but in fact they are one and the same standard language. In its various forms, standard and dialectal, Netherlandic is the indigenous language of most of The Netherlands (all but the Frisian-speaking province of Friesland), of northern Belgium, and of a small part of France immediately to the west of Belgium. It is also used as the language of administration in the former Dutch colonies of Surinam and the Netherlands Antilles. A derivative of Netherlandic, Afrikaans, is one of the two national languages (with English) of the Republic of South Africa.

Identity of "Dutch" and "Flemish"

As a written language, Netherlandic is quite uniform; it differs in Holland and Belgium no more than written English does in the United States and Great Britain. As a spoken language, however, it exists in far more varieties than does the English of North America. At one extreme is Standard Netherlandic (Algemeen Beschaafd

Nederlands, "General Cultured Netherlandic"), which is used for public and official purposes and is the language of instruction in schools and universities. It is everywhere quite uniform, though speakers usually show by their accents the general area from which they come. At the other extreme are the local dialects, used among family and friends and with others from the same village.

History. Netherlandic is descended primarily from the language of the Rhine-Weser group, especially from the language of the Franks who entered much of this area during the 4th and 5th centuries AD. At the same time, it shows many forms descended from the speech of the North Sea Germanic inhabitants of the coastal areas. For example, modern *vijf* "five" (Proto-Germanic **fimf*) shows the typical North Sea Germanic loss of a nasal sound before *f*. Modern *mond* "mouth" (Proto-Germanic **munþ-*) and *ons* "us" (Proto-Germanic **uns*), on the other hand, show preservation of *n* before *þ* and *s;* but loss of *n* before *þ* appears in such place names as *IJmuiden* "mouth of the river IJ," and loss of *n* before *s* appears in the widespread dialectal forms *us* and *os* "us."

Documents written in Netherlandic do not begin to appear until toward the end of the 12th century, in the rich literature called Middle Dutch or Middle Netherlandic. From the preceding Old Netherlandic period there are only a few glosses, names, and isolated words appearing in Latin documents. Related to Netherlandic, though not ancestral to it, are the copyings—partly running text, partly isolated words—made from a lost manuscript that apparently contained an interlinear translation from Latin into Old Low Franconian of the book of Psalms.

The development of modern Netherlandic is closely tied to the political and economic history of the area. By the middle of the 16th century the speech of Brabant and its leading cities Antwerp and Brussels was well on its way to becoming standard for the whole Netherlandic speech area. Then came the revolt against Spain, in which the northern province of Holland was split off from the southern Netherlandic provinces.

Northern and Southern Netherlandic differences

This political split between the United Provinces of the Netherlands in the north and the Spanish Netherlands in the south had far-reaching linguistic consequences. In the prosperous and vigorous north a standard language rapidly developed, based on the speech of the big cities; it also showed the influence of the culturally important refugees from Brabant, who fled to the north, above all to Amsterdam, before and especially after the fall of Antwerp (1585). In the south, French came to prevail among the upper classes. The less privileged classes continued to use dialectal Netherlandic ("Flemish"), but no supradialectal standard was developed.

The cultural predominance of French in the south increased during the period of French rule (1795–1814), abated somewhat during the years when Belgium and Holland were united independently (1815–30), and rose again after the founding of the Kingdom of Belgium in 1830. At that time French was the only official language, used exclusively in government, courts, and schools. The long struggle to give Netherlandic equal status with French ended with the Language Act of 1938, which made Netherlandic the only official language of northern Belgium. There were numerous attempts to set up a standard Flemish different from the Netherlandic of the north, but in the end the standard Netherlandic that had become established in Holland was accepted for northern Belgium as well.

Characteristics. *Consonants.* Modern Standard Netherlandic has the following consonants, given here in the usual spellings: stops, *p, b, t, d, k;* fricatives, *f, v, s, z, ch, g;* nasals, *m, n, ng;* liquids, *l, r;* glides, *w, h, j.*

The voiced stops and fricatives *b, d, v, z,* and *g* are unvoiced to *p, t, f, s,* and *ch,* respectively, in word-final position. The spelling shows this in the case of *v* and *z* (plural *dieven* "thieves," *huizen* "houses," but singular *dief* "thief," *huis* "house") but does not show it in the case of *b, d,* and *g* (plural *ribben* "ribs," *bedden* "beds," *dagen* "days," but singular *rib* "rib," *bed* "bed," *dag* "day," pronounced *rip, bet, dach*).

Vowels. Netherlandic has three classes of vowels and diphthongs: (1) six checked vowels, which are short and always followed by a consonant; (2) ten free vowels and diphthongs, most of them usually long, which need not be followed by a consonant; and (3) a vowel that occurs only in unstressed syllables. They form the system shown in Table 6 (the traditional spelling is to the left, and to the right is a notation, used by some linguists, that indicates the distinctive sounds [phonemes] of the language).

Table 6: Vowel System of Netherlandic

traditional spelling						linguistic notation					
checked			free			checked			free		
			ie	uu	oe				ĭ	ŭ	ŭ
i	u	o	ee	eu	oo	e	ö	o	ē	ō	ō
e		o	ij, ei	ui	ou, au	ɛ		ɔ	ɛi	ɔ̈ü	ɔu
	a			aa			a		ā		
unstressed: e						unstressed: ə					

Netherlandic spelling

The right is a notation, used by some linguists, that indicates the distinctive sounds [phonemes] of the language). Unlike the English spelling system, which in its basic design has remained essentially unchanged since the days of Chaucer (died 1400), the Netherlandic spelling system has undergone a series of official reforms to keep it in line with changes in pronunciation. The principal inconsistencies in the spelling of vowels are the spellings *ij* and *ei*, which both symbolize the same diphthong, pronounced somewhat between the *ai* of English "aisle" and the *ai* of English "maid" (*bijt* "he bites" rhymes with *feit* "fact"), and the spellings *ou* and *au*, which both symbolize the same diphthong, pronounced somewhat between the *ow* of English "now" and the *ow* of English "low" (*bouw* "building" rhymes with *nauw* "narrow"). Free vowels are written with double letters in closed syllables (*vuur* "fire," *boot* "boat"), but with single letters in open syllables (*vuren* "fires," *boten* "boats"). In contrast the checked vowels are always written with single letters.

Dialects. Although the standard language changes abruptly at the political border separating Holland and Belgium from Germany (Netherlandic being used to the west, German to the east), in local dialect speech there is no such abrupt change. The entire Netherlandic–German territory from the North Sea to the Alps is a single dialect area with only gradual transitions from one village to the next.

In an area bounded roughly by Amsterdam, The Hague, and Rotterdam, rural dialects are very similar to Standard Netherlandic; there are marked differences only in urban dialects, especially those of Amsterdam and Rotterdam. As one travels from this area—the source of the standard language—in any direction, however, the difference between local dialects and the standard language becomes progressively greater; as a result, throughout most of Holland rural inhabitants in effect speak two closely related but distinct languages, Standard Netherlandic and a local dialect, in varying degrees of proficiency. Dialects are traditionally named after the provinces in which they are spoken (*e.g.*, Gronings in Groningen and Limburgs in Limburg).

In Netherlandic Belgium the use of Standard Netherlandic is more limited, and that of local dialects is more extensive. Some of the better educated people speak the standard language fluently and use it regularly, while others prefer French. The less well educated use a local dialect almost exclusively and are often able to handle the standard language only with difficulty.

AFRIKAANS

In 1652 a party of Netherlanders under the leadership of Jan van Riebeeck landed at the Cape of Good Hope to establish a station for the Dutch East India Company. In the immediately following years they were joined by a wide variety of other Europeans, in particular Germans and, after 1685, French Huguenots. By 1806, a century and a half after the original settlement, the national origins of the white inhabitants are estimated to have been 53 percent Dutch, 28 percent German, 15 percent French, and 4 percent of other nationalities. Shortly be-

fore this date, in 1795, the Cape Colony came under British control, and British settlers began to arrive around 1820.

Afrikaans–Netherlandic differences

History. From the start, the dialect of the province of Zuid-Holland (South Holland), which was spoken by Van Riebeeck and his large family, seems to have set the style for what was eventually to become modern Afrikaans. As might be expected in a language used by so many non-native speakers (white and black), some simplification of sounds and forms took place. For example, whereas Standard Netherlandic shows three forms in the present tense of most verbs—*ik loop, hij loopt, wij/zij lopen* "I run, he runs, we/they run"—Afrikaans shows only one form—*ek/hy/ons/sy loop* "I/he/we/they run" (with no ending). In addition, whereas Standard Netherlandic uses stressed *die man* "that man" in contrast to unstressed *de man* "the man," Afrikaans has only *die man* 'the man"; and, whereas Standard Netherlandic distinguishes *wij/we* "we" and *ons* "us," Afrikaans has only *ons* "we/us."

The relatively small numbers of white European speakers of Afrikaans borrowed place-names and names of such cultural novelties as African plants and the like from the immensely larger numbers of black speakers of various Bantu languages, with whom they were in intimate contact. For some two centuries the gradually developing Afrikaans language existed only as a spoken dialect, alongside Standard Netherlandic (by which it was constantly influenced) and, later, English. Then, around the middle of the 19th century, the effort to make Afrikaans a medium of literary expression and a standard written language began. It came gradually to be used in newspapers. It was adopted for use in schools in 1914 and was accepted for use in the Dutch Reformed Church in 1919. In 1925 the South African Parliament declared it to be an official language, replacing Netherlandic. The first complete translation of the Bible into Afrikaans was published in 1933. Thus it came to be recognized as one of the two standard languages (beside English) of the modern Republic of South Africa. Though clearly a separate language, Afrikaans is very similar to Netherlandic. A person who knows Netherlandic can read Afrikaans with little difficulty; and, with some practice, he can easily learn to understand it when spoken.

Characteristics. *Consonants.* Afrikaans has the following consonants, given here in the conventional spellings: stops, *p, b, t, d, k, gh/g;* fricatives, *f/v, w, s, z, g;* nasals, *m, n, ng;* liquids, *l, r;* glides, *h, j.* There are numerous differences between Afrikaans and Netherlandic. Netherlandic *-g-* (*-gg-*) is a voiced fricative, but Afrikaans *-g-* (*-gg-*) is a voiced stop. Unlike Netherlandic, Afrikaans also has this voiced stop initially in a few loanwords. Netherlandic has voiced fricatives initially (*v-, z-, g-*); corresponding words have voiceless initial fricatives in Afrikaans. Afrikaans, however, has voiced *z-* in loanwords: *Zoeloe* "Zulu." Netherlandic has initial *s* plus fricative *ch* as in *schoen* "shoe"; corresponding words have *s* plus *k* in Afrikaans: *skoen.* Netherlandic has *-ft, -st,* and *-cht* as in *gift* "poison," *nest* "nest," *nacht* "night"; corresponding words show loss of *-t* in Afrikaans: *gif, nes,* and *nag.*

Vowels. Afrikaans has the system of vowels shown in Table 7 (usual spelling to the left; notation used by linguists to indicate distinctive sounds to the right). As in Netherlandic, *uu, ee, oo,* and *aa* are written with single letters in open syllables, and single consonant letters are doubled in open syllables to show that the preceding vowel is short.

GERMAN

German is spoken throughout a large area in central Europe, where it is the national language of Germany and of Austria and one of the four national languages (beside French, Italian, and Romansh) of Switzerland. From this homeland it has been carried by emigration to many other parts of the world; there are German-speaking communities in North and South America, South Africa, and Australia. In the western world it is extensively used as a second language and in this respect is next in importance (along with French) only to English.

Dispersion of German

As a written language German is quite uniform, differing in Germany, Austria, and Switzerland no more than written English does in the United States and the British Commonwealth. As a spoken language, however, German exists in far more varieties than English. At one extreme is Standard German (Hochsprache), based on the written form of the language and used in radio, television, public lectures, the theatre, schools, and universities. It is relatively uniform, although speakers often show by their accents the areas from which they come. At the other extreme are the local dialects, which differ from village to village. Between these two extremes there is a continuous scale of speech forms that, in cities, are often close to the standard language and are called Colloquial German (Umgangssprache).

History. From the point of view of local dialects the territory within which German and Netherlandic are spoken is a single speech area. It is possible to travel from Austria, northern Italy, and much of Switzerland into Germany, eastern France (Alsace and part of Lorraine), Luxembourg, northern Belgium, and The Netherlands without encountering a village where the local speech is suddenly different. The only sharp breaks occur when one enters the French-speaking parts of France and Belgium or the Frisian-speaking parts of The Netherlands and Germany.

Table 8: Results of the High German Consonant Shift

p-	*p*ound	*Pf*und	*pp*	a*pp*le	A*pf*el	*Vp*†	ho*p*e	ho*ff*en
t-	*t*en	zehn	*tt*	si*tt*ing	si*tz*en	*Vt*†	bi*t*e	bei*ss*en
k-	*c*an	*kh*ann*	*kk*	li*ck*	le*kch*en*	*Vk*†	ma*k*e	ma*ch*en

**Khann* and *lekchen*, with affricates, are southern dialect forms; standard German has stops: *kann, lecken.* †*V* represents any vowel.

The most striking dialect differences within this large area are those that divide Netherlandic-Low German in the lowlands of the north from High German in the highlands of the south. When the Germanic tribes migrated into southern Germany during the early centuries of the Christian era, their speech had the voiceless stops *p, t,* and *k* in much the same distribution as in modern English. Then, probably during the 6th century, there occurred a change customarily called the "High German consonant shift." At the beginning of words and when doubled, *p, t,* and *k* came to be pronounced as affricates; after a vowel they came to be pronounced as long fricatives. The modern results, compared with related English words, are shown in Table 8.

High German consonant shift

These changes occurred in the south of the German speech area and then spread north, some extending far-

Table 7: Vowel System of Afrikaans

usual spelling			linguistic notation		
short vowels	long vowels	diphthongs	short vowels	long vowels	diphthongs
ie uu oe	ie uu oe ee eu oo	(ee eu oo)*	i ü u	ī ṻ ū ē ō̈ ō	(iə üə uə)*
i e u o a	î ê/e û ô/o aa	y/ei ui ou ai	ə ɛ ɔ̈ ɔ a	ə̄ ɛ̄ ɔ̈̄ ɔ̄ ā	əi əü əu ai

*The spellings *ee, eu,* and *oo* are pronounced either as long vowels (ē, ō̈, ō) or as diphthongs (iə, üə, uə).

ther than others. The situation at the end of the 19th century was as indicated in Figure 2. Line 2, *maken/machen*, is generally chosen as the boundary between Low German and High German, because it is typical for the shift of *p*, *t*, and *k* after vowels to *ff*, *ss*, and *ch*, respectively (*hopen/hoffen*, *bīten/beissen*, *maken/machen*), and of *t* and *tt* to *z* and *tz*, respectively (*ten/zehn*, *sitten/sitzen*). The shift of *ik* "I" to *ich* is indicated by line 1, which



Figure 2: The Netherlandic–German dialect divisions. Numbers refer to isoglosses described in text.

shows that the shift of *k* to *ch* after a vowel in this particular word spread unusually far. Line 3, which indicates the shift of *Dorp* "village" to *Dorf* (*cf.* archaic English "thorp"), shows that shifted *p* after *r* and *l* spread less far north than did shifted *p*, *t*, *k* after a vowel. And line 4, indicating the shift of *dat* "that" to *das*, shows that the shift of *t* to *s* after a vowel spread still less far north in this word (and in a few others: *it/es* "it," *wat/was* "what"). The striking way in which these lines "fan out" in the west (in the area along the Rhine River) has led to their being called the "Rhenish fan."

The shift of *p* when doubled or at the beginning of a word occurred in a much smaller area. Line 5, showing the shift of *Appel* "apple" to *Apfel*, lies wholly within the High German speech area and is customarily used to subdivide it into Middle German (*Appel*) and Upper German (*Apfel*) areas. Line 6, which indicates the shift of *Pund* "pound" to *Pfund*, follows much the same course as does line 5 in the west, but it then runs north to join the *maken/machen* line; it is customarily used to distinguish West Middle German (*Appel*, *Pund*) from East Middle German (*Appel*, *Fund*—the latter being more common than Upper German *Pfund*).

Beginning of German

Preliterary period (to c. AD 750). As noted above (in the section on the branches of Germanic), during the early centuries of the Christian era there was only one "Germanic" language, with little more than minor dialect differences. Only after the consonant shift just described is there justification in speaking of a "German" (*i.e.*, High German) language distinct from the other Germanic languages. The fact that many early loans from Latin spread throughout all of Germanic makes it clear that the various dialects of early Germanic were mutually intelligible and that there was easy communication among them. At the same time, the modern German forms of these early loans show that they must have been borrowed before the consonant shift, because they show its effects. Examples include Latin *pondō*, English "pound," but German *Pfund;* Latin *piper*, English "pepper," but German *Pfeffer;* Latin *tegula*, English "tile," but German *Ziegel;* Latin (*via*) *strāta* "paved (road)," English "street," but German *Strasse;* Latin *catillus*, English "kettle," but German *Kessel;* and Latin *coquus*, English "cook," but German *Koch*.

Toward the end of the 4th century there began the great migrations (German *Völkerwanderung*) of Germanic tribes, resulting in an expansion of the Germanic-speaking territory. Angles, Saxons, and Jutes crossed the channel to England; Franks moved southwest into northern France and south into southern Germany; and Alamanni, Bavarians, and Langobardi moved south into southern Germany, Switzerland, Austria, and northern Italy. At the same time, the area east of the Elbe and Saale rivers was largely vacated by Germanic speakers, and Slavic speakers moved in.

Old period (c. 750–1050). In the southern area settled by Franks, Alamanni, and Bavarians, the first Old High German written records began to appear during the 2nd half of the 8th century. Their language is best described as a collection of monastery dialects; there is a certain uniformity in the writings of any given monastery, but little for the area as a whole. The first documents are translations into German of Latin word lists. Later documents include prose translations of St. Isidore of Seville (made *c.* 800) and of Tatian (*c.* 830), as well as a new verse form with end rhyme (Otfrid, *c.* 870). This literature reached its highest point in the able translations and interpretations of the Swiss monastery teacher Notker Labeo (died 1022). From the north (the Old Low German or Old Saxon speech area), the most extensive documents preserved are a life of Christ in alliterative verse (*Heliand*, *c.* 830) and a fragment of a similar Genesis translation.

In this period there were many borrowings from Latin, nearly all connected with Christianization of the Germans. Because they were made after the consonant shift, they do not show its effects. Examples of these borrowings include *predigōn* (modern German *predigen* "to preach"), from Latin *praedicāre; tempal* (modern German *Tempel* "temple"), from Latin *templum;* and *spiagal* (modern German *Spiegel* "mirror"), from Latin *speculum*. On the other hand, borrowings of this period reflect sound changes that had occurred in popular Latin, such as the change of Latin *c* before *e* from a *k* sound to *ts* in *ceila* "ceil" and *crucem* "cross," Old High German *zella*, *krūzi*, modern German *Zelle*, *Kreuz* (the letter *z* in the German and Old High German examples represents the sound of *ts*); or the change of Latin medial *-b-* to *-v-* in *tabula* "table," borrowed into Old High German as *tavala*, modern German *Tafel*.

Middle period (c. 1050–1350). Several developments justify the usual assumption of a new period, the language of which is called Middle High German, beginning around 1050. First, there were changes in the language itself, among which were the unvoicing of final *b*, *d*, and *g* (*cf.* Old High German *grab* "grave," *rad* "wheel," and *tag* "day" with Middle High German *grap*, *rat*, and *tac;* in modern German these words are again spelled *Grab*, *Rad*, and *Tag* but are pronounced with final *p*, *t*, and *k*) and the reduction of the vowels of unstressed syllables to a *ə* sound, usually spelled *e* (*e.g.*, in the plural of the word for "day," the Old High German nominative-accusative form was *taga*, the genitive was *tago*, and the dative was *tagun*, but for these Middle High German had *tage*, *tage*, and *tagen*, respectively, and modern German has *Tage*, *Tage*, and *Tagen*). Second, there were great changes in the geographical area in which German was spoken. In the west the Franks of northern France had become romanized, and the French–German language border had assumed approximately its present location; in the east, on the other hand, German began to spread into Slavic territory, a process that was to continue for many centuries and to be reversed only at the end of World War II. Third, writing became independent of the monasteries, and the number of written documents soon increased greatly in both north and south. In the south, especially, a remarkable literature developed that included courtly epic and *Minnesang*. There is clear evidence of a trend toward a standard Middle High German literary language, though it seems to have had no influence on ordinary speech. Because this literature was based largely on French models, many French words were borrowed into German.

Geographical expansion

Early modern period (c. 1350–1650). Four events—the growth of trade, the rise of a middle class, the invention of printing, and the Reformation—had great influence on the development of the language. In the north, because of the prosperity of the Hanseatic League, a standard Low German written language began to develop, though it never reached full growth and probably had little influence on everyday speech. In the south the dialects that had arisen in the recently settled East Middle German area were relatively uniform and contained elements from both West Middle German and Upper German. Gradually these East Middle German dialects came to be used as the official languages of the chancelleries of the area, including that of Saxony; and on this latter Martin Luther based the language of his widely read Bible translation (1522–34). The growth of this type of German, which developed gradually into modern Standard German, was aided by the fact that printers preferred it as a means of making their books appeal to the widest possible audience.

Three striking vowel changes are characteristic of this period. In the southeast, as early as the 12th century, the long vowels *ī*, *ū*, and *ǖ* came to be diphthongized to *ei*, *ou*, and *öü;* this is called the "New High German diphthongization." By the 15th century these new diphthongs had spread to East Middle German, and in the standard language they merged with the old diphthongs *ei*, *ou*, and *öü*. Examples include Middle High German *mîn* "my," *hûs* "house," and *hiuser* "houses" with the monophthongs *ī*, *ū*, and *ǖ*, in contrast to *ein* "a," *troum* "dream," and *tröume* "dreams" with the diphthongs *ei*, *ou*, and *öü*, but modern Standard German *mein*, *Haus*, and *Häuser* appear with the same diphthongs (*ai*, *au*, and *oi*) as *ein*, *Traum*, and *Träume*. By a specifically Middle German development, the diphthongs *iə*, *uə*, and *üə*, still preserved in the southern dialects, were monophthongized to long *ī*, *ū*, and *ǖ;* this is the "New High German monophthongization." Examples include Middle High German *tief* "deep," *vuoz* "foot," and *vüeze* "feet" with the diphthongs *iə*, *uə*, and *üə*, contrasted to modern Standard German *tief*, *Fuss*, and *Füsse* with the monophthongs *ī*, *ū*, and *ǖ*. Short vowels remained short in closed syllables before long consonants but were lengthened in open syllables before a short consonant plus an unstressed vowel. This is called "lengthening in open syllables."

Modern period (c. 1650 to the present). The outstanding developments of the modern period have been the increasing standardization of High German and its increasing acceptance as the supradialectal form of the language. In writing, it is almost the only form used (except for small amounts of dialect literature); in speech, it is the first or second language of nearly the entire population.

Standard German

Although Standard German is clearly based on the East Middle German dialects, it is not identical with any one of them; it has accepted and standardized many forms from other areas, notably the Upper German sound *pf* (*Pfund*, *Apfel*) and also large numbers of individual words in the form of other dialect areas. Since it is the only type of German taught in schools, its spoken form is based to a large extent on its written form; and the spoken form that carries the greatest prestige (that of stage, screen, radio, and so on) uses a largely Low German pronunciation of this written form. As a result, the spoken form of modern Standard German has often been aptly described as "High German with Low German sounds."

Characteristics of modern Standard German. *Consonants.* German has the following consonants, given here in phonetic symbols because the spelling often varies: stops, *p*, *b*, *t*, *d*, *k*, *g;* fricatives, *f*, *v*, *ç*~*x;* sibilants, *s*, *z*, *š*, *ž;* nasals, *m*, *n*, *ŋ*; liquids, *l*, *r;* glides, *h*, *j*. German *ç*~*x*, spelled *ch*, is the voiceless velar fricative *x* after *a*, *ā*, *o*, *ō*, *u*, *ū*, and *au* but is the voiceless palatal fricative *ç* in other phonetic environments. The German sound *ž* occurs only in loanwords.

In the orthography, German *w* always indicates a *v* sound; German *v* spells an *f* sound in native words but a *v* sound in loanwords. German *sp* and *st* spell the sounds *sp* and *st* in most positions, but they spell *šp* (*shp*) and *št* (*sht*) at the beginnings of words or word stems. In other positions the *š* (*sh*) sound is spelled *sch*—*e.g.*, *Schiff* "ship." Medial *ss* marks a preceding vowel as short, medial *ß* marks it as long; medial *ss*, however, changes to *ß* at the end of a word and before a consonant. German *z* always indicates the sound *ts*. The spelling *tz* marks a preceding vowel as short, and the spelling *z* marks it as long.

Voiced *b*, *d*, *g*, *v*, and *z* do not occur at the ends of words, at the ends of parts of compound words, before suffixes beginning with a consonant, or before endings in *s* or *t*. In these positions they are replaced in pronunciation (though not in spelling) by the corresponding voiceless consonants, namely *p*, *t*, *k*, *f*, and *s*. For example, the *g* in *Tage* "days" is pronounced as English *g*, and the *g* in *Tag* "day" is pronounced as English *k*.

Vowels. The German vowel system is given in Table 9 in phonetic symbols.

Table 9: Vowel System of German

short and lax vowels	long and tense vowels	diphthongs	unstressed vowel
i ü u	ī ǖ ū		
e ö o	ē ō ō	oi	ə
a	ā	ai au	

Though the spelling does not always indicate the difference between short and long vowels, the following devices are used more or less consistently: (1) A vowel is always short if followed by a double consonant letter—*e.g.*, *still* "still," *wenn* "if," *Rasse* "race," *offen* "open," *Hütte* "hut"—in contrast to the long vowels of *Stil* "style," *wen* "whom," *Straße* "street," *Ofen* "oven," *Hüte* "hats." (2) A vowel is always long if followed by an (unpronounced) *h*—*e.g.*, *ihnen* "to them," *stehlen* "to steal," *Kahn* "barge," *wohnen* "to dwell," *Ruhm* "fame"—in contrast to the short vowels of *innen* "inside," *stellen* "to place," *kann* "can," *Wonne* "bliss," *dumm* "dumb." (3) A vowel is always long if written double—*e.g.*, *Beet* "(flower)bed," *Staat* "state," *Boot* "boat"—in contrast to the short vowels of *Bett* "bed (for sleeping)," *Stadt* "city," *Gott* "god"; *ie* counts as the doubled spelling of *i*—*e.g.*, long *ī* in *Miete* "rent" but short *i* in *Mitte* "middle." (4) A vowel (except unstressed *e*) is always long when it stands at the end of a word.

The "plain" vowels—*a*, *o*, *u*, *ā*, *ō*, *ū*, *au*—often alternate with the "umlaut" vowels—*e*, *ö*, *ü*, *ē*, *ō̈*, *ǖ*, *oi*, respectively—as in the following examples with plain vowels in the singular but umlauted vowels in the plural: *Gast* "guest," *Gäste;* *Gott* "god," *Götter;* *Mutter* "mother," *Mütter*. As these examples show, the vowel sounds *e*, *ē*, and *oi* are spelled *ä*, *ä*, and *äu* when they are the umlaut of *a*, *ā*, and *au* sounds. *Gast–Gäste*, *Vater–Väter*, *Braut–Bräute*. Otherwise they are generally spelled *e*, *eh*, or *ee* (*beten* "to pray," *geht* "goes," *Beet* ["flower]bed"), and *eu* (*Leute* "people").

The sound *ai* is generally spelled *ei: Seite* "side," *nein* "no," though in a few words *ai: Saite* "string (of an instrument)," *Kaiser* "kaiser." The unstressed schwa sound (*ə*), as the *a* in English "sofa," is spelled *e: beginnen/bəgínən/* "to begin," *geredet/gərḗdət/* "spoken."

(Wi.G.M.)

YIDDISH

Although there were about 11,000,000 speakers of Yiddish before World War II, approximately half of them were killed in the Nazi holocaust. There are perhaps 5,000,000 Yiddish speakers today, including native speakers and those who use it as a second language. Most speakers live in the United States, Latin America, Israel, and the Soviet Union. They are served by an active press, including 11 daily newspapers.

International nature of Yiddish

History. Yiddish, although Germanic, is not a typical Germanic language; it includes not only Germanic features but also elements from Romance, Hebrew-Aramaic, and Slavic languages. A cursory examination of the

German component of Yiddish indicates that no Yiddish dialect stands in a one-to-one relationship to any German dialect. The language had its beginnings in the 10th century when Jews from northern France and northern Italy settled in the Rhineland. These early Jewish settlements were dislocated by the Crusades and later by the persecutions that followed in the wake of the Black Plague. The subsequent move to Slavic territory had enormous influence on the development of the language.

Onomastic evidence (evidence from recorded proper names) for Yiddish is known from 1096, and glosses in biblical commentaries may be several decades older. The earliest known connected text is a rhymed couplet inscribed in a Hebrew holiday prayer book from Worms and bearing the date 1272–73. The earliest extensive manuscript, known as the Cambridge Yiddish Codex, is explicitly dated November 9, 1382. It excites the interest of Germanicists for its version of "Ducus Horant" (a poem from the Hildesage of the Gudrun epic known from the Ambross Manuscript written by Hans Reid, 1502/4–1515), which antedates the earliest extant manuscript of the Hildesage by at least 130 years. The documentary history of Yiddish is unbroken thereafter to the present day. Unique evidence for spoken Yiddish is incorporated in an extensive body of rabbinical Responsa (published rabbinical opinions on matters of religious law) beginning in the 15th century. Testimony before the rabbinical court, recorded verbatim, provides unusual insight into the colloquial language.

Scholars divide the history of Yiddish into four periods: Earliest Yiddish, to 1250; Old Yiddish, 1250–1500; Middle Yiddish, 1500–1750; and Modern Yiddish, 1750 to the present. The earliest literary tradition had a Western Yiddish dialectal base; writing in this literary dialect continued into the Modern Yiddish period long after the major population centres had shifted to the East. The establishment of the modern literary language on an Eastern Yiddish base occurred only in the early 19th century. At the same time a new style in the language of Yiddish Bible translation emerged, free from the constraints of the original Hebrew syntax and of the stricture against the use of Yiddish words of Hebrew-Aramaic origin in translating from Hebrew. The continuous contact of Yiddish speakers with Hebrew-Aramaic texts and, in the European language area, with one or another Germanic or Slavic language have been important factors in the development of the language.

Characteristics. Because of the conditions under which Yiddish developed (*i.e.*, the numerous contacts it has had with other languages), it is of great interest to scholars.

Alphabet. Yiddish uses all the letters of the Hebrew alphabet, including traditional word-final variants, which have only recently been reintroduced into the orthography of Soviet Yiddish. Several letters occur only in words of Hebrew-Aramaic origin, which retain their traditional spelling except in the Yiddish of the Soviet Union.

Phonology. The vowel system of Standard Yiddish consists of the simple vowels *i, e, a, o,* and *u* and the diphthongs *ej, aj, oj*. Under Slavic influence a palatal series of consonants has emerged. Unlike German, *x* corresponding to German *ch* has no palatal variant, the ŋ sound (the *ng* in English "sing") is simply a positional variant of *n*, there is no glottal stop (a sound made by closure of the vocal cords), and word final voicing is distinctive (phonemic—*i.e.*, it carries a change in meaning). Words of Hebrew–Aramaic and Slavic origin have introduced a rich variety of consonant clusters that do not appear in German. Intonation contours, apparently related to the chant with which the Talmud is studied, convey syntactic–semantic distinctions independently.

Grammar. Case inflections, preserved only in the singular, appear in noun modifiers but only rarely in nouns themselves. The dative and accusative cases have merged in the masculine; the nominative and accusative cases have merged in the feminine and neuter. All prepositions govern the dative case. The system of noun plural formation, basically of German origin, is enriched by word elements of Hebrew origin. Many nouns differ from their German cognates both in gender and plural form. A well-developed system of diminution uses word elements largely of German origin but on a Slavic grammatical model. A semantically significant distinction between inflected and uninflected predicate adjectives has emerged, while the difference between weak and strong adjectives, a characteristic of other Germanic languages, has effectively disappeared. The verb is inflected only in the present indicative. Other tenses and moods are expressed by means of auxiliary words. In normal word order the inflected verb immediately follows the subject; any remaining part of the verb phrase occurs as close to the inflected verb as possible. The special word order of the German subordinate clause is unknown, and verb initial constructions generally express consecutiveness rather than interrogation.

Hebrew influences

Vocabulary. In the vocabulary, words and word elements borrowed from a number of different languages co-occur and often combine freely in a manner unfamiliar to the languages from which they derive. Furthermore, when words borrowed from different languages are partially alike, one of them may be analyzed and inflected in terms historically appropriate to the other, thereby yielding blends of complex etymology. In addition, a highly productive system of prefixing yields verbs that are German in form but derive their meanings from an underlying Slavic model.

Dialects. The basic dialectal division is between Western Yiddish, which occurs largely within the German language area, and Eastern Yiddish in the Slavic-speaking areas. Eastern Yiddish is traditionally subdivided into Northeastern Yiddish and Southern Yiddish, the latter consisting of Central Yiddish and Southeastern Yiddish. The phonological criteria on which this division is based are typically reflected in the variants of the phrase "to buy meat": Western Yiddish *kāfn flāš*, Central Yiddish *kojfn flajš*, Southeastern Yiddish *kojfn flejš*, Northeastern Yiddish *kejfn flejš*. Other phonological and many lexical differences reinforce the distinctness of Western Yiddish. In the East, Central Yiddish is further distinguished by a full set of contrasts in vowel length, while the varieties of Southeastern Yiddish have made changes in vowel quality that have led to the types *hont* "hand," *huz* "house," and *rign* "rain." Northeastern Yiddish is characterized by the loss of the neuter gender. Standard Yiddish adheres more closely to Northeastern Yiddish in its sound system, and more closely to Southern Yiddish in its grammatical patterns. (M.I.H.)

IV. North Germanic (the Scandinavian languages)

HISTORY

Runes

About 125 inscriptions dated from AD 200 to 600, carved in the older runic alphabet (futhark), are chronologically and linguistically the oldest evidence of any Germanic language. Most of them are from Scandinavia, but enough have been found in southeastern Europe to suggest that the use of runes was also familiar to other Germanic tribes. Most inscriptions are brief, marking ownership or manufacture, as on the Gallehus (Denmark) golden horn (*c.* AD 400): *Ek Hlewagastiz Holtijaz horna tawido* "I, Hlewagastiz, son of Holti, made [this] horn." A number of inscriptions are memorials to the dead, while others are magical in content. The earliest were carved on loose wooden or metal objects, while later ones were also chiselled in stone.

The inscriptions retain the unstressed vowels that were descended from Germanic and Indo-European but were lost in the later Germanic languages; *e.g.*, the *i*'s in *Hlewagastiz* and *tawido* (Old Norse would have been **Hlégestr* and **táða*) or the *a*'s in *Hlewagastiz, Holtijaz,* and *horna* (Old Norse **Høltir, horn*). The scantiness of the material (fewer than 300 words) makes it impossible to be sure of the relationship of this language to Germanic and its daughter languages. It is traditionally known as Proto-Scandinavian but shows few if any distinctively North Germanic features and may reflect a stage, sometimes called Northwest Germanic, prior to the splitting of North and West Germanic (but after the separation of Gothic). Only after the departure of the Angles and Jutes for England and the establishment of

the Eider River in south Jutland as a border between Scandinavians and Germans is it reasonable to speak of a clearly Scandinavian or North Germanic dialect.

Common Scandinavian: 600–1050. Inscriptions from the 7th century show North Germanic as a distinct, fairly uniform, and recognizable dialect. Information about the language is derived from runic inscriptions, which became more abundant after the creation of the short runic futhark about AD 800, from names and loanwords in foreign texts and from reconstructions based on place-names and later dialects. The expansion of Nordic peoples in the Viking Age (*c.* 750–1050) led to the establishment of Scandinavian speech in Iceland, Greenland, the Faeroes, Shetlands, Orkneys, Hebrides, and the Isle of Man, as well as parts of Ireland, Scotland, England, France (Normandy), and Russia. Scandinavian languages later disappeared in all these territories except the Faeroes and Iceland through absorption or extinction of the Scandinavian-speaking population.

Viking expansion

During the period of expansion, all Scandinavians could communicate without difficulty and thought of their language as one (sometimes called "Danish" in opposition to "German"), but the differing orientations of the various kingdoms in the Viking Age led to a number of dialectal differences. It is possible to distinguish a more conservative West Scandinavian area (Norway and her colonies, especially Iceland) from a more innovative East Scandinavian (Denmark and Sweden), the former oriented to the Atlantic, the latter to the Baltic. There were no firm borders, however, and the sea lanes formed active channels of contact. An example of a linguistic difference setting off the eastern dialect area is the monophthongization of the Common Scandinavian diphthongs *ei*, *au*, and *øy* to *ē* and *ø̄* (*e.g.*, *steinn* "stone" became *stēn*, *lauss* "loose" became *løs*, and *høyra* "hear" became *hø̄ra*). The diphthongs remained on the island of Gotland and in most North Swedish dialects, however, while they were lost in some East Norwegian dialects. The pronoun *ek* "I" became *jak* in East Scandinavian (modern Danish *jeg*, Swedish *jag*) but remained *ek* in West Scandinavian (New Norwegian and Faeroese *eg*, Icelandic *ég*); in East Norwegian it later became *jak* (dialects *je*, *jæ*, Dano-Norwegian *jeg*) but remained *ek* (dialects *a*, *æ*) in Jutland.

Old Scandinavian: 1050–1450. The establishment of the Christian church in its Roman Catholic form during the 10th and 11th centuries had considerable linguistic significance. It helped to consolidate the existing kingdoms, brought the North into the sphere of classical and medieval European culture, and introduced the writing on parchment of Latin letters. Runic writing continued in use for epigraphic purposes and messages for the general population (several thousand inscriptions are extant, from 11th-century Sweden, especially, and also all the way from Russia to Greenland). For more sustained literary efforts, the Latin alphabet was used—at first only for Latin writings but soon for native writings as well. The oldest preserved manuscripts date from *c.* 1150 in Norway and Iceland and *c.* 1250 in Denmark and Sweden. The first important works to be written down were the previously oral laws; these were followed by translations of Latin and French works, among them sermons, saints' legends, epics, and romances. Some of these may have stimulated the extraordinary flowering of native literature, especially in Iceland. One can hardly speak of distinct languages in this period, although it is customary to distinguish Old Icelandic, Old Norwegian, Old Swedish, Old Danish, and Old Gutnish (or Guthnic, spoken in Gotland) on the basis of quite minor differences in the writing traditions. Some of these were merely scribal habits resulting from local usage, but others did reflect the growing separation of the kingdoms and the centralization within each. Literary Old Icelandic is often presented in a normalized textbook form under the name of Old Norse.

Roman influence

Culture words like *caupō* "merchant" (giving Old Norse *kaupa* "buy") and *vinum* "wine" (Old Norse *vín*) had been filtering into the North from the Roman Empire for a long time. But the first great wave of such words came from the medieval church and its translations, often with the other Germanic languages as intermediaries because the first missionaries were English and German. Some religious terms were borrowed from other Germanic languages; among these are Old Norse *helviti* "hell" from Old Saxon *helliwiti* or Old English *hellewite*, and Old Norse *sál* "soul" from Old English *sāwol*. East Scandinavian borrowed the Old Saxon word *siala*, from which come later Danish *sjæl* and Swedish *själ*. In the secular field the most profound influence on Scandinavian was that exerted by Middle Low German because of the commercial dominance of the Hanseatic League and the political influence of the North German states on the royal houses of Denmark and Sweden between 1250 and 1450. The major commercial cities of Scandinavia had large Low German-speaking populations, and the wide use of their language resulted in a stock of loanwords and grammatical formatives comparable in extent to that which French left behind in English after the Norman Conquest.

Reformation and Renaissance: 1450–1550. The many local dialects that exist today developed in the late Middle Ages, when the bulk of the population was rural and tied to its local village or parish, with few opportunities to travel. The people of the cities developed new forms of urban speech, coloured by surrounding rural dialects, by foreign contacts, and by the written languages. The chanceries in which documents of government were produced began to be influential in shaping written norms that were no longer local but nationwide. The Reformation came from Germany and brought with it High German influence through Martin Luther's translation of the Bible, which was quickly translated into Swedish (1541), Danish (1550), and Icelandic (1584). That it was not translated into Norwegian was one of the major reasons that no separate Norwegian literary language arose. Until the 19th century there was no distinct written Norwegian but only a Norwegian variety of Danish. With the invention of printing and the growth of literacy, all speakers of Scandinavian dialects gradually learned to read (and eventually write) the new standard languages.

The modern languages. The six standard languages of today, in the order of their emergence as languages of culture and prestige, are Danish, Swedish, Icelandic, Faeroese, New Norwegian (Nynorsk), and Dano-Norwegian (Bokmål).

Danish. The norms of the first printed books in Danish continued the norm of the royal chancery in Copenhagen, which was not based on any particular dialect and probably reflected a state of the language closer to 1350 than 1550. Because of the influence of the written language, many speech forms used even by the aristocracy at that time were eliminated or branded as vulgar. Danish is clearly the Scandinavian language that has undergone the greatest amount of change away from the Common Scandinavian norm. In the 18th century a mildly puristic reform led to the replacement of many French loans by their native equivalents (*e.g.*, *imagination* by *indbildning; cf.* German *Einbildung*), and, in the 18th and 19th centuries, Danish became the vehicle of a classical literature. There are regional differences in the cultivated speech norm, but upper-class Copenhagen speech probably has the highest prestige. A spelling reform in 1958 eliminated the capitalization of nouns and introduced the letter *å* for *aa*, thereby bringing the spelling closer to that of Norwegian and Swedish. Danish is spoken by most of the nearly 5,000,000 inhabitants of Denmark and in a few communities south of the German border; it is taught in the schools of the Faeroe Islands, Iceland, and Greenland.

Swedish. Before the Swedish revolt of Gustav Vasa in 1525, Danish influence on the Swedish language had been strong; the new government, however, made vigorous efforts to eliminate this. The written norm was based on one that had developed in the manuscripts of central Sweden, extending from the Vadstena monastery in east Götaland to Stockholm and Uppsala. In relation to the speech of the area, many of its features were conservative (*e.g.*, silent *-t* and *-d* in words like *huset* "the house" and *kastad* "thrown"). The written language was culti-

Swedish dominance

vated energetically as a symbol of national strength, and in 1786 the Swedish Academy was established by King Gustav III. The language expanded its area at the expense of Danish and Norwegian by the conquest of southern and western provinces in the 17th century. After Sweden lost Finland in 1809, the role of Swedish was gradually reduced in that country. Since independence (1917), Finland has accepted Swedish as one of its official languages and has taught it in its schools, but only about 7 percent of its population uses it. Except for small Lappish and Finnish minorities, the entire population of Sweden (about 8,000,000) has Swedish as its daily language, and there is a rich and distinguished literature.

Icelandic. Important factors in the survival of Icelandic during the period of Danish rule were its continued use for literary purposes, the geographical remoteness of Iceland, a scattered population, and the great linguistic differences between Danish and Icelandic. In the period when the Scandinavian languages in continental Europe became essentially uninflected, Icelandic preserved Old Scandinavian grammar almost intact. The native Bible became a basis for the further development of Icelandic. Nevertheless, the circumstances of the language were highly restricted until self-government developed in the 19th century, and Icelandic was rediscovered by Scandinavian scholars. A firm orthography along etymological lines was gradually established, and the policy of not adopting foreign words was confirmed, so that Icelandic today offers a strikingly different appearance from the other Scandinavian languages.

Faeroese. Prior to modern times literary activity in the Faeroe Islands was minimal, but the local dialects continued to develop, though Danish was the official language. The Danish language scholar Rasmus Rask, who wrote the first Faeroese grammar (1811), described the language as a dialect of Icelandic, but it is actually an independent language, intermediate between West Norwegian and Icelandic but containing many Danish loanwords. Traditional dance ballads were written down after 1773 before the establishment in 1846 of an independent orthography. This orthography is etymologizing and unphonetic and gives Faeroese a strong Icelandic appearance. The establishment of home rule in 1948 led to the introduction of Faeroese as the primary language taught in the schools. The language is now spoken by about 35,000 people.

New Norwegian. Old Norwegian writing traditions gradually died out in the 15th century, after the union of Norway with Denmark and the removal of the central government to Copenhagen. After independence was achieved in 1814, the linguistic union with Danish persisted, but the ideology of National Romanticism stimulated a search for a national standard language. In 1853 a young, self-taught linguist of rural stock, Ivar Aasen, constructed a language norm from the spoken dialects that would continue the Old Norwegian tradition and, hopefully, might eventually replace Danish. After long research and experimentation, he presented this New Norwegian norm (often called Landsmål, but now officially Nynorsk) in a grammar, a dictionary, and in numerous literary texts. New Norwegian was officially recognized as a second national language in 1885. Today all Norwegians learn to read and write it, but only a fifth of the school population and an even smaller percentage of the writers actually use it as their primary language. It has been cultivated by many excellent authors and has a quality of poetic earthiness that appeals even to nonusers. Its norm has changed considerably since Aasen's time in the direction of spoken East Norwegian or written Dano-Norwegian.

Work of Ivar Aasen

Dano-Norwegian. Most Norwegian literature in the 19th century was written in a superficially Danish norm, but it was given Norwegian pronunciation and had many un-Danish words and constructions. The spoken norm was a compromise Dano-Norwegian that had grown up in the urban bourgeois environment. In the 1840s Knud Knudsen formulated a policy of gradual reform that would bring the written norm closer to that spoken norm and, thereby, create a distinctively Norwegian language without the radical disruption envisaged by the supporters of Aasen's New Norwegian. This solution was supported by most of the new writers in the powerful literary movement of the late 19th century. The official reforms of 1907, 1917, and 1938 broke with the Danish writing tradition and adopted native pronunciation and grammar as its normative base; the resultant language form was called Riksmål, later officially Bokmål. Controversial efforts to bring Dano-Norwegian and New Norwegian together into an amalgamated Pan-Norwegian (Samnorsk) have not yet led to any definite result. In its current form Dano-Norwegian is the predominant language of Norway's population of nearly 4,000,000, except in western Norway and among the small Lappish minority in the North. It is the language usually taught abroad as "Norwegian."

Dialects. The teaching of the standard languages in the schools and the high levels of literacy have tended to spread the urban norms of speaking. Nevertheless, very diverse dialects, partially unintelligible to outsiders, are spoken in many rural communities; some of them are used occasionally for the writing down of local traditions or for giving local colour in plays and novels. Dialect institutes for their study exist in each country. Boundaries between dialect areas are gradual and do not always coincide with national borders, so that the following traditional divisions are somewhat arbitrary: in Denmark, West (Jutland), Central (Fyn, Sjælland), and East (Bornholm); in Sweden, South (especially Skåne), Götaland, Svealand, North (Norrland), Gotland, East (Finland); in Norway, East (Lowland, Midland), Trönder (around Trondheim), North (Nordland), West. In the Faeroese language there are minor dialectal differences between the southern and northern islands; minor dialectal differences occur in Icelandic as well, but there are no clearly defined regional dialects. In the larger cities there is a range of social dialects from the everyday speech of the working classes (often similar to nearby rural speech) to the more cultivated forms of middle- and upper-class speech, including the highly formal style of courts and legislatures. Speakers of Danish, Norwegian, and Swedish normally use their own languages in communicating with one another; Norwegian and Swedish have a common phonetic base, and Norwegian and Danish share many vocabulary items.

CHARACTERISTICS

Differences between North and West Germanic

Common and distinctly Scandinavian characteristics. North Germanic differs from West Germanic (but not East Germanic) in having *ggj* and *ggv* for medial *jj* and *ww*, respectively (Old Norse *tveggja* "two," *hoggva* "hew"), *-t* for *-e* in the 2nd person singular of the strong preterite (Old Norse *namt* "you took"; *cf.* Old English *næme*), and a reflexive possessive *sin*.

North Germanic differs from East Germanic (but not West Germanic) in that original *ē* becomes *ā* (Old Norse *máni* "moon") and original *z* becomes *r* (Old Norse *meiri* "more"); furthermore, there is a new demonstrative pronoun *þessi* "this" (Danish, Swedish, and Norwegian *denne*), back vowels are mutated to front vowels by the influence of a following *i* or *j* ("*i*-umlaut"—*a* and *ā* become *æ* and *ǣ*, *o* and *ō* become *ø* and *ø̄* [*ø* represents umlauted *o*], *u* and *ū* become *y* and *ȳ* [*y* represents umlauted *u*], *au* becomes *ey* or *øy*), and the number of unstressed vowels is reduced to three (*a*, *i*, *u*).

North Germanic differs from both West Germanic and East Germanic in the following ways: rounding of unrounded vowels by following *u* or *w* ("*u*-umlaut"—*a* and *ā* become *ǫ* and *ǭ* [*ǫ* represents a low back rounded vowel], *e* becomes *ø*, *i* becomes *y*, *ei* becomes *ey* or *øy*); loss of initial *j* and *w* in some positions (Old Norse *ungr* "young," *ár* "year," *Óðinn* "Wodan," *ull* "wool"); loss of final nasals (Old Norse *frá* "from," *fara* "fare, go"; *cf.* Old English *faran*, German *fahren*); diphthongization ("breaking") of short *e* to *ja* or *jǫ* (Old Norse *jafn* "even," *jǫrd* "earth"). It has new pronouns for the 3rd person singular (Old Norse *hann* "he," *hon* "she"); attaches the reflexive pronoun (*sik*) to the verb to make a new mediopassive in *-sk*, *-st*, or *-s* (*finna sik* "find oneself" became

Old Norse *finnast* "be found, exist," Danish *findes*); attaches the demonstrative *inn* "that" to nouns as a definite article (Old Norse *fótrinn* "the foot," Norwegian and Swedish *foten*, Danish *foden*), except in West Jutland (possibly a later development); and uses *-t* as marker of the neuter in pronouns and adjectives (Old Norse *hvítt* "white" from *hvít-*, *eitt* "one" from *ein-*). Furthermore, North Germanic employed *es* (which changed to *er*) and later *sum* as an indeclinable relative pronoun; and it lost some Germanic prefixes, such as *ga-* (German *ge-*), and contains a considerable number of words such as *hestr* "horse," *fær* or *fár* "sheep," *gríss* "pig," *gólf* "floor," and *ostr* "cheese" that do not occur in East or West Germanic.

Orthography. The five basic vowel symbols of the Latin alphabet are supplemented by a number of special symbols, mostly for umlauted vowels: thus, there is *y* (pronounced as German *ü*), *æ* (used in Danish, Norwegian, Icelandic, and Faeroese) and the corresponding *ä* (used in Swedish), *ø* (in Danish, Norwegian, and Faeroese) and the corresponding *ö* (in Swedish and Icelandic), and *å* (also written *aa*, used in Danish, Swedish, and Norwegian).

Their present-day values are not identical; Icelandic *æ* is pronounced as the diphthong sound *ai* (as the *i* in English "ice"). Icelandic also uses accents on vowels that were long in Old Norse but are now mostly diphthongs (*á*, *é*, *í*, *ó*, *ú*, and *ý*); Faeroese has the same system except for *é*. The consonant symbols are the usual Latin ones, except that *þ* (thorn) and *ð* (edh) are used in Icelandic for voiceless and voiced *th* (*ð* in Faeroese has a different value). Loanwords containing *c*, *q*, *w*, *x*, and *z* have generally been naturalized by substituting, respectively, *k* or *s*, *kv*, *v*, *ks*, and *s* (*e.g.*, *kontakt* "contact" but Norwegian *sigar* "cigar" versus Danish and Swedish *cigar*).

Phonology. Stress is on the first syllable in native words, with sporadic exceptions for compounds. Stress on a later syllable reflects borrowing from other languages, except in Icelandic, which has stress on the first syllable of all words.

Influence of borrowing on stress

Pitch is usually high on the stressed syllable, falling at the end of a statement, rising for a yes-no question. An exception is East Norwegian and some Swedish dialects, in which the stressed syllable is low and the pitch is often rising at the end of statements. In most of Norway and Sweden and in scattered Danish dialects, there is a special word tone, by which old monosyllables have one kind of pitch while old polysyllables have another. The first pitch type is usually high or low pitch on the stressed syllable, like that in other Germanic languages, while the second is more complex and varies from region to region. In Danish the tones have been replaced by glottalization in instances in which Norwegian and Swedish have the first type.

Vowels are short before two or more consonants (with some exceptions) or when unstressed. Doubled consonants after short stressed vowels are pronounced long, except in Danish, which also does not double consonants in final position.

The Common Scandinavian vowel system contained ten vowels, each of which could be long, short, or nasalized: front unround (*i*, *e*, and *æ*), front round (*y*, *ø*, and *ǿ*), back round (*u*, *o*, and *ǫ*), and back unround (*a*). There were three falling diphthongs: front unround (*ęi*), front round (*øy*), and back round (*au*). While most of these are still present in some dialects, there have been many changes. The nasalized vowels disappeared, though they were still present in Icelandic around 1150. Diphthongs became long vowels in Danish and Swedish in the 10th century. Short low umlauted vowels coalesced with neighbouring vowels (*æ* became *e*, *ǿ* became *ø*, *ǫ* became *o*/*ø*). Long *ā* (Old Norse *á*) was rounded to *å* (pronunciation similar to the *o* in English "order"; in Icelandic and West Norwegian, pronunciation is like the *ow* in "now"). In Norwegian and Swedish the rounded vowels were shifted upwards and forwards, giving "overrounded" *o* and *u* that resemble *u* and *y* respectively. The unstressed vowels *a*, *i*, and *u* have remained in Icelandic and Faeroese but have been partially merged in New Norwegian and Swedish (written *a*, *e*, *o*), completely merged as *ə* (the schwa sound, as *a* in English "sofa") in Danish and Dano-Norwegian, and lost in Jutland and Trönder dialects. High round vowels (*y*, *ȳ*, *øy*) have been merged with the unround vowels in Icelandic and Faeroese (and in scattered dialects elsewhere) but are still distinguished in writing. Long vowels have been diphthongized not only in many dialects (*e.g.*, Jutland, Skåne, and West Norwegian) but also in standard Icelandic and Faeroese (Icelandic *é*, pronounced [je], *ó* [ou], *á* [au], *æ* [ai]; Faeroese *í* [ui], *æ* [æa], and so on). (Symbols in brackets are phonetic symbols designating actual pronunciation.) A quantity shift took place in the late Middle Ages, in which short vowels were lengthened before single consonants and long vowels were shortened before clusters, sometimes with qualitative changes that affected different dialects differently; thus, in Swedish *veta* "know" *i* became *e* (though all the other languages have *i*).

The Common Scandinavian consonant system contained voiceless stops (*p*, *t*, *k*), voiced stops (*b*, *d*, *g*), voiceless-voiced spirants (*f*/*ƀ*, *þ*/*ð*, *x*/*g*), nasals (*m*, *n*), a sibilant (*s*), liquids (*l*, *r*), and glides (*w*, *j*). The chief changes were as follows: Short voiceless stops became voiced after vowels in Danish and neighbouring dialects and then partially opened to become spirants or glides (*tapa* became *tabe* "lose," *ūt* became *ud* "out," *kakur* became *kager* "cakes"). Velar stops (*k*, *g*, *sk*) were palatalized before front vowels to merge with *kj*, *gj*, and *skj*, as still occurs in Icelandic (and Jutland dialect); in Faeroese, Norwegian, Swedish, and many Danish dialects, these were fronted to *tj*, *dj*, *stj* or even opened to spirants [ç, j, š], while in Danish they reverted to *k*, *g*, and *sk*. Voiced *f* [*ƀ*] merged with *w* to become *v*, though it is still written *f* in Icelandic; in Danish both *f* and *w* have become pronounced as *w* after vowels. Voiceless *þ* became *t* (occasionally *h* in Faeroese) and voiced *þ* [*ð*] became *d*, except in Icelandic. Voiceless *x* became *h* initially before vowels, but was lost elsewhere; voiced *x* [*g*] became *g*, except in Icelandic (in Danish it has become either [j] or [w] after vowels). The *r* sound was assimilated to following dental sounds (*l*, *n*, *s*, *t*, *d*) to make a series of retroflex consonants (*l̩*, *n̩*, *s̩*, *t̩*, *d̩*, pronounced with the tip of the tongue curled up towards the hard palate) in many Swedish and Norwegian dialects, including those of Oslo and Stockholm. In much the same area a "dark" *l* was merged with *rð* to make a new "thick *l*," defined as a "cacuminal flap" not acceptable in standard speech. The *r* sound became a uvular *r* in Danish and in the dialects of the nearest parts of Sweden and Norway in the last century or two.

Morphology. The Common Scandinavian system of inflections in nouns, adjectives, pronouns, and verbs is almost totally preserved in Icelandic, if allowance is made for some sound changes (*e.g.*, *-r* becomes *-ur* as in *situr* "sits," and *-t* becomes *ð* as in *húsið* "the house"). In Faeroese and New Norwegian the genitive case is replaced in speech by prepositional phrases or compounds. Declensions in Faeroese have been simplified in the plurals of nouns (all end in *-r*), verb plurals (*-a* in present, *-u* in preterite), verb singulars (*-i* in weak present for all persons), and the subjunctive (*-i* in the present, same ending as indicative in the preterite). In the remaining languages all case forms except the genitive (which invariably ends in *-s*) have been merged, as have markers of person and number in the verbs. This simplification began in Danish and spread to Norwegian and Swedish between 1200 and 1500, perhaps under the influence of Low German.

Loss of inflection

Major similarities. The present-day system of Danish, Dano-Norwegian, New Norwegian, and Swedish is basically identical. Nouns have singular and plural forms, to which the definite article may be suffixed; the plural suffixes vary, reflecting earlier stem, gender, and umlaut classes. Adjectives have neuter singulars marked by *-t*, plurals marked by a vowel (*-e* or *-a*), and weak forms used after determiners, usually identical in form with the plurals; the comparatives are marked by *r* and superlatives by the cluster *st*. Adverbs derived from adjectives are identical in form with the neuter singular forms

of the adjectives. Personal pronouns occur in three persons and in both singular and plural. In part, they still distinguish nominative and accusative (*e.g.*, Swedish *jag* "I"—*mig* "me"). There are polite pronouns of address that are either identical with the 2nd person plural (Swedish *ni*, Icelandic *þér*, Faeroese *tygum*, and New Norwegian *de* or *dykk*) or the 3rd person plural (Danish or Dano-Norwegian *De*); in Icelandic and Faeroese old duals have taken over the function of plurals (Icelandic *við* "we," *þið* "you"; Faeroese *vit* "we," *tit* "you"). Each personal pronoun has a corresponding possessive pronoun, the 3rd person being identical with the genitive of the pronoun and invariable. The possessive pronouns for the other persons and the reflexive *sin* are inflected for gender and number like most other pronouns and articles. Verbs inflect for tense only, with *-r* as the usual present marker (New Norwegian does not have an ending to indicate present tense in the strong verbs), while the preterites have stem-vowel ablaut changes in the strong verbs and a dental suffix in the weak verbs. Nonfinite forms of the verb have invariable suffixes (*-a* or *-e* for the infinitive, *-ande* or *-ende* for present participles, and *-at* or *-et* for perfect participles), except that Swedish and New Norwegian mark gender when the perfect participle is used adjectivally.

Major differences. New Norwegian, like Icelandic and Faeroese, and, in part, Dano-Norwegian preserve masculine, feminine, and neuter genders; Danish and Swedish combine masculine and feminine into a common (non-neuter) gender. Swedish and New Norwegian (in part) preserve non-neuter plurals in *-ar*, *-er*, and *-or*, which merged as *-er* in Dano-Norwegian; in Danish these have become *-e*, while a new plural in *-er* has arisen, primarily for loanwords. The preterite of first class weak verbs (Old Norse *-aði*) ends in *-a* in New Norwegian, *-et* in Dano-Norwegian, *-ede* in Danish, and *-ade* in Swedish (usually pronounced *-a*). In Norwegian and Swedish a new class of weak verbs with preterite ending *-dde* has arisen, including stems ending in *-d* or long vowels (Swedish *födde* "bore," *bodde* "lived"). The present tense form of strong verbs is umlauted in New Norwegian (as in Icelandic and Faeroese); it is monosyllabic in New Norwegian, has high or low pitch on the stressed syllable in Dano-Norwegian and Swedish, and glottalization in Danish (New Norwegian *kjem;* Dano-Norwegian, Swedish *kommer*, pronounced *'kåmmər;* Danish *kommer*, pronounced *kåmˀər*. New Norwegian has *-st* in the mediopassive (like Icelandic and Faeroese); Dano-Norwegian, Swedish, and Danish have *-s*.

Syntax. The reduction of morphological complexity has been accompanied by the emergence of a more rigid order of sentence elements. Normal order is subject–finite verb–indirect object–direct object. The verb must precede the subject in yes-no questions, or when any part of the predicate is put first. Contrary to German practice, the verb keeps its normal place in subordinate clauses, except that negatives and other lightly stressed adverbs usually precede the finite verb (except in Icelandic). Complex verb phrases are formed with modal auxiliaries (*e.g.*, *kan* "can") and infinitives or with the perfect auxiliaries *ha(ve)* "have" and *få* "get" (Icelandic *geta*) and the perfect participle. Instead of such durative aspect markers as the English progressive (*e.g.*, "is talking"), verbs indicating position are combined with the main verb (*e.g.*, Dano-Norwegian *han sitter* [*står, går, ligger*] *og prater* "he is sitting [standing, walking, lying] and talking."). Icelandic has special constructions for present and perfect aspects (*er að ganga* "is going" or *er buinn að ganga* "is through going").

Normal word order in sentences

Major differences in the Norwegian languages, Swedish, and Danish are few: (1) New Norwegian and Swedish use the nominative after a copula (*Det er eg/jag* "It is I"), Dano-Norwegian and Danish, the accusative (*Det er meg/mig* "It is me"). (2) A complex passive is formed either with Old Scandinavian *verða* (Swedish *varda*, New Norwegian *verta*) or Low German *bliven* (Danish *blive*, Dano-Norwegian *bli*) and the perfect participle. (3) Swedish supplements the polite pronoun of address with a pronominal use of titles: *Önskar professoren kaffe?* "Do you [professor] wish coffee?" (4) The reflexive pronoun *sin* is used with singular or plural subjects, except in Danish, in which it is used only with singular subjects. (5) A definite article is indicated by a form before the adjective and a suffix after the noun ("double definite"), except in Icelandic and Danish (*e.g.*, in Norwegian and Swedish *det store* [*stora*] *huset* "the big house," both *det* and *-et* in *huset* mean "the," in Danish the suffix *-et* is not used: *det store hus*). (6) A possessive may follow its noun in Icelandic, Faeroese, and Norwegian but not in Danish or Swedish (Icelandic *hesturinn minn* "my horse," literally, "horse mine," Swedish *min häst* "my horse"). (7) The numeral "one" is used (in unstressed form) as an indefinite article (*i.e.*. as "a," "an"), except in Icelandic, which has no indefinite article. (8) Swedish omits the auxiliary *hava* "have" in subordinate clauses (*Huset jag sett. . . .* "The house I [have] seen . . .").

Vocabulary. The everyday stock of Scandinavian words, including most of the high frequency words, is Indo-European and Germanic in its core. Of the 200,000 or more entries in the large dictionaries of each language, the vast majority are either compounds and derivatives of the simpler words or else borrowings from other languages—mostly of a scientific and cultural nature. At the present time the chief source of loanwords is English.

Borrowings

Icelandic preserved the creative powers of the older language by making it a policy not to accept new words in unassimilated form. Whenever possible, new compounds and derivatives have been created to avoid the borrowing of foreign terms. To some extent Faeroese and New Norwegian have followed the same policy but without the success of Icelandic. Danish, Swedish, and Dano-Norwegian have adopted numerous German words, along with their prefixes and suffixes; *e.g.*, Danish and Norwegian *betale* and Swedish *betala* "pay" from Low German *betalen* (*cf.* Icelandic and Faeroese *gjalda, borga*). A knowledge of German is very helpful in learning to read Norwegian, Swedish, and Danish, but this is less true for Icelandic and Faeroese, which even baffle fellow Scandinavians.

The borrowings of Danish, Swedish, and Norwegian reflect the varied contacts discussed above. Their vocabulary consists of a native core, a German middle layer (with words like Danish *skrædder* "tailor"; *cf.* Icelandic and Faeroese *klæðskeri*, literally "cloth-cutter"), and an international outer layer (with words like *psykologi* "psychology"; *cf.* Icelandic and Faeroese *sálfrædi*, literally, "soul science"). While there are some differences among the languages in the exact composition of these layers, there is also considerable agreement. Differences occur especially in words of local origin (slang, humour, endearments, abuse) and in borrowings of different origin; *e.g.*, Norwegian *etasje*/Swedish *våning*/Danish *sal* "story" (in a hotel), from French *étage*, Middle Low German *woninge*, and Old Scandinavian *salr* (but with its meaning from North German *Saal*).

(Ei.H.)

BIBLIOGRAPHY

History and classification of the Germanic languages: The reconstruction of Proto-Germanic and the derivation of Proto-Germanic from Proto-Indo-European are treated in FRANS VAN COETSEM (ed.), *Toward a Grammar of Proto-Germanic* (1972); EDUARD PROKOSCH, *A Comparative Germanic Grammar* (1939); HERMANN HIRT, *Handbuch des Urgermanischen*, 3 vol. (1931–34); ANTOINE MEILLET, *Caractères généraux des langues germaniques*, 7th ed. (1949); WILHELM STREITBERG, *Urgermanische Grammatik* (1896, reprinted 1963).

East Germanic and Gothic: (*General survey*): JAMES W. MARCHAND, "The Gothic Language," *Orbis*, 7:492–515 (1958). (*Texts*): WILHELM STREITBERG (ed.), *Die gotische Bibel*, 3rd ed. (1950). (*Grammar*): JOSEPH WRIGHT, *Grammar of the Gothic Language . . .*, 2nd ed. (1954); FERNAND MOSSE, *Manuel de la langue gotique*, 2nd ed. (1956); THEODOR WILHELM BRAUNE, *Gotische Grammatik*, 16th ed. (1961); WOLFGANG KRAUSE, *Handbuch des Gotischen* (1953). (*Dictionary and concordance*): ERNST SCHULZE, *Gothisches Glossar* (1848). (*Etymological dictionary*): SIGMUND FEIST, *Vergleichendes Wörterbuch der gotischen Sprache*, 3rd ed. (1939). (*Ostrogothic*): FERDINAND WREDE, *Über die Sprache der Ostgoten in*

Italien (1891). (*Vandalic*): FERDINAND WREDE, *Über die Sprache der Wandalen* (1886). (*Bibliography*): FERNAND MOSSE, "Bibliographia Gotica," *Mediaeval Studies*, 12:237–324 (1950), supplements in 15:169–183 (1953), and 19:174–196 (1957).

Frisian: (*General survey*): BO SJOLIN, *Einführung in das Friesische* (1969). (*Old Frisian, grammar and phonology*): WALTHER STELLER, *Abriss der altfriesischen Grammatik* (1928). (*Dictionary*): FERDINAND HOLTHAUSEN, *Altfriesisches Wörterbuch* (1925). (*Modern West Frisian, grammar*): K. FOKKEMA, *Beknopte Friese Spraakkunst*, 2nd ed. (1967). (*Phonology*): ANTONIE COHEN *et al.*, *Fonologie van het Nederlands en het Fries*, 2nd ed. (1961). (*Dictionary*): H.S. BUWALDA, G.A.G. MEERBURG, and Y.R. POORTINGA (eds.), *Frysk Wurdboek*, 2 vol. (1952–56), Netherlandic-Frisian, Frisian-Netherlandic. (*Dialects*): NILS ARHAMMER, "Friesische Dialektologie," in *Germanische Dialektologie*, vol. 1, pp. 264–317 (1968).

Netherlandic: (*General*): WALTER LAGERWEY, *Guide to Dutch Studies: Bibliography of Textual Materials for the Study of Dutch Language, Literature, Civilization* (1961); C.B. VAN HAERINGEN, *Netherlandic Language Research*, 2nd ed. (1960). (*History*): MORITZ SCHONFELD, *Historische Grammatica van het Nederlands*, 6th ed. (1960); C.G.N. DE VOOYS, *Geschiedenis van de Nederlandse Taal*, 5th ed. (1952). (*Old Low Franconian*): ROBERT L. KYES (ed.), *The Old Low Franconian Psalms and Glosses* (1969). (*Phonology*): ANTONIE COHEN *et al.*, *Fonologie van het Nederlands en het Fries*, 2nd ed. (1961); B. VAN DEN BERG, *Foniek van het Nederlands*, 2nd ed. (1960); EDGAR BLANCQUAERT, *Practische Uitspraakleer van de Nederlandse Taal*, 4th ed. (1953). (*Grammar*): ETSKO KRUISINGA, *A Grammar of Modern Dutch* (1924). (*Dialects*): A.A. WEIJNEN, *Nederlandse Dialectkunde* (1966).

Afrikaans: (*History*): G.G. KLOEKE, *Herkomst en Groei van het Afrikaans* (1950). (*Phonology*): MEYER DE VILLIERS, *Afrikaanse Klankleer: Fonetiek, Fonologie en Woordbou*, 3rd ed. (1965). (*Grammar*): H.J.J.M. VAN DER MERWE, *An Introduction to Afrikaans* (1952). (*Dictionary*): D.B. BOSMAN, I.W. VAN DER MERWE, and L.W. HIEMSTRA, *Tweetalige Woordeboek*, vol. 1, *Afrikaans-Engels* and vol. 2, *Engels-Afrikaans*, 5th ed. (1964; 7th ed., 1 vol., 1967). (*Comparison with Netherlandic*): JAMES L. WILSON, *Some Phonological, Morphological and Syntactic Correspondences Between Standard Dutch and Afrikaans* (1967; available from University Microfilms).

German: (*History*): W.B. LOCKWOOD, *An Informal History of the German Language, with Chapters on Dutch and Afrikaans, Frisian and Yiddish* (1965); JOHN T. WATERMAN, *A History of the German Language* (1966); ADOLF BACH, *Geschichte der deutschen Sprache*, 8th ed. (1965). (*Pronunciation*): WILLIAM G. MOULTON, *The Sounds of English and German* (1962); THEODOR SIEBS, *Deutsche Aussprache: Bühnenaussprache*, 19th ed. (1969); MAX MANGOLD (ed.), *Duden Aussprachewörterbuch* (1962). (*Spelling*): *Duden: Rechtschreibung der deutschen Sprache und der Fremdwörter*, 16th rev. ed. (1967). (*Grammar*): HERBERT LEDERER, *Reference Grammar of the German Language* (1969), trans. and adapted from HEINZ GRIESBACH and DORA SCHULZ, *Grammatik der deutschen Sprache*, 6th ed. (1967); HERBERT L. KUFNER, *The Grammatical Structures of English and German* (1962); PAUL GREBE (ed.), *Duden Grammatik der deutschen Gegenwartssprache*, 2nd ed. (1966). (*Dialects*): R.E. KELLER, *German Dialects* (1961); ADOLF BACH, *Deutsche Mundartforschung*, 2nd ed. (1950); V.M. SCHIRMUNSKI, *Deutsche Mundartkunde* (1962; orig. pub. in Russian, 1956).

Yiddish: The main works up to 1958 are listed in URIEL and BEATRICE WEINREICH, *Yiddish Language and Folklore: A Selective Bibliography for Research* (1959). Significant recent works include: *For Max Weinreich on His Seventieth Birthday: Studies in Jewish Languages, Literature, and Society* (1964); *The Field of Yiddish: Studies in Language, Folklore and Literature*, 2nd collection, ed. by URIEL WEINREICH (1965); *The Field of Yiddish: Studies in Language, Folklore, and Literature*, 3rd collection, ed. by MARVIN I. HERZOG, WITA RAVID, and URIEL WEINREICH (1969); JOSHUA A. FISHMAN, *Yiddish in America: Socio-linguistic Description and Analysis* (1965); MARVIN I. HERZOG, *The Yiddish Language in Northern Poland: Its Geography and History* (1965); URIEL WEINREICH, *Modern English-Yiddish, Yiddish-English Dictionary* (1968), and "Yiddish Language," *Encyclopaedia Judaica*, vol. 16, col. 789–798 (1971).

Scandinavian (North Germanic): A survey of research on Scandinavian languages since 1918 by EINAR HAUGEN and THOMAS L. MARKEY is presented in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 11 (forthcoming). Important recent contributions are treated in *The Nordic Languages and Modern Linguistics*, ed. by H. BENEDIKTSSON (1970). (*Histories*): Compact histories of all the languages are presented in ELIAS WESSEN, *Die nordischen Sprachen* (1968); and EINAR HAUGEN, *The Scandinavian Languages* (forthcoming). Detailed histories include: D.A. SEIP, *Norsk språkhistorie til omkring 1370*, 2 vol. (1955; German trans., 1971), for Old Norwegian; PETER SKAUTRUP, *Det danske sprogs historie*, 4 vol. and index (1944–68), for Danish; ELIAS WESSEN, *Svensk språkhistoria*, 3 vol. (1965), for Swedish; GUSTAV L. INDREBO, *Norsk målsoga* (1951), for Norwegian; and EINAR HAUGEN, *Language Conflict and Language Planning* (1966), for the modern Norwegian period. (*Grammars*): The classic grammars of the older languages are ADOLF G. NOREEN, *Altisländische und altnorwegische Grammatik*, 4th ed. (1923); and *Altschwedische Grammatik* (1904); and JOHANNES BRONDUM-NIELSEN, *Gammeldansk grammatik*, 5 vol. (1950–65). For introductory purposes the best grammar is E.V. GORDON, *An Introduction to Old Norse*, 2nd ed. (1957). Grammars of the modern languages are: PAUL DIDERICHSEN, *Elementaer dansk grammatik*, 2nd ed. (1957), for Danish, and Diderichsen's compendium *Essentials of Danish Grammar* (1964), as well as AAGE K. HANSEN, *Moderne dansk*, 3 vol. (1967); ADOLF G. NOREEN, *Vårt språk*, 9 vol. (1903–24), for Swedish; AUGUST WESTERN, *Norsk riksmåls grammatikk* (1921), for Dano-Norwegian; OLAV T. BEITO, *Nynorsk grammatikk* (1970), for New Norwegian; W.B. LOCKWOOD, *An Introduction to Modern Faroese* (1955), for Faeroese; STEFAN EINARSSON, *Icelandic: Grammar, Texts, Glossary* (1945). Introductory textbooks for English-speaking users are (beside the two just preceding) ELIAS BREDSDORFF, *Danish* (1956); EINAR HAUGEN and KENNETH G. CHAPMAN, *Spoken Norwegian*, rev. ed. (1964); NILS-GUSTAV HILDEMAN and ANN-MARI BEITE (eds.), *Learn Swedish*, 2nd ed. (1964); and volumes in the "Teach Yourself Series." (*Dictionaries*): There are many-volumed native dictionaries for each language. Only some bilingual dictionaries are listed here: HERMANN VINTERBERG and JENS AXELSEN, *Dansk-engelsk ordbog*, 7th ed., 2 vol. (1967); RICHARD CLEASBY and GUDBRAND VIGFUSSON, *An Icelandic-English Dictionary*, 2nd ed. (1957); W.E. HARLOCK, *Svensk-engelsk ordbok* (1944); the technical I.E. GULLBERG, *Svensk-engelsk fackordbok* (1964); and EINAR HAUGEN (ed.), *Norwegian-English Dictionary* (1965).

(Wi.G.M./Ei.H./M.I.H.)

Germanic Law

Germanic law is a designation that covers the laws of the various peoples of Germanic stock from the time that the earliest "barbarian" tribes came into contact with the Romans until their tribal laws developed into national territorial laws—a development that occurred at different times with different peoples. Thus some of the characteristics of Scandinavian legal collections of the 12th century are similar to those in the Visigothic laws of the 6th century.

Knowledge of the early Germanic period is derived mainly from the observations of tribal life contained in Julius Caesar's *Gallic War* and Tacitus' *Germania*. The first written collections of Germanic law are the so-called *Leges Barbarorum*, which date from the 5th century until the 9th century. They are all written in Latin and show Roman influence in the use of technical terms of Roman law. The Anglo-Saxon laws and the laws of the North Germanic group, on the other hand, are in the vernacular and owe their written form largely to the advent of Christianity.

For all of the Germanic peoples, law (West German, *reht, êwa;* High German, *wizzôd;* North German, *lagh*, from which the English word "law" is derived) was basically not something laid down by a central authority such as the king but rather the custom of a particular nation (tribe). It was essentially unwritten, being derived from popular practices, and was not sharply distinguished from morality; it was personal in the sense that it applied only to those who belonged to the nation. Each man thus followed his own law, a notion appropriate to a nomadic people who originally did not live in clearly defined territory. When, after the fall of the Roman Empire in the West, Germanic tribes took over former Roman provinces, they did not attempt to apply their laws to their Roman subjects, to whom Roman law was still applicable.

Thus the earliest Germanic code, that of Euric, king of the Visigoths in Spain and southwestern Gaul in the late 5th century, applied exclusively to Visigoths. The *Lex*

Visigothic Laws

Romana Visigothorum, or Breviary of Alaric, was issued in AD 506 for their Roman subjects. It was a compilation of "vulgar law"—Roman law adapted to fit the social and economic conditions of the late Roman Empire—and was later the main source of Roman law in the Frankish kingdom. Only in the 7th century was Visigothic law applied to Visigoths and Romans alike, the two peoples having by then substantially fused.

The *Leges Barbarorum*, then, are not legislation in the modern sense but rather the records of customs that were first collected and then declared as law. The prologue to the Salic Law (the law of the west or Salic Franks) recounts how four chosen men collected the original practices in particular cases after having discussed them with the presidents of the local popular assemblies. Nor did the *Leges* seek, as do modern codes, to set out all of the main rules of law. They were not concerned with what everyone took for granted but concentrated on matters that, perhaps as a result of migration or conquest, had become doubtful and needed authoritative exposition. They dealt with specific situations rather than general rules, particularly with court procedure, money compensation for acts of violence, and succession on death.

The initiative for declaring law usually came from the king, but the resulting laws normally required approval by the popular assemblies. Because of this collaboration between king and people, a compilation was sometimes referred to as an "agreement" or *pactus*. The Visigothic laws were an exception; they appear always to have been formulated by the king and chief landowners without popular participation. Gradually, in any case, first the Lombard and then the Frankish kings overcame their people's aversion to central government and legislated unilaterally. The Lombards, who invaded Italy in 568, had no single code of custom, but their kings issued edicts from the mid-7th century onward. In the Frankish kingdom, the Merovingian kings called their legislation *edicta* or *praecepta*, but the succeeding Carolingians characterized them as *capitularia; i.e.*, royal ordinances divided into articles (*capitula*). These included modifications of the *Leges* of the Franks or other nations in the Frankish kingdom; administrative orders to officials; and independent legislation. Charlemagne and his successors claimed the power of the Roman emperors to make laws for all their subjects, irrespective of nation, and without the consent of any assembly. The validity of the law depended solely on the oral act of the king in promulgating it.

Germanic institutions and legal organization. Germanic law recognized a distinction between free and unfree persons. Only the former had legal capacity, and they were subdivided into nobles and ordinary freemen. The nobles enjoyed a larger share in land distribution, were preferentially chosen for public office, and were protected by a larger money compensation if they were injured. Certain west Germanic tribes recognized an intermediate status of half-free persons, who could enter into legal transactions and marry but had no political rights.

Role of the king

Basically a Germanic tribe was a league of clans. Its main institutions of government were the king, his council, and the tribal assembly (*mallus*, *witan*, *mot*, *ding*, or *thing*). The king was military leader, chief priest, and president of the assembly, and he was assisted in the routine business of government by his council of elders and higher nobles. The assembly was composed of all free members of the tribe grouped in clans. It elected kings, declared war, outlawed freemen, and generally controlled the membership of the tribe by its supervision of the manumission of slaves, the emancipation of minors, and the adoption of strangers.

The dominant social institution was the "sib" (*sippe*), a term that could mean both a clan—the extended family composed of all those related by blood, however remotely, and subject to a clan chief—and also a household or narrow family, whose members were under the *mund* of the family head. A boy remained in his father's *mund* until he was emancipated on attaining physical maturity; a girl remained until she married, when she passed into the *mund* of her husband. Marriage commonly took the form of the sale of the bride to her groom for a price, which developed into a settlement for her benefit. The husband could divorce his wife at will but risked being penalized financially.

Laws of property

The main notion in the law of property was the *gewere* or power exercised by the owner, which did not clearly distinguish legal title and physical control. Various forms of limited ownership were recognized. Land was treated differently from movables. Originally it belonged to each family collectively; gradually family ownership developed into the private ownership of the family head, but for a long time he could alienate land only with the consent of the nearest heirs. Land transfer required much formality, and among the west Germanic peoples a glove or spear was handed over as a symbol of *gewere*.

On the death of the family head, his property passed to his descendants in the nearest degree of proximity, with a preference for males. (The declaration in the Salic Law that daughters could not inherit land was used by 16th-century French lawyers as additional support for the long-standing practice of excluding women or their descendants from succeeding to the crown.) In the absence of descendants, several *Leges* provided that property deriving from the father's side should return to that side and property from the mother's side to her side. The order of succession could not be altered by will.

When trade was still on a cash or barter basis, there was little need for formal contract law. A family could obligate itself to another either by pledging a thing as security (*wadium*, *gage*) or by surrendering a hostage (*gijzel*, *born*).

Later, a debt was guaranteed by formal oath accompanied by surrendering a staff to the creditor (*effestucatio*). Contractual obligation was then constituted either by oath, enforced by an action for perjury, or by delivery of a thing, in which case it was enforced by an action for theft.

Laws to control crime

Offenses against the community, such as treason, secret killing, and secret theft were punished by outlawry, pronounced by the tribal assembly. The convicted person could then be killed by anyone. Offenses against individuals, which included open killing and open robbery, became the subject of a blood feud if the criminal and victim belonged to different family groups. Peace could be bought by the payment of compensation, known as *wergild* in homicide cases and *bot* in others, which was at first voluntary and only later became obligatory. Even in the 7th century, Visigothic law still allowed retaliation in kind for all injuries except those to the head. The *Leges* contained elaborate tariffs of compensation for different kinds of injury, the amount varying according to the social status of the victim. The private feud was eventually restricted with the growth of royal authority in the Frankish period and the notion of the king's peace, breach of which was punishable by the king's court.

When the parties had appeared before a court and stated their case, the court decided the method of proof, which could be either by oath of the parties, supported by *compurgatores* (literally "oath-helpers"), the number required depending on the gravity of the case, or by ordeal or by battle. A successful claimant had to enforce judgment himself on the person or property of the defendant.

The abandonment of personal laws. With the division of the Frankish kingdom in the late 9th century, government became highly decentralized. Already before then the pattern of landholding, which determined the more important legal relationships, had begun to take on the characteristics of feudalism. Before the end of the Roman Empire much of the land had been concentrated in the hands of magnates, secular and ecclesiastical. But, unlike their predecessors under the Romans, the holders of secular land in the Germanic states became largely independent of the central government. By the 9th century, many lords had become strong enough to challenge the power of the Carolingian kings of the Frankish Empire and to make the inhabitants of their own areas their vassals. These vassals held their land from the lords as

tenants of a so-called feud or fee. Each feudal lord held a court for his tenants, and in the feudal court the lord applied the same law to all of his tenants, irrespective of their racial or national origin. Thus the old Germanic personal principle was abandoned in favour of the territorial principle, which meant the application of the custom of the region. This was usually based partly on Germanic law and partly on the Roman law of the *Lex Romana Visigothorum*, adapted in the interests of the feudal lords.

Influence of canon law

In the same period the Roman Church became the main unifying force in western Europe and began to claim jurisdiction over many matters that earlier had been considered secular rather than ecclesiastical. Church courts had existed since the Roman Empire, and their power in matters of faith was recognized by the secular authorities. The personal law of the church as an institution was always Roman, and indeed the law of the Ripuarian Franks on the Rhine expressly declared: "The church lives by Roman law." The canon law applied in the church courts was largely influenced by Roman law and contained very few Germanic elements. Now the church courts applied this law to matters that had previously been dealt with by the secular courts, such as marriage, adultery, wills, and succession. In many countries these matters remained withdrawn from Germanic law and subject to the church law even after the Reformation.

Merchants also found that the old Germanic customary law was inadequate to cope with the problems created by the rapid growth of commerce that had occurred by the 12th century. A special commercial law, based mainly on Roman law as developed by the Mediterranean seaborne traders, was developed to settle disputes between merchants, without regard to their nationality or place of residence.

These developments reduced the range of matters subject to the jurisdiction of the local county courts. In Germany some of the earlier codifications of customary law were forgotten, partly because the local judges were unable to understand the Latin in which they were written and partly because the rules that they contained were unsuited to the new social and economic conditions. The local courts applied an unwritten customary law based on the dominant tribal law of the area, and it was this that formed the basis of such codifications as the *Sachsenspiegel* ("mirror of the Saxons") in the 13th century.

Regional variations

In France the legal development in the North differed from that in the South. The regional customs in the North were made up of Germanic and Roman law, the Carolingian capitularies, and canon law, but Germanic elements predominated. In the South, the so-called *pays de droit écrit* ("land of written law"), where Gallo-Romans had been far more numerous than Franks, the custom of each district was based mainly on the vulgar law of the *Lex Romana Visigothorum*. In Italy this law existed side by side with Lombard law. In the 7th and 8th centuries that law had been subjected to a relatively sophisticated codification, whose form shows Roman influence.

In England the Norman conquerors continued the movement toward legal unity begun by the Anglo-Saxons and imposed on the country a centralized form of government more powerful than any on the Continent. In the 12th century, Henry II made the king's court a permanent court of professional judges, who assumed jurisdiction over many matters that earlier had been dealt with by other courts. The common law developed by this court was largely Germanic law.

Scandinavian developments. For several reasons—because they were isolated from the literate tradition of Rome, because their kingdoms were not unified until the 9th century, and because feudalism was less prevalent—the Scandinavian countries (including Iceland) developed forms of law distinct from the mainstream of Germanic law.

As elsewhere, the Scandinavian tribes originally had no written laws; indeed, customary laws were handed down by word of mouth until the 11th century or even later. Those learned in the traditional customs were known as "law men." They gave private counsel; advised the local assembly or *thing*, which had decisive authority to declare the law; and, in effect, acted as judges. In some areas they gave a periodic, public recital of the law (*laghsagha*, or "law-saying"). In Norway and Iceland they collectively retained this semi-official character for centuries, but in Sweden the law man became an elected public official who presided over the local *landsthing* and generally acted as the guardian of the *land*'s independence against the encroachments of the king's authority. Consolidation of several provinces under one king did not at first change this situation significantly, because cohesion was feeble and the central government's influence on legislation was restricted. A trend toward greater legal unity asserted itself, however, with the formation of more comprehensive regions with a common *thing*.

Between the 11th and the 13th centuries the provincial customary laws were first put into writing on private initiative and later officially codified, the most famous codification being the *Jydske Lov*, enacted for Jutland in 1241. Whereas the unwritten laws, in common with other early Germanic laws, had tolerated and regulated blood feuds, the new codes were in several respects more progressive. Thus, the code of the Norwegian King Magnus (1263–80) abolished private vengeance, declaring that the king's officials should initiate criminal proceedings and provide for the punishment of evildoers; crime was thus made a matter of public prosecution rather than of personal action. Further, presumably under the influence of Christianity, legal provisions were introduced for assisting paupers and the helpless. Landholding was allodial rather than feudal; that is, the ownership of the land was in the holder himself rather than in a lord from whom he held it. The freedom from outside influences, whether of Roman, canon, or other Germanic laws, largely prevailed in Scandinavia until the 18th and 19th centuries.

BIBLIOGRAPHY. For sources, see *A General Survey of Events, Sources, Persons and Movements in Continental Legal History*, "Continental Legal History Series," vol. 1 (1912); for an exhaustive survey, K. VON AMIRA and K.A. ECKHARDT, *Germanisches Recht*, vol. 1 (1960); for discussion, R. BUCHNER, "Die Rechtsquellen," in W. WATTENBACH and W. LEVISON, *Deutschlands Geschichtsquellen im Mittelalter: Vorzeit und Karolinger*, vol. 2 (1953); and E. JENKS, *Law and Politics in the Middle Ages*, 2nd ed. (1913). For substantive law, see H. BRUNNER, *Deutsche Rechtsgeschichte*, 2nd ed., vol. 1 (1906), the classic treatment; H. CONRAD, *Deutsche Rechtsgeschichte*, vol. 1, *Frühzeit und Mittelalter*, 2nd ed. (1962); and C.V. SCHWERIN and H. THIEME, *Grundzüge der deutschen Rechtsgeschichte* (1950); for very early law, see M. SCOVAZZI, *Le origini del diritto germanico: fonti, preistoria, diritto pubblico* (1957).

For Visigothic and Burgundian law, see E.A. THOMPSON, "The Barbarian Kingdoms in Gaul and Spain," *Nottingham Mediaeval Studies*, 7:3–33 (1963); and P.D. KING, *Law and Society in the Visigothic Kingdom* (1972). For Anglo-Saxon law, see F.W. MAITLAND, "The Laws of the Anglo-Saxons," *Collected Papers*, vol. 3 (1911); and H.G. RICHARDSON and G.O. SAYLES, *Law and Legislation from Aethelberht to Magna Carta* (1966). For north Germanic law, see L.B. ORFIELD, *The Growth of Scandinavian Law* (1953).

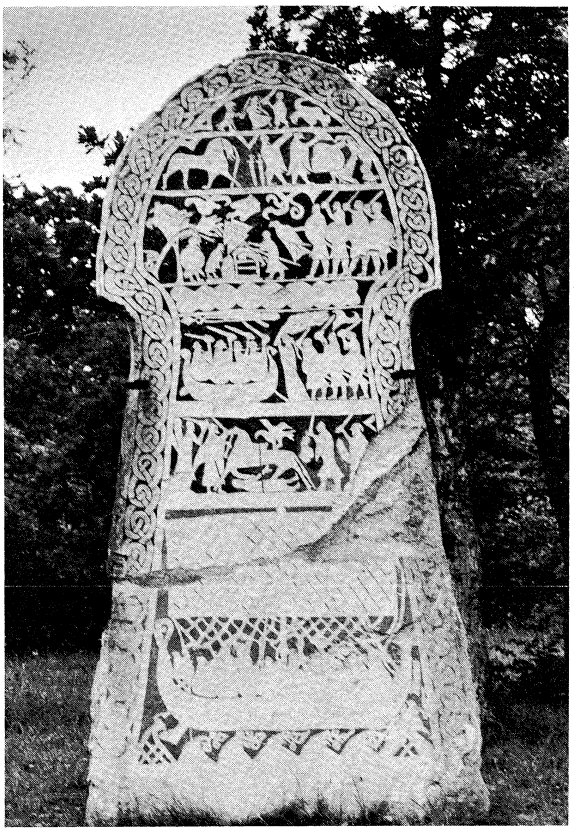
(P.G.S.)

Germanic Religion and Mythology

The term Germanic religion is used for the mythology, religious beliefs, and cults practiced by peoples who spoke one or another of the Germanic dialects before they were converted to Christianity. Germanic culture extended, at various times, from the Black Sea to Greenland, or even the American continent. Germanic religion played an important role in shaping the civilization of Europe. But since the Germanic peoples of the Continent and of England were converted to Christianity in comparatively early times, it is not surprising that little is known about the gods whom they used to worship and the forms of their religious cults. Vivid memories of Germanic religion have survived only in Scandinavia.

SOURCES

Classical and early medieval sources. The works of classical authors, written mostly in Latin and occasion-



Memorial stone from Gotland, Sweden, showing battle scenes and ships. In the third panel from the top, a warrior is being hanged in a tree as a sacrifice to Odin, whose cult is represented by an eagle and a twisted knot. Late 8th century.
Antikvarisk-Topografiska Arkivet

ally in Greek, throw some light on the religion of Germanic peoples; however, few authors showed interest in the religious practices of Germanic peoples. Tacitus, who wrote *Germania* c. AD 98, was an exception; he provided a lucid picture of customs and religious practices of continental Germanic tribes. He described some of their sacrifices, and occasionally named a god or a goddess. It is doubtful whether Tacitus had ever visited Germany; his book was based largely on older ones now lost.

Caesar (*De bello Gallico* VI, 21–4) made some remarks on the political and social organization of Germans, as well as cursory remarks on their religion. Contrasting Germans with the Celts of Gaul, Caesar was struck by the poverty of the German religion: they had no druids (*druides*), no zeal for sacrifice, and counted as gods only the sun, the fire god (Vulcanus), and the moon. For all his knowledge of the Celts, Caesar had no more than a superficial knowledge of Germans.

Early medieval records. As the power of Rome declined, records grew poorer, and nothing of great importance survives before *Getica* ("The History of the Goths"), written by the Gothic historian Jordanes c. 550; it was based partly on a larger work of Cassiodorus and partly on the work of Ablavius. The *Getica* is short and scrappy, but it incorporates valuable records of Gothic tradition, the origin of the Goths, and some important remarks about the gods whom the Goths worshipped and the forms of their sacrifices, human and otherwise.

Origin of the Lombards

A story about the origin of the Lombards is given in a tract, *Origo gentis Langobardorum* ("Origin of the Nation of Lombards"), of the late 7th century. It relates how the goddess Frea, wife of Godan (Wodan) tricked her husband into granting the Lombards victory over the Vandals. The story shows that the divine pair, recognizable from Scandinavian sources as Odin and Frigg, was known to the Lombards at this early time. A rather similar story about this pair is told in a Scandinavian source. Paul the Deacon, working late in the 8th or early in the 9th century, repeated the tale just mentioned in his fairly comprehensive *Historia Langobardorum* ("History of the Lombards"). Paulus used written sources available to him and seemed also to draw upon Lombard tradition in prose and verse.

The Venerable Bede, writing his *Historia ecclesiastica gentis Anglorum* (*Ecclesiastical History of the English People*) early in the 8th century, showed much interest in the conversion of the English, and some in the pagan religion practiced by the English before it. The lives of Irish and English missionaries who worked among Germanic peoples on the Continent (*e.g.*, Columbanus, Willibrord, Boniface) provide some information about heathens and their sacrifices.

The first detailed document touching upon the early religion of Scandinavia is the *Vita Anskarii* ("Life of St. Anskar of Rimbert"). St. Anskar (died 888) twice visited the royal seat, Björkö, in eastern Sweden, and noticed some religious practices, among them the worship of a dead king. Anskar was well received by the Swedes, but it was much later that they adopted Christianity.

Some two centuries later, c. 1072, Adam of Bremen compiled his *Gesta Hammaburgensis ecclesiae pontificum* (*History of the Archbishops of Hamburg-Bremen*), which included a description of the lands in the north, then part of the province of Hamburg. Adam's work is particularly rich in descriptions of the festivals and sacrifices of the Swedes, still largely pagan in his day.

German and English vernacular sources. Learned sources, such as those just mentioned, may be supplemented by a few written in vernacular in continental Germany and England. Among the most interesting are two charms, the so-called Merseburg Charms, found in a manuscript of c. 900, in alliterating verse. The charms appear to be of great antiquity, and the second, intended to cure sprains, contains the names of seven or perhaps eight gods and goddesses. While some of these are unknown from other sources, three or perhaps four are also known from Scandinavian records, viz., Wodan (Odin), Friia (Frigg), Volla (Fulla), Balder (Baldr). The last is somewhat doubtful. The charm also names two other deities, Phol and Sinthgut, who cannot be identified.

A manuscript of the 9th century contains a baptismal vow in the Saxon dialect, probably dating from the 8th century. The postulant is made to renounce the devil and all his works, as well as three gods, Thunaer (Thor), Woden (Odin), and Saxnot. The meaning of the last name is not clear but a Seaxnet, son of Woden, appears in a genealogy of the kings of Essex.

Old English sources

Vernacular sources in Old English are rich, but reveal little about the pre-Christian religion. The poem *Beowulf* is based upon heroic traditions, ultimately of Scandinavian origin, but it is thoroughly Christian. It does, however, retain something of Germanic morality, heroism, and defiance of death. The same might be said of minor historical poems, such as the *Battle of Brunanburh* and the *Battle of Maldon*. Old English literature also includes numerous charms, intended as safeguards against illnesses and misfortunes, but these can hardly be called religious. It is noticeable that Woden (Odin) is mentioned repeatedly in Old English sources; he is frequently named among ancestors of the royal houses.

Scandinavian literary sources. The greater part of scholarly knowledge of Germanic religion comes from literary sources written in Scandinavia. These sources are mostly written in the Old Norse language and, strange to say, they are nearly all preserved in manuscripts written in Iceland from the 12th to 14th centuries or in later copies of manuscripts written at that period. This implies a surviving tradition and an antiquarian revival in that distant outpost of Scandinavian culture.

The oldest of the sources found in the Icelandic manuscripts are in verse. Although remembered and written

down in Iceland, some of these verses originated elsewhere, some in Norway and a few in Denmark and Sweden. Some of them may well be older than the settlement of Iceland, which took place toward the end of the 9th century. The Icelanders remained pagan until the year AD 1000 (or 999).

The verse preserved in the Icelandic manuscripts may be divided broadly into two classes, the so-called Eddaic and the scaldic. The Eddaic is mostly composed in free alliterative measures, much like that of the Old English *Beowulf*. Much of the Eddaic poetry is preserved in a manuscript now called the *Elder Edda* or *Poetic Edda*, written in Iceland *c.* 1270 and containing material centuries older. The meaning of the name *Edda* is disputed; it was not originally applied to this book but to another mentioned below.

The *Elder Edda* consists of a number of lays, which may be divided into two classes, the mythological and the heroic. The mythological section contains stories about heathen gods; words of wisdom; and descriptions of the cosmos, the beginning of the world, and the Ragnarök, the end of the pagan world. The ages and places of origin of the various lays preserved in the *Edda* and minor collections are disputed. The first lay in the *Edda* is called *Völuspá* ("Sibyl's Prophecy") and, in comparatively few lines, covers the history of the world of gods from the beginning to the Ragnarök. Scholars detect the influence of Christian imagery in this poem. The scenery described is that of Iceland, and it is commonly thought that it was composed in Iceland about the year 1000, when Icelanders perceived the fall of their ancient gods and the approach of Christianity.

Hávamál

The *Hávamál* ("Words of the High One") is given as the words of Odin. It is not, in fact, one poem but consists of fragments of six or more. Some sections cannot be Icelandic, for the scenery and conditions described are rather those of Norway in the Viking Age, the 9th or 10th century. The *Hávamál* contains two interesting myths about the erotic affairs of Odin, his theft of the precious mead, and, most interesting of all, an account of how Odin hanged himself on the World Tree, Yggdrasill, a name apparently meaning "Odin's Horse."

In another poem the god Odin engages in a contest of wits with an immensely wise giant (Vafthrúdnir). The poem is in the form of question and answer, and is thus didactic; it tells of the cosmos, gods, giants, the beginning of the world and its end.

The second section of the *Elder Edda* tells of traditional Germanic heroes. Many of the stories told there are also known from continental Germany and England, but the Norse sources preserve them in an older and purer form. They are of some interest for the study of religion because the gods often intervene in the lives of heroes.

The Icelandic and, to a lesser extent, the Norwegian manuscripts of the 13th and 14th centuries contain a great bulk of poetry of a quite different kind. This is commonly, if unjustifiably, called scaldic poetry. The scaldic verse forms were perhaps devised in Norway in the 9th century. They differ fundamentally from the traditional Germanic and Eddaic forms in that the syllables are strictly counted and the lines must end in a given form. The scalds also used a complicated system of alliteration, as well as internal rhyme and consonance. Their strophes usually consist of eight lines, falling neatly into half-strophes of four lines. Scaldic poetry was often made in praise of chieftains of Norway and other Scandinavian lands. Its authors are frequently named and their approximate ages are known. The scalds allude to myths and pagan worship, and their testimony is the most reliable of all. The strict verse forms provided some safeguard against corruption, although the complex syntax and abstruse diction often make this professional poetry difficult to interpret.

After the Icelanders were converted to Christianity, much of their ancient poetry survived this religious change, as did traditions about pagan gods and their worship. Icelanders of the 12th century travelled widely and were perhaps the most lettered people in Europe, studying and translating homilies, saints' lives, and other learned literature of Europe. The 13th century saw a revival of the Icelanders' interest in the practices of their pagan ancestors, as well as in those of their kinsfolk in Norway and, to a lesser extent, in Sweden.

The *Edda* and other writings of Snorri Sturluson

The name chiefly associated with this revival is that of Snorri Sturluson (1179–1241). Snorri acquired great wealth and received the best education available in those days. He became a powerful man in Icelandic politics, and political intrigue led to his assassination in 1241. The first of Snorri's works and one of the most memorable was his *Prose Edda*, written *c.* 1220. It is to this book that the title *Edda*, whatever its meaning, originally belonged.

It is likely that Snorri wrote the various sections of this book in an order opposite to that which they now have. He began with a poem exemplifying 102 different forms of verse, addressed to Haakon, the young king of Norway, and his uncle Jarl Skúli. He then furnished a section on the speech of poetry, explaining the diction of the scalds and their abstruse allusions to gods and ancient heroes. After this, Snorri wrote the section of the *Edda* most widely read today, the *Gylfaginning* ("Beguiling of Gylfi"), which is an introduction to the mythology of the north. All the major gods and their functions are described. Snorri worked partly from Eddaic and scaldic poetry still extant, but partly from sources that are now lost. He presents a clear, if not altogether reliable, account of the gods, the creation of the world, and the Ragnarök. Another important work ascribed to Snorri is the *Heimskringla* ("Orb of the World"), a history of the kings of Norway from the beginning to the mid-12th century. The first section of this book, the *Ynglinga Saga*, is of particular interest, for in it, Snorri described the descent of the kings of Norway from the royal house of Sweden, the Ynglingar, who, in their turn, were said to descend from gods. Snorri used such written sources as were available, but he also used scaldic poems, in which he placed great faith, some of which are among the oldest known. Snorri visited Norway twice and Sweden once, and he probably used popular traditions that he heard in both countries.

About the beginning of the 13th century Icelanders began to write so-called Family Sagas; *i.e.*, lives of their ancestors who had settled in Iceland in the late 9th century, and lived through the 10th and 11th centuries. A good deal had already been written about these men in summary form by Ari the Learned (1067–1148) and other scholars of the early 12th century, but much more had been preserved in tradition handed down in verse and prose.

The reliability of Family Sagas as sources of history has long been debated and no simple answer can be given. Each saga has to be studied separately, and consideration must be given not only to the author's sources but also to his aims. Some of the authors were antiquarians and tried to relate faithfully the history of a district, a family, or a hero. Others aimed at re-creating the past by writing historical fiction.

Saxo Grammaticus

About the time when the first Family Sagas were written, the Dane Saxo Grammaticus, secretary of Archbishop Absalon (died 1201), was compiling in Latin a history of ancient gods and heroes, which formed the first nine books of his great history of the Danes (*Gesta Danorum*). These nine books deal with the prehistory of the Danes, and especially with pagan gods, whom Saxo regarded chiefly as crafty men of old. It may be that Saxo drew on Danish traditions and Danish poetry now lost, but much of his information came from vagrant Icelanders, of whom he speaks with some respect.

Material such as Saxo used was also used by Icelanders some generations later in the so-called Heroic Sagas (*fornaldar sögur*). Sagas of this kind describe the adventures of heroes who lived, or were supposed to have lived, in Scandinavia or on the Continent before Iceland was peopled. The gods, and particularly Odin, are frequently said to take part in the affairs of men, but since few of the Heroic Sagas were written before the 14th

century, and the aim of their authors was often entertainment rather than instruction, these sagas can be used as sources only with utmost discrimination.

Archaeology. The archaeological finds of Scandinavia are rich, and information about religious beliefs may be drawn especially from the grave goods and forms of burial. It may, in fact, be possible to trace continuity of belief from the Bronze Age to the Viking Age in the 9th and 10th centuries. Archaeological finds, however, are difficult to interpret from a religious point of view. Graves in which a ship or boat is found need not necessarily imply that the dead were believed, by those who buried or cremated them, to sail to the Other World. Such practices could have become simply conventional.

A number of small images in silver or bronze, dating from the Viking Age, have also been found in various parts of Scandinavia. One, found in Sweden, shows a fertility god with full erection, perhaps Freyr; another, found in Iceland, shows Thor with his hammer. Several silver images that appear to represent Thor's symbol, the hammer, have also been found in various places.

Runic inscriptions. The runic alphabet was used throughout the Germanic world beginning in about the 1st century AD. Occasionally one god or another is named; the god Thor may be called upon to hallow a grave. The runes had magical and sacral significance, as will be explained below.

Place-names. Theophoric place-names (derived from or bearing the name of a god) are found in all Germanic lands. Poor scholarly knowledge of pagan religion in continental Germany and England may be partly supplemented, especially in England, by the names of places with which gods' names are compounded. The theophoric place-names of Norway and Sweden are richer and have been carefully sifted. The names of numerous gods have come to light, but evidence drawn from them is treacherous. If scholars find a place called *Thofslundr* ("Thor's Grove") it need not necessarily mean that Thor was worshipped there, for names are often transferred by settlers from one place to another, as from England to America and, in the Viking Age, from the Scandinavian mainland to Iceland. Groups of theophoric place-names may, however, provide evidence of the cult of one god or another.

MYTHOLOGY

The beginning of the world of giants, gods, and men. The story of the beginning is told, with much variation, in three poems of the *Elder Edda*, and a synthesis of these is given by Snorri Sturluson in his *Prose Edda*. Snorri adds certain details that he must have taken from sources now lost.

Völuspa

The most rational of the older accounts is that of the *Völuspa*, defective as it is. The story is there told by a sibyl of such great age that she could remember the primeval giants who fostered her, although she does not disclose the origin of these giants. In the beginning there was nothing but Ginnungagap, a mighty void. Three gods, Odin and his brothers, raised up the earth, perhaps from the sea into which it will afterward sink. The sun shone on the barren rocks and the earth was overgrown with green herbage.

Later, Odin and two other gods came upon two figures, Askr and Embla, apparently lifeless tree trunks. They endowed them with breath, countenance, and the other qualities of life, evidently making them man and woman.

A very different story is told in the didactic poem *Vafthrúdnismál* ("The Words of Vafthrúdnir"). The poet speaks of a giant, Aurgelmir, who appears sometimes to be called Ymir. Drops of poison fell from the waters Elivagar (Élivágar; "stormy waves") and they grew into a giant. One of the giant's legs begat a six-headed son with the other leg, and under his arm grew a maid and a youth. The earth was formed from the body of the giant Ymir who, according to Snorri, was slaughtered by Odin and his brothers. Ymir's bones were the rocks, his skull the sky, and his blood the sea. Another didactic poem, *Grímnismál* ("The Words of Grímnir [Odin]"), adds further details The trees were the giant's hair and his brains the clouds. Snorri quotes the three poetic sources just mentioned, giving a more coherent account and adding some details. One of the most interesting and primitive is that about the primeval cow Audumla (Auðumla), formed from drops of melting rime. She was nourished by licking salty, rime-covered stones. Four rivers of milk flowed from her udders and thus she fed the giant Ymir. The cow licked the stones into the shape of a man; this was Buri (Búri), who was to be grandfather of Odin and his brothers. This story bears some resemblance to that of the Egyptian mother-goddess Hathor, who had the form of a cow. Parallels from other lands also have been noted.

A central point in the cosmos is the evergreen ash, Yggdrasill, on which the welfare of the world seems to depend. Its three roots stretch to the worlds of death, frost-giants and men. A hart (stag) is biting its foliage, its trunk is rotting, and a cruel dragon is gnawing its roots. When the Ragnarök approaches, the tree will shiver and, presumably, fall. Beneath the tree stands a well, the fount of wisdom. Odin got a drink from this well and had to leave one of his eyes as a pledge.

The gods. Old Norse sources name a great number of gods and goddesses. The evidence of place-names suggests that one cult succeeded another. Names, especially those in southeast Norway and southern Sweden, suggest that there was once widespread worship of a god Ull (Ullr). Indeed, an early poem reports an oath on the ring of Ull, suggesting that he was once the highest of the gods at least in some areas. Beyond that, little is known about Ull; he was god of the shield, the bow, and snowshoes.

Two tribes of gods: Aesir and Vanir

The gods can be divided roughly into two tribes, Aesir and Vanir. At one time, according to fairly reliable sources, there was war between the Aesir and the Vanir. Peace was made between them and hostages were exchanged. In this way, the specialized fertility gods, the Vanir, Njörd (Njörðr), his son Freyr, and presumably his daughter, Freyja, came to dwell among the Aesir and to be accepted in their hierarchy.

Odin (Óðinn). The foremost of the Aesir, according to literary sources, was Odin, although his worship, on the evidence of place-names and other records, does not appear to have been widespread. Probably Odin owes his pre-eminence in literary sources to the tradition that he was god of poetry. Various stories are told of the origin of poetry, but it was Odin who brought it to the world of gods. Poetry, called the sacred mead, was first brewed from the blood of a wise god, Kvasir, who was murdered by dwarfs. It later came into the hands of a giant and was stolen by Odin, who flew from the giant's fastness, apparently in the form of an eagle, carrying the sacred mead in his crop in order to regurgitate it in the dwelling of the gods. Because of this story, early poets refer to poetry as "blood of Kvasir" or "Odin's theft."

Odin is more than the god of poetry. He is god of occult wisdom, which, as he himself tells in the *Hávamál*, was acquired by hanging for nine nights on the World Tree. Odin was pierced with a spear; he was a sacrifice to himself. It was probably because of this story that the tree was called Yggdrasill; *i.e.*, "Odin's Horse." Odin was dead, or nearly dead, and seemed to acquire that wisdom that belongs only to the dead, as well as the runes that he carved and painted.

Odin, since he hanged himself, was god of the hanged, and stories are told of men hanged as a sacrifice to him; he was a necromancer, and could make hanged men talk. He is god of heroes and of war, promoting battle between princes. Those who fell in battle were said to go to his castle Valhalla (Valhöll), probably meaning "hall of the slain." There fallen warriors would live in bliss until the Ragnarök, when they would join Odin in his fight against a monstrous wolf. As god of the dead, Odin was accompanied by carrion beasts, two ravens, and two wolves. He had only one eye, for he had sacrificed the other for knowledge.

While place-names suggest that Odin was worshipped

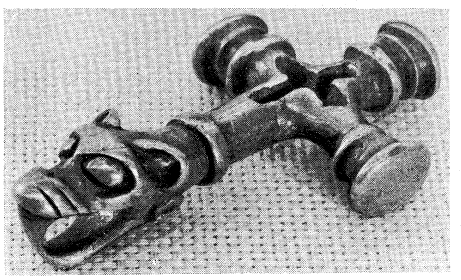
from early times, especially in south Sweden and Denmark, his cult appears to have spread during the Viking Age. He is god of lawless men and vikings, boasting that he himself breaks the most sacred of oaths, the oath on the holy ring. Icelandic poets refer frequently to Odin, but it is doubtful whether he was ever worshipped in Iceland.

English and continental sources show that Odin (Woden, Wotan) was well known in the southern lands. His name is found as an element in many place-names, especially those of England and, as already mentioned, in the royal genealogies of England.

Tacitus describes a god, the *regnator omnium deus* (the god governing all), worshipped by the tribe Semnones in a forest. They would sacrifice a man to him in a horrifying ritual. There are reasons to believe that the *regnator* was Odin, often identified by writers in Latin with Mercury, to whom, according to Tacitus, human sacrifice was offered.

The name Odin is related to Icelandic, *óðr*, German *Wut*, meaning "rage, fury."

By courtesy of the Natioial Museum, Reykjavik; photograph, Gjsli Gestsson



Silver image showing the integration of pagan (Thor's hammer) and Christian (cross) symbols; found in southern Iceland. In the National Museum, Reykjavik.

Thor (*þórr*). Thor is a god of very different stamp. Place-names, personal names, poetry, and prose show that he was worshipped widely, at any rate toward the end of the pagan period. Thor is called the son of Odin, but is in many ways his antithesis. He is god of order and chief antagonist of the giants, demons of chaos. Were it not for him, the whole of our world would be peopled by these monsters, and many stories are told of Thor's contests with them. His chief weapon was his short-handled hammer, of which several images have been found. Among his chief enemies was the World Serpent Jörmungand (Jörmungandr), symbol of evil, who surrounds the world. It is told how Thor failed to smash the skull of this monster; in the Ragnarök they will fight and kill each other.

The World Serpent: Jörmungand

The place-names of England suggest that Thor, or Thunor as he was called in English, was fairly well known, at any rate in Saxon and Jutish areas. As already noted, Thor was among the gods whom continental Saxons renounced on baptism.

Thor was sometimes equated with Jupiter, *dies Jovis* (Jove's day) becoming Thor's Day (Thursday). The name Thor is related to thunder (German *Donner*). Thor travelled in a chariot drawn by goats, and later evidence suggested that thunder was thought of as the sound of this chariot. As god of thunder Thor was god of rainfall and fertility, as is supported by place-names found in eastern Scandinavia and in England.

Balder (*Baldr*). The west Norse sources name another son of Odin, Balder, the spotless, innocent, suffering god. When Balder had dreams foreboding his death, his mother, Frigg, took oaths from all creatures, as well as from fire, water, metals, trees, stones, and illnesses, not to harm Balder. Only the mistletoe was thought too young and slender to take the oath. The guileful Loki tore up the mistletoe and, under his guidance, the blind god Höd (Höðr) hurled it as a shaft through Balder's body. The gods sent an emissary to Hel, goddess of death; she would release Balder if all things would weep for him. All did, except a giantess, who appears to be none other than Loki in disguise. There is another version of this story, to which allusion is made in a west Norse poem (*Baldrs draumar*). According to this Loki does not seem to be directly responsible for Balder's death but Höd alone.

Balder's name occurs rarely in place-names, and it does not appear that his worship was widespread.

The Danish historian Saxo gives an entirely different picture of Balder: he is not the innocent, Christlike figure of the west Norse sources but vicious and lustful, not a god but a demigod. He and Höd were rivals for the hand of Nanna, said in west Norse sources to be the wife of Balder. After many adventures, Höd pierced Balder with a sword. In order to secure vengeance, Odin raped a princess, Rinda (Rindr), who bore a son, Bous, who killed Höd.

Saxo's story has many details in common with the west Norse sources, but his views of Balder were so different that he may have been following a Danish rather than a west Norse tradition. Much of Saxo's story is placed in Denmark.

English sources show that Balder was known among the Scandinavian settlers in England, but there is no strong evidence that he was worshipped by the English themselves. The Merseburg Charm, mentioned above, may suggest that he was worshipped in continental Germany, but those lines are difficult to interpret.

Loki. There is no more baffling figure in Norse mythology than Loki. He is counted among the Aesir, but is not one of them. His father was a giant (Fárbauti; Cruel Striker). Loki begat a female, Angrboda (Angrboða; Distress-Bringer), and produced three evil progeny, the goddess of death, Hel, the evil serpent surrounding the world (Jörmungand) and the wolf Fenrir (Fenrisúlfr), who lies chained until he will break loose in the Ragnarök, to fall upon the dwellings of gods and men. Loki himself lies bound, but will break his bonds in the Ragnarök to join the giants in battle against the gods.

Origin of monsters

Loki deceived the gods and cheated them, but sometimes he got them out of trouble. He is seen in company with Odin and an obscure god Hoenir; he is called the friend of Thor. Loki is bisexual or homosexual; he was said to bear as well as to beget children. He was said to be mother, not father, of Odin's eight-legged horse, Sleipnir. According to one poem, Loki ate the heart of an evil woman and grew pregnant; many evil monsters are thus descended from him.

Like Odin, Loki was a changer of shape; he could take the form of a hawk and, apparently, that of a seal. According to an early poem, Odin and Loki had mixed their blood as foster brothers. It has been suggested that Loki was a hypostasis of Odin, or at least that he represents one side of Odin, the most evil side.

Minor Aesir. Though the Vanir form a major group, some minor gods and goddesses should be mentioned who do not, apparently, belong to the tribe of the Vanir. The god Heimdall (Heimdallr or Heimdalr) is interesting, though obscure. A scald of the 10th century described a struggle between Loki and Heimdall. Snorri relates that in the Ragnarök Loki and Heimdall will kill each other. Heimdall is of mysterious origin; he was the son of nine mothers, said to be sisters, all of whom are named. Heimdall was said to live in Himinbjörg ("Heavenly Rocks"?), merrily drinking his splendid mead. Snorri adds some details, which are supported by the older poetry; he is the watchman of the gods and, when the Ragnarök draws near, he will blow his ringing horn (Gjallarhorn).

There is another myth that, according to many scholars, makes Heimdall the father of men. He came to the world of man and called himself Rigr (Rígr), a name probably derived from the Irish *rí* (king). He consorted with three women, from whom descend the three classes of men—thrall (serf), freeman, and prince.

Germanic religion is judged chiefly from poetry composed late in the pagan period and from remarks of outside observers, who generally had little interest in what they considered to be heathendom. There were many

gods whom these authors have nearly forgotten; Ull was mentioned earlier as one of these. Tyr (Týr) was probably a major god in early times, but memories of him have faded. He is the one-handed god, because one of his hands had been bitten off by the wolf Fenrir. He is brave and warlike; in the Ragnarök he will face the wolf Garm (Garmr), and they will kill each other. Like other gods, Tyr is said to be a son of Odin, but according to one early poem, he was the son of a giant. Tyr's cult is remembered in place-names, particularly those of Denmark. He was equated with Mars and hence *dies Martis* (Mars' day) became "Tuesday" (Icelandic *Týs dagr*).

Tyr

Bragi. Not only Odin but also Bragi was called god of poetry, at any rate in later sources. It is remarkable that the first recorded scald, living in the 9th century, was also called Bragi. Since there is no record of a cult of the god Bragi, some have suspected that the god and the poet are identical.

Frigg. Frigg is remembered as wife of Odin and mother of Balder. She is named as Friia in the second Merseburg Charm. Some writers describe her as unchaste, but more often she is depicted as the weeping mother of Balder. She was sometimes equated with Venus, and her name survives in "Friday" (Old English *Frigedæg*) from *dies Veneris*, Venus' day.

Idun (Iðunn). Idun was said to be wife of Bragi. Little is told of her, but one of the oldest Norwegian scalds refers to a myth that she kept the apples that prevented the gods from growing old. She was raped by a giant and was recovered with her apples by Loki. The story of the apples has many parallels in foreign sources, those in Irish being the closest.

Jörd (Jörð). Jörd, whose name means "earth," was mother of Thor and thus mistress of Odin. She appears under other names, Fjörgyn, which may also mean "earth," and Hlodyn (Hloðyn). A *dea Hludana* is also remembered in votive inscriptions of lower Germany and Holland. When one goddess has several names, it often means that several goddesses have been combined as one.

The Vanir. As mentioned above, the Vanir formed a tribe of gods distinct from the Aesir. If the Aesir were largely gods of battle, the Vanir were gods of fertility and riches, although they could fight too.

The best known Vanir, Njörd, Freyr, and probably Freyja, came as hostages to the Aesir. Njörd was father of the god Freyr and goddess Freyja. Tacitus, in his *Germania* (Ch. XL) described the worship of a goddess, Nerthus, on an island, probably in the Baltic Sea. Her idol was stored in a grove and conveyed in a chariot; when the goddess was present there was rejoicing and peace, and weapons were laid aside.

Although she was a goddess, *Terra Mater* (earth mother), the name Nerthus corresponds with that of the god Njörd (older Nerthuz). The change of sex is hard to explain. Some say that Nerthus was originally a hermaphrodite. Others point out that feminine nouns of the type Njörd practically died out in the Norse language and therefore the original goddess must be seen as a god.

Njörd

Njörd (Njörðr) had been married to his own sister before he came to the Aesir. Since such incestuous unions were not allowed among the Aesir, Njörd afterward married Skadi (Skaði), daughter of a giant. He was said to have immense wealth and, at least in western Scandinavia, he was god of the sea and its riches. Evidence from place-names shows that Njörd was worshipped widely in Sweden and Norway, and he was one of the gods whom Icelanders invoked when they swore their most sacred oaths.

Freyr. Much more is told of Freyr, the son of Njörd. His name probably means "Lord" (compare Old English *Frea*), but Freyr had other names as well; he was called Yngvi or Yngvi-Freyr, and this name suggests that he was the eponymous father of the north Germans whom Tacitus calls Ingvæones (Ingævones). Ing is also named in the Old English *Runic Poem*, where it is said that he was seen first among the eastern Danes; he departed eastward over a wave and his chariot went after him. It is remarkable how the chariot persists in the cult of the Vanir, Nerthus, Ing, Freyr. A comparatively late source tells how the idol of Freyr was carried in a chariot to bring fertility to the crops in Sweden. In an early saga of Iceland, where crops were little cultivated, Freyr still appears as the guardian of a sacred cornfield. Freyr's name often is found as the first element of a place-name, especially in eastern Sweden; the second element often means "cornfield, meadow."

The Eddaic poem *Skírnismál* ("The Words of Skírnir") relates a myth of Freyr taking his bride, Gerd (Gerðr), from the world of giants. Many have seen this story as a fertility myth. Gerdr (from *garðr*, "field") is held fast in the clutches of the frost-giant demons of winter. Freyr is thus the god of the sun.

Several animals were sacred to Freyr, particularly the horse and, because of his great fertility, the boar.

The centre of Freyr's cult was Uppsala, and he was once said to be king of the Swedes. His reign was one of peace and plenty. While Freyr reigned in Sweden a certain Frodi ruled the Danes, and the Danes attributed this age of prosperity to him. Frodi (Fróði) was also conveyed ceremoniously in a chariot, and some have seen him as no other than a doublet of Freyr. Freyr was said to be ancestor of the Ynglingar, the kings of the Swedes; many have regarded these kings as semidivine figures, although this view has lately been challenged.

Freyja. Freyja, the sister of Freyr, resembles her brother in many ways. She was goddess of love and fertility. One of her nicknames was *Syr* or "sow." She was goddess of wealth, owning the famous Brísing necklace mentioned in the *Thrymskviða* ("Lay of Thrym") and many other sources. It was said to be forged by dwarfs. Freyja wept tears of gold. She was also goddess of magic (*seiðr*) and taught this to the Aesir. Freyja has some of the qualities of a death goddess, taking half of those who fall in battle.

Guardian spirits. Besides gods and goddesses, medieval writers frequently allude to female guardian spirits called *dísir* and *fylgjur*. The conceptions underlying these two certainly differed originally, although some of the later writers used the words interchangeably.

Dísir and *fylgjur*

Reference is made several times to sacrifice to the *dísir*, held at the beginning of winter. The sacrifice seems generally to be a private one, suggesting that the *dísir* belonged to one house, one district, or one family. The *dísir* are sometimes described as "dead women," and in actuality they may have been dead female ancestors, assuring the prosperity of their descendants.

There is no record of a cult of the *fylgja* (plural *fylgjur*), a word best translated as "fetch, wraith." The *fylgja* may take the form of a woman or an animal that is rarely seen except in dreams or at the time of death. They may be the companions of one man or of a family and are transferred at death from father to son.

The elves (*álfar*) also stood in fairly close relationship to men. An Icelandic Christian poet of the 11th century described a sacrifice to the elves early in winter among the pagan Swedes. The elves lived in mounds or rocks. An old saga tells how the blood of a bull was smeared on a mound inhabited by elves.

A good deal is told of land spirits (*landvaettir*). The prosperity of the land depends largely upon them. According to the pre-Christian law of Iceland, no one must approach the land in a ship bearing a dragonhead, lest he frighten the land spirits. An Icelandic poet, cursing the king and queen of Norway, enjoined the *landvaettir* to drive them from the land.

Dwarfs. Dwarfs (*dvergar*) play a part in Norse mythology. Snorri says that they originated as maggots in the flesh of the slaughtered giant Ymir. This crude statement finds some support in the older poetic sources, obscure as they are. The dwarfs were very wise and expert craftsmen. Among the treasures they forged was Thor's hammer.

WORSHIP AND PLACES OF WORSHIP

Sacrifice often was conducted in the open or in groves and forests. As already noted, Tacitus speaks of a forest

in which men were sacrificed to the *regnator omnium deus* and gives other examples of groves dedicated to gods. Tacitus does, however, mention temples in Germany, although they were probably few. Old English laws mention fenced places around a stone, tree, or other object of worship. In Scandinavia, men brought sacrifice to groves and waterfalls.

A common word for a holy place in Old English is *hearg* and in Old High German *harug*, occasionally glossed as *lucus* ("a sacred grove") or *nemus* ("forest"). In Old Norse this word has the form *hörgr*. It seems to be applied to a pile of stones, or primitive altar, but was occasionally used for roofed temples. *Vé* (compare *vígja*, "to consecrate") was also applied to sacred places in Scandinavia, and forms an element in many place-names; *e.g.*, Odense (older Óðinsvé).

Temples

Although worship was originally conducted in the open, temples also developed with the art of building. Bede shows that some temples in England were built well enough to be used as churches and mentions a great one that burned.

The word *hof*, commonly applied to temples in the literature of Iceland, seems to belong to the later rather than to the earlier period. A detailed description of a *hof* is given in one of the sagas. It consisted of two compartments, perhaps analogous to the chancel and the nave of a church. The idols were kept in the chancel. If the description is valid, it implies that Icelandic temples of the 10th century were modelled on churches. A building believed to be a temple has been excavated in northern Iceland, and its outline agrees closely with that described in the saga.

Temples on the mainland of Scandinavia were probably built of wood, of which nothing survives, although some see the influence of pagan temples on the so-called stave-churches. At the close of the pagan period, the most splendid temple of all was at Uppsala. It was richly described by Adam of Bremen, though he knew it only by hearsay and may have been influenced in his description of it by the description of the temple of Solomon in the Bible. Idols of Thor, Wodan, and Fricco (Freyr) stood together within it. There were also famous temples in Norway, but detailed descriptions are not given.

Sacrifice took many different forms. Jordanes mentions the sacrifice of prisoners of war among the Goths. The idol of Nerthus was washed by thralls who alone were allowed to see her; immediately afterward they were drowned. A detailed description of a sacrificial feast is given in a saga about a king of Norway. All kinds of cattle were slaughtered, and blood was sprinkled inside and out; the meat was consumed and toasts were drunk to Odin, Njörd, and Freyr. The most detailed description of a sacrifice is that given by Adam of Bremen. Every nine years a great festival was held at Uppsala, and sacrifice was conducted in a sacred grove that stood beside the temple. The victims, human and animal, were hung on trees. One of the trees in this grove was holier than all the others and beneath it lay a well into which a living man would be plunged.

There also were sacrifices of a more private kind. A man might sacrifice an ox to a god or smear an elf mound with bull's blood.

ESCHATOLOGY AND DEATH CUSTOMS

No unified conception of the afterlife is known. Some may have believed that fallen warriors would go to Valhalla to live happily with Odin until the Ragnarök, but it is unlikely that this belief was widespread. Others seemed to believe that there was no afterlife. According to the *Hávamál*, any misfortune was better than to be burnt on a funeral pyre, for a corpse was a useless object.

More often people believed that life went on after death for a time, but was inseparable from the body. If men had been evil in life, they could persecute the living when dead; they might have to be killed a second time or even a third before they were finished.

The presence of ships or boats in graves, and occasionally of chariots and horses, may suggest that men were thought to go on a journey to the Other World, but this is questionable.

Some records show that the dead needed company; a wife, mistress, or servant would be placed in the grave with them. The famous Oseberg grave contained the bones of two women, probably a queen and her servant. Some stories suggest that there was an ancient belief that men might be born again, but this, says a medieval writer, is an old wives' tale.

On the whole, beliefs in death seem negative. Men pass, perhaps by slow stages, to a dark, misty world called Niflheim (Niflheimr).

Cosmic destruction: Twilight of the Gods

The word Ragnarök is found in two forms. The older is Ragnarök, meaning "Fate of the Gods"; the later form, used by Snorri and some others, is Ragnarøkkr, "Twilight of the Gods." Allusions to the Ragnarök, the impending disaster, are made by several scalds of the 10th and 11th centuries, but fuller descriptions of it are given chiefly in the *Völuspá* and the didactic poems of the *Poetic Edda*, which form the basis of Snorri's description in his *Edda*.

Only a brief summary of this rich subject can be attempted here. Through their own work, and especially because of the strength of Thor, gods have kept the demons of destruction at bay. The savage wolf Fenrir is chained, as is the guileful Loki, but they will break loose. Giants and other monsters will attack the world of gods and men from various directions. If we follow the *Völuspá*, the most rational of the sources, Odin will fight the wolf and lose his life, to be avenged by his son *Vidar* (Víðarr) piercing the beast to the heart. Thor will fight the World Serpent and they will kill each other. The sun will turn black, the stars vanish, and fire will play against the firmament. The earth will sink into the sea, but will rise again refreshed. Unsown fields will bear corn. Balder and his innocent slayer, Höd, will return to inhabit the dwellings of gods. There will be a hall, roofed with gold, in which just men will live throughout ages. It is as if men found it impossible to believe in a universe in which nothing at all is left.

Many have seen the influence of Christian thought, and especially of apocryphal Christian literature about the end of the world, in the description given in the *Völuspá* of the Ragnarök. According to another Eddaic poem, the wolf will swallow Odin and, in revenge, his son will tear the jaws of the beast asunder. Several more details are given in other sources, generally cruder than those of the *Völuspá*.

THE END OF PAGANISM

The Germanic peoples were converted to Christianity in different periods, many of the Goths in the 4th century, the English in the 6th and 7th centuries, the Saxons, under force of Frankish arms, in the late 8th century, and the Danes, under German pressure, in the course of the 10th century. The pagan religion held out longest in the most northerly lands, Iceland, Norway, and Sweden.

The story of the conversion of Iceland is known best because of the wealth of historical documents written in that country during the Middle Ages. Icelanders were, in many ways, the most international of northern Scandinavians. Among those who settled in Iceland in the late 9th century were men and women partly of Norse stock from Christian Ireland. Some of these were Christians; some were mixed in their beliefs, worshipping Christ and Thor at once. There were others who believed in no gods at all. Lack of faith in the heathen gods seemed to grow during the 10th century. Influence of Christian thought on some Icelandic poets is noticeable, as is a tendency toward monotheism. Occasional missions to Iceland in the later 10th century are recorded, but little progress was made until Olaf I Trygvason, king of Norway, sent out the German priest Thangbrand *c.* 997. Thangbrand was a ruthless, brutal man; he was outlawed and returned to Norway *c.* 999. But, unlike Norway and Sweden, Iceland was ripe for conversion. In the year after Thangbrand left (*c.* 1000) the Icelandic parliament (Althingi) resolved, at the instigation of King Olaf, that all should be

Conversions to Christianity

baptized, although concessions were made to those who wished to practice heathen rites in private. Many of the hereditary pagan chieftains became leaders of the church and, largely for this reason, tradition survived in Iceland as in no other Scandinavian land.

The conversion of Norway was far less peaceful, and much is known about it, chiefly from Icelandic records, which are highly coloured. Olaf Trygvason had come to Norway from England *c.* 995, and quickly overcame the arch-pagan ruler Haakon Sigurdarson. Paganism was deeply rooted in the minds of hereditary landowners, whose social system depended largely upon it. Olaf converted the whole of Norway in his short reign of five years. He used fire and sword rather than persuasion. When he died in a naval battle, *c.* 1000, many of Olaf's subjects were Christians only in name. When Olaf the Saint came to the throne, about 15 years later, some of the Norwegians were baptized and some not, and everyone believed whatever he liked. St. Olaf completed the work of his namesake, and his methods were of the same kind. He was such a tyrant that his own subjects, Christian though they were, drove him into exile in Russia. When he returned with a motley army, *c.* 1030, he met his death and was soon regarded as a saint. For all his faults, Olaf had established Christianity firmly in Norway.

Very little is known about the conversion of Sweden. It was a slow and complicated process. The people of West Gautland were, apparently, converted earlier than the rest, but public pagan sacrifice persisted in the temple of Uppsala until late in the 11th century. Kings who were Christian were driven out, probably because of their religious activities. Sweden was hardly a Christian country before *c.* 1100.

Germanic paganism was, in many ways, a vague religion. Some were devoted to one god, some to another. Religion was often a personal relationship between a man and his favourite god. Under pagan civilization in the north, the standard of morals was high, but this had little to do with religious belief. An offense against the god would be punished by the god. Other crimes—theft, adultery, seduction—could be prosecuted under the civil law. The law may have originated under pagan religion but, in the sources extant, religion had little to do with morals.

Paganism was a religion without force that could not resist the pressure of positive, monotheistic Christianity, a religion with forceful, definitive creed, embracing morals as well as faith.

BIBLIOGRAPHY

General works: J. GRIMM, *Deutsche Mythologie* (1835; Eng. trans., *Teutonic Mythology*, 4 vol., trans. by J.S. STALLYBRASS, 1883–1900, reprinted 1966), an indispensable source; J. DE VRIES, *Altgermanische Religionsgeschichte*, 2nd ed., 2 vol. (1956–57), a thorough account of Germanic heathendom in Scandinavia, Germany, and England; G. DUMEZIL, *Les Dieux des Germains* (1959), a short account of Germanic mythology seen from the author's viewpoint of the community of Indo-European religion; H.R. ELLIS DAVIDSON, *Gods and Myths of Northern Europe* (1964), a readable sketch, somewhat marred by inaccuracy; E.O.G. TURVILLE-PETRE, *Myth and Religion of the North* (1964), the fullest account in English of Norse myth and religious practice; F. STROM, *Nordisk Hedendom* (1967), a short but thoroughly reliable work; O. BRIEM, *Heiðinn Siður á Íslandi* (1945), an excellently balanced and objective survey of heathendom in Iceland.

Gods and spirits: H.M. CHADWICK, *The Cult of Othin* (1899), a useful and reliable study, though somewhat outdated; E.O.G. TURVILLE-PETRE, *Um Óðinsdýrkun á Íslandi* (1958), a study of the worship of Odin in Iceland; G. NECKEL, *Walhall* (1931), a consideration of the beliefs in Valhalla and their relation to comparable beliefs outside the Germanic world, and *Die Überlieferungen vom Gotte Balder* (1920), an interesting study of Balder, emphasizing similarities with gods of the Near East; H. LJUNGBERG, *Tor I* (1947), interesting ideas about the god Thor and his relations with Lappish, Finnish, and other non-Germanic deities; J. DE VRIES, *The Problem of Loki* (1933), a long and somewhat inconclusive study of this puzzling mythological figure; D. STROMBACK, *Tidrande och Diserna* (1949), a brief and perceptive account of guardian spirits and their powers of good and evil; F. STROM, *Diser, Nornor, Valkyrjor* (1954), a splendid study of beliefs in attendant spirits and other minor deities.

Conversion of the pagans: K. MAURER, *Die Bekehrung des norwegischen Stammes zum Christenthume*, 2 vol. (1855–56), a summary of nearly all the literary sources about the conversion of west Norse peoples to Christianity; H. LJUNGBERG, *Den nordiska Religionen och Kristendomen* (1938), on the relations between Christians and pagans in Scandinavia during the Viking Age; S.U. PALME, *Kristendomens genembrott i Sverige* (1959), a stimulating account of the conversion of the Swedes.

(E.O.G.T.-P.)

Germans, Ancient

The Germanic, or Teutonic, peoples are a branch of the Indo-Europeans. Their origin is a problem of profound obscurity. A major cause of the difficulty is the paucity of archaeological finds relating to them in northern Germany and southern Sweden between the end of the Bronze Age (*c.* 500–400 BC) and the 2nd century BC. It may be supposed, however, that in the Late Bronze Age Germanic peoples inhabited southern Sweden, the Danish peninsula, and northern Germany between the Ems and Oder rivers and the Harz Mountains. The Vandals, Gepidae, and Goths migrated from southern Sweden in the closing centuries BC and occupied the area of the southern Baltic coast between the Oder and the Vistula and even beyond to the Passarge (Pasłęka) River. At an early date there was also migration toward the south and west at the expense of the Celtic peoples who then inhabited much of western Germany; the Helvetii, for example, who were confined to Switzerland in the 1st century BC, had once extended as far as the Main River.

By the time of Julius Caesar, Germans were established west of the Rhine and had reached the Danube in the south. Their first great clash with Romans came at the end of the 2nd century BC, when the Cimbri and Teutoni (Teutones) invaded southern Gaul and northern Italy and were annihilated by Marius in 102 and 101. Although individual travellers from the time of Pytheas onward had visited Teutonic countries in the north, it was not until the middle of the 1st century BC that the Romans learned to distinguish precisely between the Germans and the Celts, a distinction that is made with great clarity by Julius Caesar. It was Caesar who incorporated within the frontiers of the Roman Empire those Germans who had penetrated west of the Rhine, and it is he who provides the earliest extant description of Germanic culture. In 9 BC the Romans pushed their frontier eastward from the Rhine to the Elbe, but in AD 9 a revolt of their subject Germans headed by Arminius ended in the destruction of the occupying army of P. Quinctilius Varus in the Teutoburg Forest and in the withdrawal of the Roman frontier to the Rhine. In this period of occupation and during the numerous wars fought between Rome and the Germans in the 1st century AD, enormous quantities of information about the Germans reached Rome. When Tacitus published in AD 98 the book now known as the *Germania*, he had reliable sources of information on which to draw. The book is one of the most valuable ethnographical works in existence; archaeology has in many ways supplemented Tacitus' information but in general has tended only to confirm his accuracy and to illustrate his insight into his subject.

Tacitus' *Germania*

Tacitus relates that according to their ancient songs the Germans were descended from the three sons of Mannus (perhaps "Man") the son of the god Tuisto, the son of Earth. Hence they were divided into three groups: the Ingaevones, the Herminones, and the Istaevones. The first of these groups embraced the peoples of the northwest, the second those of the interior, and the third, probably, those near the Rhine. These names, which are also mentioned by the elder Pliny and in a 6th-century Frankish document, are clearly of great antiquity, but Tacitus gives no further information about them. What this grouping amounted to in practice or what it was originally based on is unknown. Indeed the historian records a variant form of the genealogy according to which Mannus had a larger number of sons,

who were regarded as the ancestors of the Suebi, the Vandals, and others. At any rate the existence of these poems suggests that in Tacitus' time the various Germanic peoples were conscious of their relationship with one another. Although individual Germans in Roman service would sometimes refer to themselves as Germani, the free Germans beyond the Rhine had no collective name for themselves until the 11th century AD, when the adjective *diutisc* ("popular," modern *deutsch*) came into fashion. A famous sentence in which Tacitus attempts to account for the origin of the word Germani as a generic term has been interminably discussed. Its general sense is that Germani was originally the name of one Germanic people, the people later known as the Tungri. The Tungri first crossed the Rhine at an unknown date into the Meuse Valley (hence the name Tongeren or Tongres, Belgium) and expelled the Gauls there from their habitations. Their name, which was Germani at that time, was applied to all their fellow countrymen. (Similarly, the Hellenes were called Graeci by the Romans, perhaps after a small people in Epirus, and the French call the Germans *les allemands* after the Alemanni.) The meaning of the word Germani and even of the language to which it belongs have never been determined.

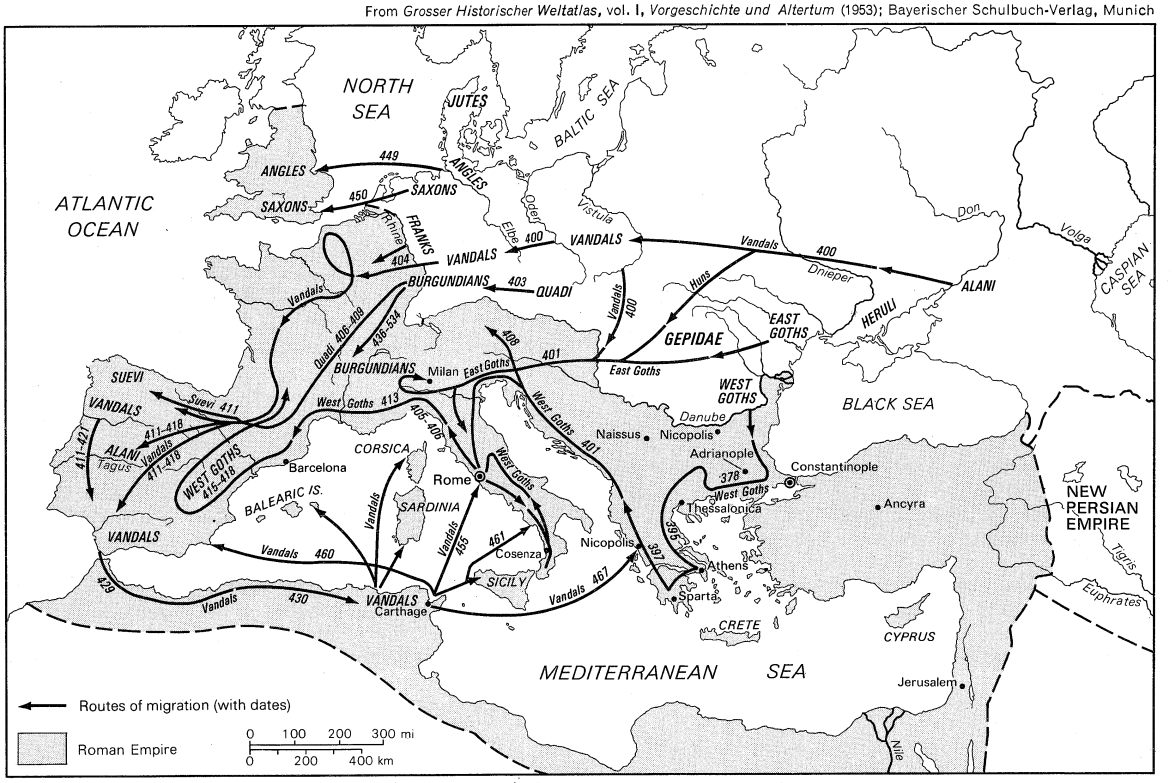
Distribution. The principal Germanic peoples were distributed as follows in the time of Tacitus. The Chatti lived in what is now Hesse. The Frisii inhabited the coastlands between the Rhine and the Ems. The Chauci were at the mouth of the Weser, and south of them lived the Cherusci, the people of Arminius. The Suebi, who have given their name to Schwaben, were a group of peoples inhabiting Mecklenburg, Brandenburg, Saxony, and Thuringia; the Semnones, living around the Havel and the Spree rivers, were a Suebic people, as were the Langobardi (Lombards) who lived northwest of the Semnones. Among the seven peoples who worshipped the goddess Nerthus were the Angli (Angles), centred on the peninsula of Angeln in eastern Schleswig. As for the Danubian frontier of the Roman Empire, the Hermunduri extended from the neighbourhood of Regensburg northward through Franconia to Thuringia. The Marcomanni, who had previously lived in the Main Valley, migrated during the last decade BC to Bohemia (which had hitherto been occupied by a Celtic people called the Boii). To the east were the Quadi in Moravia. On the lower Danube were a people called the Bastarnae, who are usually thought to have been Germans. The Goths, Gepidae, and Vandals on the Baltic coast have already been mentioned. Tacitus mentions the Suiones and the Sitones as living in Sweden. He also speaks of several other peoples of less historical importance than those listed here but he knew nothing of the Saxons, the Burgundians, the Franks, and others who became prominent after his time.

New tribes

By the end of the 3rd century AD important changes had taken place. East of the Rhine lived three great confederacies of peoples unknown to Tacitus. The Roman frontier on the lower Rhine now faced the Franks. The Main Valley was occupied from *c.* 260 by the Burgundians, while the *agri decumates* (the area south of the Main River and immediately east of the Rhine) were held by the Alemanni. The Burgundians appear to have been immigrants from eastern Germany. The Franks and the Alemanni may have been confederacies of peoples who had lived in these respective areas in Tacitus' day, though perhaps with an admixture of immigrants from the east. The peoples whom Tacitus mentions as living on the Baltic coast had moved southeastward in the second half of the 2nd century. The Goths now controlled the Ukraine and a large section of the present-day country of Romania; the Gepidae were in the mountains north of Transylvania with the Vandals as their neighbours on the west.

By the year 500 further striking changes had taken place. The Angles and Saxons were in England and the Franks controlled northeastern Gaul. The Burgundians were in the Rhône Valley with the Visigoths as their western neighbours. The Ostrogoths were established in Italy and the Vandals in Africa. In 507 the Franks expelled the Visigoths from most of their Gallic possessions, which had stretched from the Pyrenees to the Loire River, and the Visigoths thereafter lived in Spain until their extinction by the Muslims in 711. In 568 the Lombards entered Italy and lived there in an independent kingdom until they were overthrown by Charlemagne (774). The areas of eastern Germany vacated by the Goths and others were filled up by the Slavs, who ex-

From *Grosser Historischer Weltatlas*, vol. I, *Vorgeschichte und Altertum* (1953); Bayerischer Schulbuch-Verlag, Munich



Distribution and migration of Germanic tribes (AD 378–*c.* 470).

tended westward as far as Bohemia and the basin of the Elbe. After the 8th century the Germans recovered eastern Germany, lower Austria, and much of Styria and Carinthia from the Slavs.

Material civilization. According to Julius Caesar the Germans were not primarily agriculturists; they were pastoralists, and the bulk of their foodstuffs—milk, cheese, and meat—came from their flocks and herds. But agriculture was not unknown, and Caesar's description of their system of land tenure has been much debated. His words are:

Land tenure

> No one has a fixed measure of land as private property; but the magistrates and leading men distribute every year to the clans and groups of kinsmen [what follows in the Latin text is corrupt; the words further defined the kinsmen] as much land and in such place as has seemed right to them, and in the following year force them to move on.

It does not follow from this description that any given piece of land was tilled for one year only and was thereafter deserted by the community; it cannot be inferred that the entire community abandoned its dwellings and moved on each year to completely new habitations after reaping their crops. Such a type of migratory agriculture had almost completely disappeared from Europe at a far earlier date, and such a hypothesis could hardly be reconciled with the archaeological evidence, which shows German settlements continuously inhabited for very long periods of time. The correct inference would seem to be that the same piece of land was used by the same community indefinitely but that the arable land was redistributed annually so that the one kindred did not till the one piece of land for more than a year. A further inference is that land was not privately owned in the middle of the 1st century BC but was in some sense the property of the community: in the case of the Suebi, Caesar emphasizes this point. What exactly is to be understood by the "community" in this context, Caesar does not make clear. Is he speaking of villages or of *pagi* (these appear to have been adjacent groups of kindreds, but information about them is very meagre) or of some other unit? At all events, it seems certain that the leading men distributed each piece of arable land not to individual cultivators but to kindreds, which tilled their allotment in common. This system of annual redistribution ensured the general equality of the kindreds insofar as agriculture was concerned, and Caesar comments that it guaranteed that the masses of the warriors would be contented and disinclined to engage in factional strife and revolt.

As regards Tacitus' description of Germanic land tenure, it must be remembered that he is discussing the situation as it existed more than 100 years after Caesar had been on the Rhine and at a date when the proximity of Roman civilization had possibly brought about considerable changes in Germanic society. His words are:

> Lands proportionate in extent to the number of cultivators are occupied in turn by the whole body of them, and these they then divide among themselves according to social standing: distribution is easy owing to the wide expanse of ground. They change their arable every year and yet there is land in abundance.

Whatever is to be made of this, one point seems clear: although the annual redistribution mentioned by Caesar was still practiced, the arable land was now allotted not to kindreds but to individuals—the term social standing (*dignationem*) makes this certain. In other words, although the private ownership of land was not yet common in Germany, it had come perceptibly closer than it had been in Caesar's day and individual cases of private ownership undoubtedly existed. The kindred, at any rate in matters relating to agriculture, was no longer the basic unit of Germanic society; its place had been taken by the individual. When Tacitus goes on to say that grain alone was grown by the Germans, he is in error; a great variety of root crops and vegetables was known to them, though of fruits only the apple was cultivated by them in the Roman period. As for the techniques of agriculture, Caesar and Tacitus are silent on these, but even the Bronze Age rock carvings at Bohuslän in southern Sweden include a picture of an ox-drawn plow.

It must be stressed, however, that cattle were the main source of food for the Germans. There is no reason to think that in the historical period cattle were owned by the clan collectively; they were the private property of individuals, and a man's status was reckoned on the basis of the number of cattle he owned. Both the cattle and the horses of the Germans were of poor quality by Roman standards.

The Iron Age had begun in Germany about four centuries before the days of Caesar, but even in his time metal appears to have been a luxury material for domestic utensils, most of which were made of wood, leather, or clay. Of the larger metal objects used by them, most were still made of bronze, though this was not the case with weapons. Pottery was for the most part still made by hand and pots turned on the wheel were distinctly unusual.

Trade

The degree to which trade was developed in early Germany is very obscure. There was certainly a slave trade and many slaves were sold to the Romans. Such potters as used the wheel—and these were relatively few—and smiths and miners no doubt sold their products. But in general the average Germanic village is unlikely to have used many objects that had not been made at home. Foreign merchants dealing in Italian as well as Celtic wares were active in Germany in Caesar's time and supplied prosperous warriors with wine, bronze vessels, and so on. But from the reign of Augustus onward there was a huge increase in German imports from the Roman Empire. The German leaders were now able to buy whole categories of goods—glass vessels, red tableware, Roman weapons, brooches, statuettes, ornaments of various kinds, and other objects—that had not reached them before. These Roman products brought their owners much prestige, but how the Germans paid for them is not fully known. The amber trade, however, became important after the middle of the 1st century BC, though for the most part it affected eastern Germany alone.

Form of government. No trace of autocracy can be found among the Germans whom Caesar describes. The leading men of the *pagi* would try to patch up such disputes as arose but they acted only in those disputes that broke out between members of their own *pagus*. When disputes arose between persons belonging to different *pagi* there appears to have been no mediatory body at this date. In fact in peacetime there seems to have been no central authority that could issue orders to, or exercise influence over, all the *pagi* of which any one people (*civitas, Stamm*) was composed. Evidently the clans or groups of clans were more or less independent of one another in internal affairs, and on at least some occasions they could act independently even in their foreign relations. Some clans or *pagi* could, for example, negotiate with the Romans without reference to the other clans of which their people was composed. Caesar relates that in wartime a number of confederate chieftains were elected, but he gives no indication that any one of them had greater authority than the others; they were joint leaders and they held office only in time of war. It was still the case among some Germanic peoples in the 4th century AD and even later there was normally no one overall peacetime chieftain. Nor was the multiple war leadership quick to disappear in all parts of Germany; it could still be found among the Franks, Burgundians, and Alemanni in the 4th and 5th centuries.

Military chieftains

A new type of military chieftainship had come into being by the time Tacitus wrote. This type of chieftain was also elected, but not all the warriors were eligible for election. The office was confined to the members of a recognized "royal clan," such as is known to have existed among the 1st-century Cherusci and Batavians and the 6th-century Heruli. Any member of this royal clan was eligible for election and the chieftainship was in no way hereditary. There is no conclusive evidence that any Germanic chieftain of the 1st century AD was followed as leader of his people by his son. A chief of this type held

office for life and had religious as well as military duties, but he was in no sense an autocrat. He could be overruled by the council of the leading men, and the proposals he laid before the general assembly of the warriors might be rejected by them. On the battlefield, too, he had no powers of coercion; he could only set an example and give advice. The degree of his influence, in fact, depended largely on his own personal qualities. Among the peoples who retained the type of chieftainship described by Caesar, it became common to appoint two leaders to hold office simultaneously, but the chief of the type mentioned by Tacitus had no colleague.

The council of the leading men (*principes*) dealt with matters of minor importance affecting the people as a whole, while the most weighty business was decided by the general assembly of the warriors, though this too received preliminary consideration in the council of the leading men. Little is known of how these leading men were appointed or of their numbers or whether each *pagus* necessarily had one or more representatives on the council. As for the general assembly, all the warriors had the right to attend its meetings except those who had thrown away their shields in battle. The assembly could not initiate measures nor does it seem to have had the right to debate any proposals put to it by the chief or the leading men; it could only adopt or reject their proposals, but its decision was final. It is not known what happened when opinion was divided in the assembly, but certainly there were no means by which a substantial minority of the warriors could have been coerced into a course of action of which they strongly disapproved. The assembly exercised judicial functions in cases that were felt to affect the community as a whole. Tacitus lists treachery, desertion, cowardice, and homosexuality as examples, and for all these offenses the penalty was death.

As for the administration of justice in general, it has been seen that in Caesar's day the leading men of each *pagus* tried to settle disputes that arose within the *pagus*. There is no evidence that they could oblige the disputants to appear before them or that they could enforce their decisions; they could, it seems, only use their influence. But in Tacitus' time the general assembly elected a number of the leading men to act as judges and these judges travelled through the villages to hear private suits. Each of them was accompanied by 100 attendants to lend authority to his decisions. That is to say, the judge now came into the community often as a stranger to those among whom he was to adjudicate. A rudimentary judicial apparatus had come into existence. A person who was found guilty by these judges had to pay a number of horses or cattle proportionate to the gravity of his offense. But many disputes—*e.g.*, those arising from homicide, wounding, or theft—continued to be settled by the kindreds themselves, and the blood feuds to which they gave rise might continue from generation to generation. Long after the conversion to Christianity the German rulers found it difficult to stamp out the blood feud.

Growth of the *comitatus*

It is possible, then, to detect a considerable development in Germanic society between the time of Caesar and that of Tacitus. The rudiments of the state had now made their appearance. Power, insofar as it existed at all, was tending to become concentrated and wealth to accumulate in private hands. Perhaps the most remarkable development concerns the "retinue" (*comitatus*). In Caesar's day one of the leading men would announce in the assembly that he proposed to undertake a foray and would call for followers. Whoever was attracted by the proposal would volunteer his services and when he had done so public opinion would not allow him to withdraw. The relationship between the leader and his followers was a purely temporary one, lasting only for the duration of the raid, and the followers could not be described as dependents of the leader. But in the time of which Tacitus speaks the relationship between the leader and his "companions" (*comites*) had become a permanent one. The leader fed them and kept them about him in peacetime as well as in war. He supplied them with their weapons and horses and with a share in the booty taken during their raids, though in these early times he could not supply them with land, for full private ownership of land hardly existed as yet. The retinue leader thus acquired a military force over which the other warriors had little or no control, and his followers were prepared to fight for him to the death—it was a disgrace for them to survive their leader. The members of the retinues seem nearly always to have been drawn from among the more well-to-do warriors, so that in Tacitus' time the tribal aristocracy appeared to be well on the way to overthrowing the primitive democracy described above and to establishing something like state power among the various peoples. But in fact only one Germanic chieftain is known who was able before AD 100 to set up a personal tyranny over his people. This was Maroboduus, who led the Marcomanni from their homes in the Main Valley *c.* 9 BC and settled them in Bohemia. From there he conquered a considerable number of other Germanic peoples between the Elbe and the Vistula, including the Semnones, the Lombards, and the Lugii. But the Cherusci, joined by some of the king's subjects, attacked him in AD 17, overthrew him, and drove him into Roman territory. All other chiefs who attempted in this period to establish monarchies were, so far as is known, defeated.

Maroboduus' kingdom

Many of the peoples who had been prominent in western Germany in the days of Tacitus disappeared from history during the 2nd century and their place was taken by the Franks, Burgundians, Alemanni, and others. Sources for their internal and social history during the 3rd and later centuries are very fragmentary. In general it is not easy to detect any substantial difference in their material civilization or social organization from what Tacitus had described, except that there is little or no evidence for the existence of the general assembly of the warriors among any of the great peoples living along the Rhine and the Danube. Information about the Visigoths is more abundant than information about any of the other peoples and yet nothing is said by the ancient authorities about the Visigothic warriors having the right to assemble together in order to approve or veto the recommendations of their leaders. This suggests that in the 4th century the rank and file of the population had less control over their own affairs than they had had 300 years earlier. Even so, there are few reports in this period of the overthrow of such democratic institutions as still existed or of the establishment of monarchies. The only certain example is that of Ermanaric who ruled the Ostrogoths in the Ukraine as a king in the middle of the 4th century. The monarchy did not become fully established in the Germanic world until German peoples had settled as federates inside the Roman Empire, after which the leaders of the Ostrogoths in Italy, the Visigoths in Gaul and Spain, and the Vandals in Africa became the first Germanic kings. Other famous German chieftains in this period, such as Athanaric and Alaric who lived outside the Roman frontier or whose peoples were not federates settled in the provinces under a treaty (*foedus*) to defend the frontier, seem to have had little more personal authority than the war leaders described by Tacitus.

Warfare. In the period of the early Roman Empire German weapons, both offensive and defensive, were characterized by shortage of metal. The warrior, whether mounted or on foot, had as his chief weapon a long lance with one end hardened by fire or else fitted with a short, narrow iron point, which could be hurled or used for thrusting. Very few carried swords. Helmets and breastplates were almost unknown and the German went into battle naked or wearing a short cloak. His only defensive weapon was a light wooden or wicker shield, sometimes fitted with an iron rim and sometimes strengthened with leather. This lack of adequate equipment explains the swift, fierce rush with which the Germans would charge the ranks of the heavily armed Romans. Their only hope of overwhelming a Roman army in open country was to break it by the impetus of their first attack, for if they became entangled in a prolonged, hand-to-hand grapple, where their light shields and thrusting spears were confronted with Roman arms and armour, they had little

Weaponry

hope of success. Their best plan was to catch the Romans on an open plain surrounded by woods and then to launch incessant, short, sharp attacks on them from all sides, using the woods as cover.

In the period of the later Roman Empire the equipment of the German warriors does not seem to have improved very much. In the 3rd century they began to use bows and arrows extensively. Their armies were still mainly infantry forces though the tribal nobles, as previously, were mounted. Even in the 6th century there is no reason to think that helmets and breastplates were not something of a rarity in the armies of the Germanic peoples who were now living in what had been the Roman provinces. The Ostrogothic army in Italy, which fought against Belisarius, was an army of mounted spearmen supported by unmounted archers. Such, too, were the army of the Vandals in Africa in the 6th century and the army of the Visigoths in Spain in the 6th and 7th centuries. These peoples seem to have had a higher percentage of cavalry in their armed forces than the Germans had had in Tacitus' day, but in defensive armour little progress had been made. The main strength of the Frankish army lay not in mounted spearmen but in its infantry, whose characteristic weapon was a short-handled, double-headed ax used as a missile weapon. They, too, wore little defensive armour. None of these peoples evolved a military force adequate to deal with the heavily armed mounted archers of Justinian I. Nor did they master the art of siege warfare; throughout the first six centuries AD there is hardly a single report of a Roman town skillfully and successfully besieged by a German army (the Fall of Philippopolis in 250–251 was due to the treachery of the governor).

Conversion to Christianity. Christianity was first brought to the Visigoths by some of the prisoners whom they took during their raids on Anatolia in the mid-3rd century. The Visigoths do not seem to have been converted to Christianity until they were living as federates in the Roman province of Moesia (south of the lower Danube) in the years 382–395. The evidence suggests that before the fall of the Western Roman Empire in 476 none of the great Germanic peoples was converted to Christianity while still living outside the Roman frontier but that all the Germanic peoples who moved into the Roman provinces before that date were converted to Christianity within a generation. The Vandals seem to have been converted when in Spain in 409–429, the Burgundians when in eastern Gaul in 412–436, and the Ostrogoths when in the province of Pannonia *c.* 456–472. The only certain exception is the people known as the Rugii, who were already Christian before 482 while living north of the Danube in lower Austria; in what circumstances they had been converted is not known. In all these cases it seems likely that the conversion was carried through by German-speaking and not by Roman missionaries, and Visigothic priests are likely to have played a major part in the process.

In all these cases the Germans embraced the Arian form of Christianity; none of the major Germanic people became officially Catholic until the conversion of the Franks under Clovis (496) and of the Burgundians under Sigismund. The reason for their adoption of Arianism rather than Catholicism is very obscure. Unhappily, the books produced by the Arian Germans have all disappeared with the exception of the fragments of Ulfilas' Bible, some leaves of an anonymous Gothic commentary on St. John's Gospel, and a fragment of a church calendar written in Gothic. It is clear, however, that their theology depended on a literal interpretation of the Scriptures and refrained from drawing practical lessons from the texts. Although the pagan Germans living outside the Roman frontier once or twice persecuted the Christians in their midst, the Arian German kings, with the exception of the Vandal kings of Africa, were extraordinarily tolerant both of Catholics and of Jews.

The last Germanic people on the continent to be converted to Christianity were the Old Saxons (second half of the 8th century), while the Scandinavian peoples were converted in the 10th century, though several districts remained pagan until late in the 11th century. England had been converted in the 7th century, while Christianity obtained public recognition in Iceland in the year 1000.

BIBLIOGRAPHY. For the older archaeological evidence, see especially the excellent bibliography to G. EKHOLM, *The Cambridge Ancient History*, vol. 11, ch. 2, sect. 1–5 (1936); and for more recent finds, consult the periodical *Germania* (semi-annual). On Roman objects of trade imported into Germany, HANS J. EGGERS, *Der römische Import im freien Germanien* (1951), is fundamental. On the finds of Roman coins, see HANS GEBHART and KONRAD KRAFT (eds.), *Die Fundmünzen der römischen Zeit in Deutschland* (1960–). On German weapons and armour, see MARTIN JAHN, *Die Bewaffnung der Germanen . . .* (1916). The best commentary on Tacitus' *Germania* is that of RUDOLF MUCH, 3rd ed. (1967); see also EDUARD NORDEN, *Die germanische Urgeschichte in Tacitus Germania*, 4th ed. (1959), and *Alt-Germanien*, 2nd ed. (1962); and on the social history of the early Roman period, E.A. THOMPSON, *The Early Germans* (1965). For the period of the later Roman Empire, see OTTO SEECK, *Geschichte des Untergangs der antiken Welt*, 6 vol. (1895–1921, reprinted 1966); J.B. BURY, *History of the Later Roman Empire from the Death of Theodosius I. to the Death of Justinian*, 2 vol. (1923, reprinted 1958); and *The Invasion of Europe by the Barbarians* (1928, reprinted 1967); ERNST STEIN, *Histoire du Bas-Empire*, 2 vol. (1949–59); and above all LUDWIG SCHMIDT, *Geschichte der deutschen Stämme . . .*, 2nd rev. ed., 2 vol. (1941–42). For the conversion to Christianity, see KURT D. SCHMIDT, *Die Bekehrung der Germanen zum Christentum*, 2 vol. (1935–39); and HEINZ E. GIESECKE, *Die Ostgermanen und der Arianismus* (1939).

(E.A.T.)

Germany, Federal Republic of

Commonly referred to as West Germany, the Federal Republic of Germany (Bundesrepublik Deutschland) is a federation of 10 *Länder* (states) plus the territory of West Berlin. It is a self-governing republic created in 1949 through the amalgamation of the British, United States, and French zones of occupation that were established following the surrender of the German *Reich* in May 1945.

Location and boundaries of the Federal Republic

West Germany occupies the regions of northwest central Europe, extending from the upland approaches of the Northern Alps in the south to the North Sea, the Jutland Peninsula, and the western Baltic in the north. Except for a stretch of the Upper Rhine River extending south of Karlsruhe to form part of its border with eastern France and continuing due east of Basel to form a portion of its boundary with Switzerland, the republic is largely lacking in natural frontiers with the eight countries on which it borders. To these eight neighbours must be added the separate state created out of the former *Reich*, the German Democratic Republic (Deutsche Demokratische Republik), commonly known as East Germany, the former Soviet zone of occupation.

The Federal Republic is bounded at its extreme north on the Jutland Peninsula by Denmark. To its west it borders on The Netherlands, Belgium, and Luxembourg; to the southwest, with France. It shares its entire southern boundaries with Austria and Switzerland. Its eastern frontiers, continuing northward with Austria and past Czechoslovakia, recede to the west at the juncture of Czechoslovakia and East Germany and meander northward in a common border with its rival German state as far as the Baltic Sea.

Until recently, West Germany held itself to be the sole successor to the former *Reich* and alone entitled to act in the name of the latter's defunct government. It is linked politically to the three Western sectors of Berlin, the former capital of the *Reich*, technically under quadripartite occupation by the British, U.S., French, and Soviet forces since the end of World War II and at present an enclave within the German Democratic Republic. Connected to West Germany by precisely defined overland, railway, water, and air routes, West Berlin by treaty agreement is no longer claimed by the Federal Republic as a part of its constituent territory but rather as an administrative adjunct, whose inhabitants are counted for all purposes—except for their franchise to

vote in federal German elections—as citizens of West Germany.

Backgrounds of the federal *Länder*

The 96,094 square miles (248,882 square kilometres) of the German Federal Republic, including the 185 square miles of West Berlin, had a population of over 61,280,000 by mid-year 1971; the capital is at Bonn. The constituent *Länder* follow roughly the historic lines of the political divisions making up the former *Reich*. Most of these date back to consolidations brought about largely by the Napoleonic Wars and the founding of the *Reich* in 1871 out of an intricate maze of kingdoms, principalities, duchies, bishoprics, free cities, and the like that existed under the Holy Roman Empire. The *Länder* enjoy considerable political autonomy within the federative structure, especially in such areas as education, finance, and law enforcement; and each has its own equivalent of a prime minister, a parliament or diet, and its own provincial ministries. Their strong voice in the upper levels of the federal government is unique among Western republics.

The largest of the *Länder* is Bayern (English Bavaria), the wealthiest and most populous, Nordrhein-Westfalen (North Rhine-Westphalia). In the extreme north lies Schleswig-Holstein, south of which is Niedersachsen (Lower Saxony). The ancient free Hanseatic cities of Hamburg and Bremen rank as *Länder* in their own right. The states of Rheinland-Pfalz (Rhineland-Palatinate) and Hessen (Hesse) cover the central area of the Federal Republic; the mineral-rich Saarland (or often the Saar), a pocket in the southwestern corner of the Pfälzerwald, was an administrative unit of France after 1945 but reverted to Germany in 1957. In 1951 the small southwestern states of Baden and Württemberg with Hohenzollern were amalgamated into the *Land* Baden-Württemberg.

West Germany's recovery from its total economic and political prostration at the end of World War II and the devastation of its cities and capital industries have been of such dramatic proportions as to become a modern legend. The *Wirtschaftswunder* ("economic miracle") of the 1950s had by the early 1970s catapulted Germany into the position of the world's fourth-largest industrial power, after the U.S., the Soviet Union, and Japan. Its currency, the Deutsche Mark, had become so strong that by 1972 it had been revalued on four occasions, to an increase of almost 25 percent of its value in relation to the U.S. dollar.

Development of a national identity

Domestically, the fledgling West German state, with only the abortive 14 years of the Weimar Republic (1919–33) to serve as a precedent, developed into one of the most stable of the Western democracies. As it gradually relaxed its predilection inherited from the past for the authoritarian or, at best, paternalistic leadership of elderly statesmen, a younger and more heterogeneous cast of political leaders won the public confidence, rising to power through maturation in political service throughout the republic's postwar growth or, often, through their technical expertise. The three major political parties largely have overcome their origins along class, regional, or sectarian lines. In large measure because of the 5 percent of votes required for representation in the federal parliament and provincial diets, extremist parties of the right and left call forth little parliamentary representation or significant popular support.

A careful and intricate balance of power between the federal government and the historically divisive regional interests, now in the custody of the *Länder*, has all but eliminated the factionalism that for centuries was Germany's greatest obstacle to unity. The West Germans as citizens have moved far from their accustomed apathy or even perverse boast of being apolitical to becoming intensely involved and well informed on the issues, domestic and foreign, that affect their personal lives. Of the some 42,000,000 eligible voters, roughly 75 to 90 percent go to the polls in the national elections.

From the ignominy of its position in the early postwar years, the Federal Republic has grown from strength to strength and now commands a status of great respect and, through its economic might, of no small power in the councils of the world. It has undertaken a series of independent diplomatic initiatives, the most significant of which have been aimed at a political rapprochement not only with its adversary German state but also with the Soviet Union and other nations of the eastern European bloc.

The Federal Republic is a major force in the North Atlantic Treaty Organization (NATO), of which it became a member upon achievement of full sovereignty in 1955. Apart from the other intra-European organizations to which it belongs, it is a charter member of the European Coal and Steel Community (ECSC) of 1952, the European Atomic Energy Community (Euratom), and the European Economic Community (EEC), both dating from 1957. The anomalous position of both East and West Germany in international law has long prevented either from becoming a formal member of the United Nations, but it has sent a permanent observer and participates in the working agencies and organizations of the U.N.

The history of both West and East Germany is covered in GERMANY, HISTORY OF, while that of the region in the ancient world is in GERMANS, ANCIENT. The article EUROPE places West Germany in its geographical and cultural contexts, while separate articles focus more closely on the life and institutions of each of the West German *Länder* (see under their German spellings; *e.g.*, BAYERN). See also the articles BERLIN; COLOGNE; FRANKFURT AM MAIN; MUNICH; and GERMAN DEMOCRATIC REPUBLIC. German contributions to the arts may be found in such articles as LITERATURE, WESTERN; MUSIC, WESTERN; MODERN DANCE; and VISUAL ARTS, WESTERN.

This article is divided into the following major sections:

I. The natural and human landscape
  Features of the natural environment
  Character of the traditional regions
  Patterns of human settlement
II. The people of West Germany
  Diverse groupings within the population
  Contemporary demography
III. The national economy
  Extent and distribution of resources
  Sources of national income
  Management of the economy
  Contemporary economic policies
  Economic problems and prospects
IV. Transportation
  Patterns of traffic movement
  Components of the systems
V. Administration and social conditions
  Structure of government
  Political institutions
  The armed forces
  The social milieu
VI. Cultural life and institutions
  The cultural milieu
  Current state of artistic production
  Cultural institutions
  The communications media
VII. Prospects

I. The natural and human landscape

FEATURES OF THE NATURAL ENVIRONMENT

In its major topographical features, West Germany shares with the countries on its eastern and western borders a profile that ascends in altitude from north to south. The North German Plain, or Lowland, comprising approximately the northern third of the republic, is but a segment of a vast lowland coastal area extending from extreme southwestern France across the Low Countries and throughout the Baltic areas to the Soviet Union. The central third of the territory, commonly known as the Central German Uplands, is a complex system of forested lowland mountain ranges, river basins, and plateaus. These follow a sequence of upland regions that range in a wide arc from the Massif Central of south central France, through Belgium and Luxembourg, across West Germany and the southern portion of East Germany, and through the entirety of Czechoslovakia to the Carpathian Mountains of eastern Europe. The southern third of Germany, although in its outer appearance often similar to the Central Uplands, is, in effect, a northeast-

The upland regions

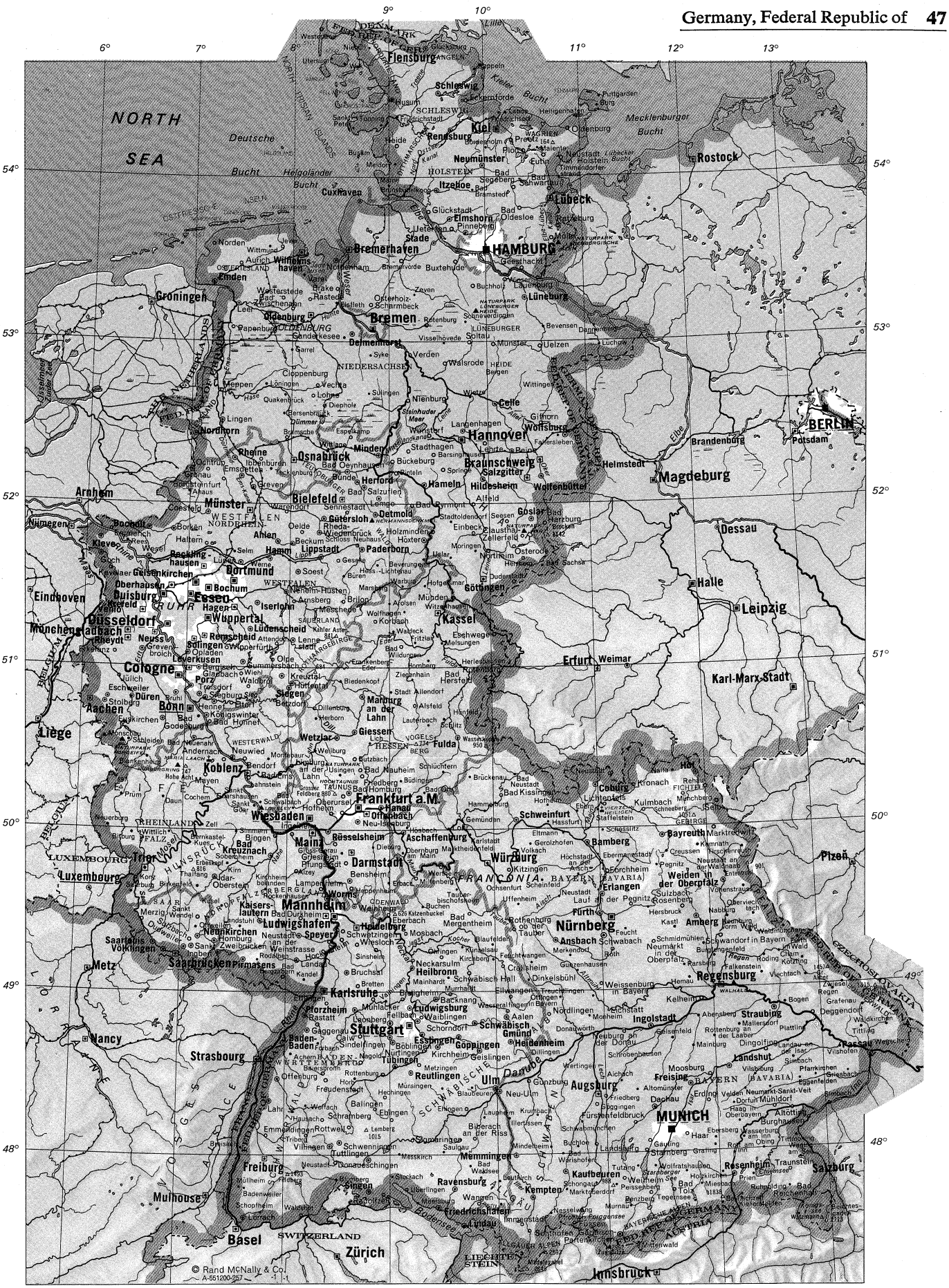
MAP INDEX

Political subdivisions

Baden-Württemberg 48·30n 9·00e
Bayern (Bavaria) 49·00n 11·30e
Bremen 53·05n 8·50e
Hamburg 53·35n 10·00e
Hessen 50·30n 9·15e
Niedersachsen 52·50n 9·00e
Nordrhein-Westfalen 51·30n 7·30e
Rheinland-Pfalz 50·00n 7·00e
Saarland 49·20n 7·00e
Schleswig-Holstein 54·00n 10·30e
West Berlin 52·30n 13·15e

Cities and towns

Aachen 50·47n 6·05e
Aalen 48·50n 10·05e
Abensberg 48·49n 11·51e
Achern 48·37n 8·04e
Ahaus 52·04n 7·00e
Ahlen 51·46n 7·53e
Aichach 48·28n 11·08e
Alfeld 51·59n 9·50e
Alsfeld 50·45n 9·16e
Altomünster 48·28n 11·58e
Altötting 48·13n 12·40e
Alzey 49·45n 8·07e
Amberg 49·27n 11·52e
Andernach 50·26n 7·24e
Ansbach 49·17n 10·34e
Arolsen 51·23n 9·01e
Arnsberg 51·24n 8·03e
Aschaffenburg 49·59n 9·09e
Attendorn 51·07n 7·54e
Augsburg 48·23n 10·53e
Aurich 53·28n 7·29e
Backnang 48·56n 9·25e
Bad Bergzabern 49·07n 8·00e
Bad Bramstedt 53·55n 9·53e
Bad Dürkheim 49·28n 8·10e
Bad Ems 50·20n 7·43e
Baden-Baden 48·46n 8·14e
Badenweiler 47·48n 7·40e
Bad Godesberg 50·41n 7·10e
Bad Hersfeld 50·52n 9·42e
Bad Homburg [vor der Höhe] 50·13n 8·37e
Bad Honnef [am Rhein] 50·39n 7·13e
Bad Kissingen 50·12n 10·04e
Bad Kreuznach 49·52n 7·51e
Bad Mergentheim 49·30n 9·46e
Bad Nauheim 50·22n 8·44e
Bad Neuenahr 50·32n 7·05e
Bad Neustadt [an der Saale] 50·19n 10·13e
Bad Oeynhausen 52·12n 8·48e
Bad Oldesloe 53·48n 10·22e
Bad Orb 50·14n 9·20e
Bad Pyrmont 51·59n 9·15e
Bad Reichenhall 47·43n 12·52e
Bad Sachsa 51·36n 10·32e
Bad Salzuflen 52·05n 8·44e
Bad Schwalbach 50·08n 8·04e
Bad Schwartau 53·55n 10·40e
Bad Segeberg 53·56n 10·17e
Bad Tölz 47·46n 11·34e
Bad Waldsee 47·55n 9·45e
Bad Wildungen 51·07n 9·07e
Bad Wörishofen 48·00n 10·36e
Bad Zwischenahn 53·11n 8·00e
Baiersbronn 48·30n 8·22e
Balingen 48·16n 8·51e
Bamberg 49·53n 10·53e
Barsinghausen 52·18n 9·27e
Bayreuth 49·57n 11·35e
Bayrischzell 47·40n 12·00e
Beckum 51·45n 8·02e
Bendorf 50·25n 7·34e
Bensheim 49·41n 8·37e
Bergen 52·48n 9·58e
Bergisch Gladbach 50·59n 7·07e
Bernkastel-Kues 49·55n 7·04e
Bersenbrück 52·33n 7·56e
Betzdorf 50·47n 7·53e
Bevensen 53·05n 10·34e
Beverungen 51·39n 9·22e
Biberach an der Riss 48·06n 9·47e
Biedenkopf 50·55n 8·32e
Bielefeld 52·01n 8·31e
Bietigheim 48·54n 8·14e
Bingen 48·07n 9·16e
Birkenfeld 49·39n 7·10e
Bitburg 49·58n 6·31e
Blankenheim 50·26n 6·39e
Blaubeuren 48·24n 9·47e
Blaufelden 49·18n 9·58e
Blumberg 47·50n 8·31e
Böblingen 48·41n 9·01e
Bocholt 51·50n 6·36e
Bochum 51·28n 7·13e
Bogen 48·55n 12·43e
Bonn 50·44n 7·05e
Bordesholm 54·11n 10·01e
Borken 51·51n 6·51e
Brake 53·19n 8·28e
Bramsche 52·24n 7·58e
Braunschweig 52·16n 10·31e
Breisach 48·01n 7·40e
Bremen 53·04n 8·49e
Bremerhaven 53·33n 8·34e
Bremervörde 53·29n 9·08e
Bretten 49·02n 8·42e
Brilon 51·24n 8·34e
Bruchsal 49·07n 8·35e
Brückenau 50·18n 9·47e
Brühl 50·48n 6·54e
Brunsbüttelkoog 53·54n 9·08e
Brunswick, see Braunschweig
Buchen 49·32n 9·17e
Buchholz 53·20n 9·52e
Buchloe 48·02n 10·44e
Bückeburg 52·16n 9·02e
Büdingen 50·17n 9·07e
Bünde 52·12n 8·35e
Büren 51·33n 8·33e
Burg 54·26n 11·12e
Burghausen 48·09n 12·49e
Burglengenfeld 49·13n 12·03e
Burgsteinfurt 52·08n 7·20e
Büsum 54·08n 8·51e
Butzbach 50·26n 8·40e
Buxtehude 53·28n 9·41e
Calw 48·43n 8·44e
Celle 52·37n 10·05e
Cham 49·13n 12·41e
Clausthal-Zellerfeld 51·48n 10·20e
Cloppenburg 52·50n 8·02e
Coburg 50·15n 10·58e
Cochem 50·11n 7·09e
Cologne 50·56n 6·59e
Coesfeld 51·56n 7·10e
Crailsheim 49·08n 10·04e
Creussen 49·51n 11·37e
Cuxhaven 53·52n 8·42e
Dachau 48·15n 11·27e
Dannenberg 53·06n 11·05e
Darmstadt 49·53n 8·40e
Daun 50·11n 6·50e
Deggendorf 48·51n 12·59e
Delmenhorst 53·03n 8·38e
Detmold 51·56n 8·52e
Dieburg 49·54n 8·50e
Diepholz 52·35n 8·21e
Dillenburg 50·44n 8·17e
Dillingen [an der Donau] 48·34n 10·29e
Dingolfing 48·38n 12·31e
Dinkelsbühl 49·04n 10·19e
Donaueschingen 47·57n 8·29e
Donauwörth 48·43n 10·46e
Dorfen 48·17n 12·08e
Dortmund 51·31n 7·28e
Duderstadt 51·31n 10·16e
Dudweiler 49·17n 7·02e
Duisburg 51·25n 6·46e
Düren 50·48n 6·28e
Düsseldorf 51·12n 6·47e
Eberbach 49·28n 8·59e
Ebermannstadt 49·23n 11·13e
Ebern 50·05n 10·47e
Ebersberg 48·05n 11·58e
Ebingen 48·13n 9·01e
Eckernförde 54·28n 9·50e
Eggenfelden 48·25n 12·46e
Ehingen 48·17n 9·43e
Eichstätt 48·54n 11·12e
Einbeck 51·49n 9·52e
Eitorf 50·46n 7·26e
Ellwangen 48·57n 10·07e
Elmshorn 53·45n 9·39e
Elsfleth 53·14n 8·28e
Eltmann 49·58n 10·40e
Emden 53·22n 7·12e
Emmendingen 48·07n 7·50e
Emmerich 51·50n 6·15e
Emsdetten 52·10n 7·31e
Erbach 49·40n 8·59e
Erding 48·18n 11·54e
Erkelenz 51·05n 6·19e
Erlangen 49·36n 11·01e
Eschwege 51·11n 10·04e
Eschweiler 50·49n 6·16e
Espelkamp 52·25n 8·36e
Essen 51·28n 7·01e
Esslingen 48·45n 9·16e
Ettlingen 48·56n 8·24e
Euskirchen 50·39n 6·47e
Eutin 54·08n 10·37e
Falkenstein 49·06n 12·30e
Fallersleben 52·25n 10·43e
Fellbach 48·48n 9·15e
Feucht 49·22n 11·13e
Feuchtwangen 49·10n 10·20e
Flensburg 54·47n 9·26e
Forbach 48·41n 8·21e
Forchheim 49·43n 11·04e
Frankenberg-Eder 51·03n 8·48e
Frankfurt am Main 50·07n 8·40e
Freiburg [im Breisgau] 47·59n 7·51e
Freising 48·23n 11·44e
Freudenstadt 48·28n 8·25e
Friedberg 48·21n 10·58e
Friedberg 50·20n 8·45e
Friedrichshafen 47·39n 9·28e
Friedrichsort 54·24n 10·11e
Friedrichstadt 54·22n 9·05e
Fritzlar 51·08n 9·16e
Fulda 50·33n 9·41e
Fürstenfeldbruck 48·10n 11·15e
Fürth 49·28n 10·49e
Furth im Wald 49·18n 12·51e
Füssen 47·34n 10·42e
Gaggenau 48·48n 8·19e
Ganderkesee 53·02n 8·32e
Garmisch-Partenkirchen 47·29n 11·05e
Garrel 52·57n 8·01e
Gauting 48·04n 11·23e
Geesthacht 53·26n 10·22e
Geisenfeld 48·41n 11·37e
Geislingen [an der Steige] 48·36n 9·50e
Geisweid 50·55n 8·01e
Gelsenkirchen 51·31n 7·07e
Gemünden 50·03n 9·41e
Gerolzhofen 49·54n 10·21e
Geseke 51·38n 8·31e
Giessen 50·35n 8·40e
Gifhorn 52·29n 10·33e
Glücksburg 54·50n 9·33e
Glückstadt 53·47n 9·25e
Göggingen 48·20n 10·52e
Göppingen 48·42n 9·40e
Goslar 51·54n 10·25e
Göttingen 51·32n 9·55e
Grafenau 48·52n 13·25e
Grafing [bei München] 48·02n 11·59e
Greven 52·05n 7·36e
Grevenbroich 51·05n 6·35e
Griesbach 48·28n 13·11e
Griesheim 49·50n 8·34e
Gronau 52·05n 9·46e
Gross-Gerau 49·55n 8·29e
Gummersbach 51·02n 7·34e
Günzburg 48·27n 10·16e
Gunzenhausen 49·07n 10·45e
Gütersloh 51·54n 8·23e
Haag in Oberbayern 48·10n 12·11e
Haar 48·06n 11·44e
Hagen 51·22n 7·28e
Haltern 51·46n 7·10e
Hamburg 53·33n 9·59e
Hameln 52·06n 9·21e
Hamm 51·41n 7·49e
Hammelburg 50·07n 9·53e
Hanau 50·08n 8·55e
Hannover 52·24n 9·44e
Hassfurt 50·02n 10·31e
Hausach 48·17n 8·10e
Hechingen 48·21n 8·58e
Heide 54·12n 9·06e
Heidelberg 49·25n 8·43e
Heidenheim [an der Brenz] 48·40n 10·08e
Heilbronn 49·08n 9·13e
Heiligenhafen 54·22n 10·58e
Hemau 49·03n 11·47e
Hennef 50·46n 7·16e
Heppenheim [an der Bergstrasse] 49·39n 8·38e
Herborn 50·40n 8·17e
Herford 52·06n 8·40e
Herleshausen 51·00n 10·09e
Hersbruck 49·30n 11·26e
Herzberg [am Harz] 51·39n 10·20e
Hessisch Lichtenau 51·12n 9·43e
Hildesheim 52·09n 9·57e
Höchstadt an der Aisch 49·42n 10·44e
Hockenheim 49·19n 8·33e
Hof 50·18n 11·55e
Hofgeismar 51·30n 9·22e
Hofheim 50·07n 8·26e
Hofheim 50·08n 10·31e
Hohentwiel, see Singen
Holzkirchen 47·52n 11·42e
Holzminden 51·50n 9·27e
Homberg 51·02n 9·24e
Homburg 49·19n 7·20e
Horb 48·26n 8·41e
Hösbach 50·00n 9·12e
Höxter 51·46n 9·23e
Hünfeld 50·40n 9·46e
Husum 54·28n 9·03e
Hüttental 50·54n 8·02e
Ibbenbüren 52·16n 7·43e
Idar-Oberstein 49·42n 7·19e
Illertissen 48·13n 10·06e
Immenstadt 47·33n 10·13e
Ingelheim 49·59n 8·05e
Ingolstadt 48·46n 11·27e
Iserlohn 51·22n 7·41e
Isny 47·41n 10·02e
Itzehoe 53·55n 9·31e
Jever 53·34n 7·54e
Jülich 50·55n 6·21e
Kaiserslautern 49·26n 7·46e
Kandel 49·05n 8·11e
Kappeln 54·40n 9·56e
Karlsruhe 49·03n 8·24e
Karlstadt 49·57n 9·45e
Kassel 51·19n 9·29e
Kastl 49·22n 11·42e
Kaufbeuren 47·53n 10·37e
Kelheim 48·55n 11·52e
Kemnath 49·52n 11·54e
Kempten [Allgäu] 47·43n 10·19e
Kiefersfelden 47·37n 12·11e
Kirchberg 49·12n 9·58e
Kirchheimbolanden 49·40n 8·00e
Kirchheim [unter Teck] 48·39n 9·27e
Kirn 49·47n 7·28e
Kitzingen 49·44n 10·09e
Kleve 51·48n 6·09e
Kevelaer 51·35n 6·15e
Koblenz 50·21n 7·35e
Königswinter 50·40n 7·11e
Konz 49·42n 6·34e
Korbach 51·16n 8·52e
Kötzting 49·11n 12·52e
Krefeld 51·20n 6·34e
Kreuztal 50·58n 7·59e
Kronach 50·14n 11·20e
Krumbach 48·14n 10·22e
Kulmbach 50·06n 11·27e
Künzelsau 49·16n 9·41e
Kusel 49·32n 7·24e
Laboe 54·24n 10·15e
Lahnstein 50·18n 7·37e
Lahr 48·20n 7·52e
Lampertheim 49·35n 8·28e
Landau 49·12n 8·07e
Landau an der Isar 48·40n 12·43e
Landsberg [am Lech] 48·05n 10·55e
Landshut 48·33n 12·09e
Landstuhl 49·25n 7·34e
Langenhagen 52·27n 9·44e
Lauenburg 53·22n 10·33e
Lauf an der Pegnitz 49·30n 11·17e
Laupheim 48·14n 9·52e
Lauterbach 50·38n 9·24e
Leer 53·14n 7·26e
Lehrte 52·22n 9·59e
Lemgo 52·02n 8·54e
Lennestadt 51·09n 8·04e
Leonberg 48·48n 9·01e
Leutkirch 47·49n 10·01e
Leverkusen 51·03n 6·59e
Lich 50·33n 8·50e
Lichtenfels 50·09n 11·04e
Limburg an der Lahn 50·23n 8·04e
Lindau 47·33n 9·41e
Lingen 52·31n 7·19e
Lippstadt 51·40n 8·19e
Lohne 52·42n 8·12e
Lohr [am Main] 50·00n 9·34e
Löningen 52·44n 7·46e
Lörrach 47·37n 7·40e
Lübeck 53·52n 10·40e
Lüchow 52·58n 11·10e
Lüdenscheid 51·13n 7·38e
Ludwigsburg 48·53n 9·11e
Ludwigshafen 49·29n 8·26e
Lüneburg 53·15n 10·23e
Lünen 51·36n 7·32e
Mainburg 48·38n 11·47e
Mainhardt 49·04n 9·33e
Mainz 50·01n 8·16e
Malente 54·10n 10·33e
Mallersdorf 48·47n 12·16e
Mannheim 49·29n 8·29e
Marburg an der Lahn 50·49n 8·46e
Marktheidenfeld 49·50n 9·36e
Marktoberdorf 47·47n 10·37e
Marktredwitz 50·00n 12·06e
Marne 53·57n 9·00e
Mayen 50·19n 7·13e
Meersburg 47·41n 9·16e
Meldorf 54·05n 9·05e
Melsungen 51·08n 9·32e
Memmingen 47·59n 10·11e
Meppen 52·41n 7·17e
Merkendorf 49·12n 10·42e
Merzig 49·27n 6·36e


NORTH SEA
Deutsche Bucht
Helgoländer Bucht
Kieler Bucht
Mecklenburger Bucht
FEDERAL REPUBLIC OF GERMANY
NIEDERSACHSEN
SCHLESWIG-HOLSTEIN
NORDRHEIN-WESTFALEN
HESSEN
RHEINLAND-PFALZ
SAARLAND
BADEN-WÜRTTEMBERG
BAYERN
FRANCONIA
SCHWÄBISCHE ALB
SCHWÄBISCHE DONAU
ALLGÄUER ALPEN
HAMBURG
BERLIN
MUNICH
Flensburg
Schleswig
Kiel
Rostock
Lübeck
Bremerhaven
Cuxhaven
Itzehoe
Neumünster
Rendsburg
Emden
Wilhelmshaven
Oldenburg
Bremen
Lüneburg
Groningen
Delmenhorst
Hannover
Celle
Wolfsburg
Braunschweig
Salzgitter
Helmstedt
Magdeburg
Potsdam
Brandenburg
Osnabrück
Minden
Rheine
Bielefeld
Herford
Hameln
Hildesheim
Wolfenbüttel
Münster
Detmold
Paderborn
Lippstadt
Hamm
Arnhem
Nijmegen
Enschede
Bocholt
Recklinghausen
Gelsenkirchen
Bottrop
Oberhausen
Bochum
Essen
Duisburg
Dortmund
Eindhoven
Düsseldorf
Wuppertal
Remscheid
Solingen
Leverkusen
Cologne
Mönchengladbach
Neuss
Krefeld
Aachen
Liège
Bonn
Siegen
Lüdenscheid
Iserlohn
Hagen
Marburg an der Lahn
Giessen
Wetzlar
Kassel
Göttingen
Halle
Leipzig
Dessau
Erfurt
Weimar
Karl-Marx-Stadt
Eisenach
Fulda
Koblenz
Trier
Luxembourg
Frankfurt a.M.
Wiesbaden
Mainz
Kreuznach
Darmstadt
Aschaffenburg
Hanau
Offenbach
Rüsselsheim
Würzburg
Schweinfurt
Bamberg
Bayreuth
Coburg
Hof
Plzeň
Weiden in der Oberpfalz
Erlangen
Nürnberg
Fürth
Ansbach
Amberg
Regensburg
Mannheim
Ludwigshafen
Heidelberg
Speyer
Kaiserslautern
Saarbrücken
Saarlouis
Völklingen
Neunkirchen
Pirmasens
Metz
Nancy
Strasbourg
Karlsruhe
Pforzheim
Baden-Baden
Freiburg
Mulhouse
Basel
Zürich
SWITZERLAND
LIECHTENSTEIN
Heilbronn
Ludwigsburg
Stuttgart
Esslingen
Göppingen
Heidenheim
Schwäbisch Gmünd
Aalen
Ulm
Reutlingen
Tübingen
Sindelfingen
Böblingen
Nürtingen
Kirchheim
Geislingen
Augsburg
Ingolstadt
Straubing
Landshut
Freising
Dachau
Rosenheim
Traunstein
Salzburg
Kempten
Ravensburg
Friedrichshafen
Lindau
Konstanz
Singen
Villingen
Schwenningen
Tuttlingen
Rottweil
Sindelfingen
Innsbruck
AUSTRIA
BAVARIA
Size of symbol indicates relative size of town
Elevations in metres
0 25 50 100 150 km
0 25 50 100 mi
© Rand McNally & Co.
A-551200-257

MAP INDEX (continued)

Meschede.....51·20n 8·17e
Messkirch.....47·59n 9·07e
Metzingen.....48·32n 9·17e
Miesbach.....47·47n 11·50e
Miltenberg.....49·42n 9·15e
Mindelheim.....48·03n 10·29e
Minden.....52·17n 8·55e
Mittenwald.....47·27n 11·15e
Mölln.....53·37n 10·41e
Mönchengladbach.....51·12n 6·28e
Monheim.....48·50n 10·51e
Monschau.....50·33n 6·14e
Montabaur.....50·26n 7·50e
Moosburg.....48·29n 11·57e
Moringen.....51·42n 9·52e
Mosbach.....49·21n 9·08e
Mühlacker.....48·57n 8·50e
Mühldorf.....48·15n 12·32e
Müllheim.....47·48n 7·38e
Münchberg.....50·11n 11·47e
Münden.....51·25n 9·39e
Munich.....48·08n 11·34e
Münsingen.....48·25n 9·29e
Münster.....51·57n 7·37e
Munster.....52·59n 10·05e
Murnau.....47·40n 11·12e
Murrhardt.....48·59n 9·34e
Nabburg.....49·28n 12·11e
Naila.....50·19n 11·42e
Neckarsulm.....49·12n 9·13e
Neheim-Hüsten..51·27n 7·57e
Nesselwang.....47·37n 10·30e
Neuburg an der Donau.....48·44n 11·11e
Neuerburg.....50·00n 6·17e
Neu-Isenburg....50·03n 8·41e
Neumarkt in der Oberpfalz.....49·16n 11·28e
Neumarkt-Sankt-Veit.....48·22n 12·30e
Neumünster.....54·04n 9·59e
Neunburg vorm Wald.....49·21n 12·24e
Neunkirchen [Saar].....49·20n 7·10e
Neuss.....51·12n 6·41e
Neustadt.....47·54n 8·13e
Neustadt [an der Aisch].....49·34n 10·37e
Neustadt an der Waldnaab.....49·44n 12·11e
Neustadt an der Weinstrasse....49·21n 8·08e
Neustadt [bei Coburg]....50·19n 11·07e
Neustadt in Holstein.....54·06n 10·48e
Neu-Ulm.....48·23n 10·01e
Neuwied.....50·25n 7·27e
Niebüll.....54·48n 8·50e
Niedermarsberg.51·28n 8·50e
Nienburg.....52·38n 9·13e
Norden.....53·36n 7·12e
Nordenham.....53·29n 8·28e
Nordhorn.....52·27n 7·05e
Nördlingen.....48·51n 10·30e
Northeim.....51·42n 10·00e
Nürnberg.....49·27n 11·04e
Nürtingen.....48·38n 9·20e
Oberhausen.....51·28n 6·50e
Obernburg am-Main.....49·50n 9·08e
Oberursel.....50·11n 8·35e
Oberviechtach...49·28n 12·25e
Obing.....48·00n 12·24e
Ochsenfurt.....49·40n 10·03e
Ochtrup.....52·13n 7·11e
Oelde.....51·49n 8·08e
Offenbach.....50·08n 8·47e
Offenburg.....48·28n 7·57e
Öhringen.....49·12n 9·29e
Oldenburg.....53·08n 8·13e
Oldenburg.....54·17n 10·52e
Olpe.....51·02n 7·52e
Opladen.....51·04n 7·00e
Osnabrück.....52·16n 8·02e
Osterholz-Scharmbeck....53·14n 8·47e
Osterode.....51·43n 10·14e
Öttingen in Bayern.....48·57n 10·36e
Ottweiler.....49·24n 7·09e
Paderborn.....51·43n 8·45e
Papenburg.....53·05n 7·23e
Parsberg.....49·09n 11·43e
Passau.....48·35n 13·28e
Pegnitz.....49·45n 11·33e
Peine.....52·19n 10·13e
Peissenberg.....47·48n 11·04e
Penzberg.....47·45n 11·23e
Pfarrkirchen.....48·27n 12·56e
Pforzheim.....48·54n 8·42e
Pfronten.....47·34n 10·33e
Pfungstadt.....49·48n 8·36e
Pinneberg.....53·40n 9·47e
Pirmasens.....49·12n 7·36e
Plattling.....48·47n 12·53e
Plön.....54·09n 12·25e
Porz.....50·53n 7·03e
Preetz.....54·14n 10·16e
Prien [am Chiemsee].....47·51n 12·20e
Prüm.....50·12n 6·25e
Puttgarden.....54·30n 11·13e
Quakenbrück....52·40n 7·57e
Radolfzell.....47·44n 8·58e
Rastatt.....48·51n 8·12e
Rastede.....53·15n 8·11e
Ratzeburg.....53·42n 10·46e
Ravensburg.....47·47n 9·37e
Recklinghausen..51·36n 7·13e
Rees.....51·45n 6·23e
Regen.....48·59n 13·07e
Regensburg.....49·01n 12·06e
Rehau.....50·15n 12·02e
Remscheid.....51·11n 7·11e
Rendsburg.....54·18n 9·40e
Reutlingen.....48·29n 9·11e
Rheda-Wiedenbrück...51·50n 8·20e
Rheine.....52·17n 7·26e
Rheydt.....51·10n 6·25e
Rinteln.....52·11n 9·04e
Rockenhausen...49·38n 7·49e
Rodalben.....49·14n 7·38e
Roding.....49·12n 12·43e
Rosenheim.....47·51n 12·07e
Rotenburg.....53·06n 9·24e
Rotenburg [an der Fulda].....51·00n 9·45e
Roth [bei Nürnburg].....49·14n 11·04e
Rothenburg ob der Tauber.....49·23n 10·10e
Rott am Inn.....47·59n 12·07e
Rottenburg.....48·28n 8·56e
Rottenburg an der Laaber.....48·42n 12·02e
Rottweil.....48·10n 8·37e
Ruhpolding.....47·45n 12·38e
Rüsselsheim.....50·00n 8·25e
Saarbrücken.....49·14n 6·59e
Saarburg.....49·36n 6·33e
Saarlouis.....49·21n 6·45e
Salzgitter.....52·10n 10·25e
Sankt Goar.....50·09n 7·43e
Sankt Goarshausen...50·09n 7·43e
Sankt Ingbert....49·17n 7·06e
Sankt Peter.....54·18n 8·38e
Sankt Wendel....49·28n 7·10e
Saulgau.....48·01n 9·30e
Scheinfeld.....49·40n 10·27e
Schesslitz.....49·59n 11·01e
Schleiden.....50·31n 6·28e
Schleswig.....54·31n 9·33e
Schlitz.....50·40n 9·33e
Schloss Neuhaus.....51·44n 8·43e
Schlüchtern.....50·20n 9·31e
Schmidmühlen ..49·16n 11·56e
Schneverdingen.53·07n 9·47e
Schongau.....47·49n 10·54e
Schopfheim.....47·39n 7·49e
Schorndorf.....48·48n 9·31e
Schramberg.....48·13n 8·23e
Schroben-hausen.....48·33n 11·17e
Schwabach.....49·20n 11·01e
Schwäbisch Gmünd.....48·48n 9·47e
Schwäbisch Hall.....49·07n 9·44e
Schwab-münchen.....48·11n 10·45e
Schwandorf in Bayern.....49·20n 12·08e
Schweinfurt.....50·03n 10·14e
Schwenningen...48·04n 8·32e
Schwetzingen....49·23n 8·34e
Seesen.....51·53n 10·10e
Selb.....50·10n 12·08e
Selm.....51·42n 7·28e
Sennestadt.....51·57n 8·35e
Siegburg.....50·47n 7·12e
Siegen.....50·52n 8·02e
Sigmaringen.....48·05n 9·13e
Simbach.....48·34n 12·45e
Simmern.....49·59n 7·31e
Sindelfingen.....48·42n 9·00e
Singen (Hohentwiel)....47·46n 8·50e
Sinsheim.....49·15n 8·53e
Sobernheim.....49·47n 7·38e
Soest.....51·34n 8·07e
Solingen.....51·10n 7·05e
Soltau.....52·59n 9·49e
Sonthofen.....47·31n 10·17e
Speyer.....49·19n 8·26e
Springe.....52·12n 9·32e
Stadt Allendorf..50·50n 9·01e
Stadthagen.....52·19n 9·13e
Stadtoldendorf...51·53n 9·37e
Staffelstein.....50·06n 11·00e
Starnberg.....48·00n 11·20e
Stockach.....47·51n 9·00e
Stolberg.....50·46n 6·13e
Straubing.....48·53n 12·34e
Stuttgart.....48·46n 9·11e
Sulingen.....52·41n 8·47e
Sulzbach.....49·18n 7·07e
Sulzbach-Rosenberg.....49·30n 11·45e
Syke.....52·54n 8·49e
Tauberbischofsheim.....49·37n 9·40e
Tecklenburg.....52·13n 7·48e
Tegernsee.....47·43n 11·45e
Thalfang.....49·45n 6·59e
Timmendorfer-strand.....54·00n 10·46e
Tirschenreuth...49·53n 12·21e
Tittling.....48·44n 13·23e
Tönning.....54·19n 8·56e
Traunstein.....47·52n 12·38e
Treuchtlingen...48·57n 10·54e
Triberg.....48·08n 8·13e
Trier.....49·45n 6·38e
Troisdorf.....50·49n 7·08e
Tübingen.....48·31n 9·02e
Tuttlingen.....47·59n 8·49e
Tutzing.....47·54n 11·17e
Überlingen.....47·46n 9·10e
Uelzen.....52·58n 10·33e
Uetersen.....53·41n 9·39e
Uffenheim.....49·32n 10·14e
Ulm.....48·24n 10·00e
Usingen.....50·20n 8·32e
Uslar.....51·39n 9·38e
Utersum.....54·43n 8·24e
Vaihingen [an der Enz].....48·56n 8·58e
Varel.....53·22n 8·10e
Vechta.....52·43n 8·16e
Velden.....48·19n 12·16e
Verden.....52·55n 9·13e
Viechtach.....49·05n 12·53e
Villingen [im Schwarzwald]...48·03n 8·27e
Vilsbiburg.....48·27n 12·12e
Vilshofen.....48·39n 13·12e
Visselhövede....52·59n 9·35e
Vohenstrauss....49·37n 12·21e
Volkach.....49·52n 10·13e
Völklingen.....49·15n 6·50e
Waging am See.....47·56n 12·43e
Waiblingen.....48·50n 9·19e
Waldbröl.....50·53n 7·37e
Waldeck.....51·12n 9·04e
Waldkirchen....48·44n 13·37e
Waldmünchen...49·23n 12·43e
Waldshut.....47·37n 8·13e
Walsrode.....52·52n 9·35e
Wangen [im Allgäu].....47·41n 9·50e
Warburg.....51·29n 9·08e
Warendorf.....51·57n 7·59e
Wasseralfingen..48·52n 10·06e
Wasserburg am Inn.....48·04n 12·13e
Wegscheid.....48·36n 13·48e
Weiden in der Oberpfalz.....49·41n 12·10e
Weilburg.....50·29n 8·15e
Weilheim.....47·50n 11·08e
Weinheim.....49·33n 8·39e
Weissenburg in Bayern.....49·01n 10·58e
Werne [an der Lippe].....51·40n 7·40e
Wertingen.....48·34n 10·41e
Wertheim.....49·46n 9·31e
Wesel.....51·40n 6·38e
Westerland.....54·54n 8·18e
Westerstede.....53·15n 7·55e
Wetzlar.....50·33n 8·29e
Wiehl.....50·57n 7·31e
Wiesbaden.....50·05n 8·14e
Wiesloch.....49·17n 8·42e
Wietze.....52·39n 9·50e
Wilhelmshaven...53·31n 8·08e
Winsen.....53·22n 10·12e
Wipperfürth.....51·07n 7·23e
Wittingen.....52·43n 10·44e
Wittlage.....52·19n 8·22e
Wittlich.....49·59n 6·53e
Wittmund.....53·34n 7·47e
Witzenhausen...51·20n 9·51e
Wolfach.....48·17n 8·13e
Wolfenbüttel.....52·10n 10·32e
Wolfhagen.....51·19n 9·10e
Wolfratshausen..47·54n 11·25e
Wolfsburg.....52·25n 10·47e
Worms.....49·38n 8·22e
Wunstorf.....52·25n 9·26e
Wuppertal.....51·16n 7·11e
Würzburg.....49·48n 9·56e
Wyk.....54·42n 8·34e
Zell.....50·01n 7·10e
Zeven.....53·18n 9·16e
Ziegenhain.....50·55n 9·15e
Zweibrücken.....49·15n 7·21e
Zwiesel.....49·01n 13·14e

Physical features and points of interest

Aisch, *river*.....49·46n 11·01e
Allgäu, *physical region*.....47·35n 10·10e
Allgäuer Alpen, *mountains*.....47·24n 10·15e
Altmühl, *river*....48·55n 11·52e
Amrum, *island*...54·39n 8·21e
Bayerische Alpen (Bavarian Alps), *mountains*.....47·30n 11·00e
Black Forest, see Schwarzwald
Bodensee, *lake*...47·35n 9·25e
Bohemian Forest, see Böhmer Wald
Böhmer Wald, *mountains*.....48·55n 13·30e
Borkum, *island*...53·35n 6·41e
Chiemsee, *lake*...47·54n 12·29e
Constance, Lake, see Bodensee
Danube, *river*....48·31n 13·41e
Deutsche Bucht, *bay*.....54·30n 7·30e
Dill, *river*.....50·33n 8·29e
Dithmarschen, *physical region*..54·10n 9·15e
Dortmund-Ems-Kanal, *canal*....51·32n 7·27e
Dümmer, *lake*....52·30n 8·21e
Eder, *river*.....51·13n 9·27e
Eifel, *mountains*..50·15n 6·45e
Elbe, *river*.....53·50n 9·00e
Elbe-Lübeck Kanal, *canal*....53·49n 10·38e
Ems, *river*.....51·09n 9·26e
Emsland, *physical region*..52·50n 7·20e
Entenbuhl, *mountain*.....49·46n 12·24e
Erft, *river*.....51·11n 6·44e
Fehmarn, *island*..54·28n 11·08e
Fehmarnbelt, *strait*.....54·35n 11·15e
Feldberg, *mountain*.....47·52n 8·00e
Fichtelgebirge, *mountains*.....50·11n 11·58e
Föhr, *island*.....54·43n 8·30e
Forggensee, *lake*.47·36n 10·44e
Franconia, *historic region*...49·43n 10·10e
Fulda, *river*.....51·25n 9·39e
Grosser Arber, *mountain*.....49·07n 13·01e
Grosser Feldberg, *mountain*.....50·14n 8·26e
Halligen, *islands*.54·35n 8·35e
Harz, *mountains*.51·50n 10·20e
Harz, Naturpark, *park*.....51·44n 10·30e
Hase, *river*.....52·41n 7·18e
Heiligon, Vierzehn, *shrine*.....50·08n 11·02e
Helgoland, *island*.....54·12n 7·53e
Helgoländer Bucht, *bay*.....54·10n 8·04e
Hermannsdenkmal, *monument*.51·55n 8·50e
Hoohtaunus, Naturpark, *park*.50·20n 8·20e
Hohe Acht, *mountain*.....50·23n 7·00e
Hunsbrück, *mountains*.....49·50n 6·40e
Hunte, *river*.....52·30n 8·19e
Iller, *river*.....48·23n 9·58e
Inn, *river*.....48·35n 13·28e
Isar, *river*.....48·49n 12·58e
Jadebusen, *bay*..53·30n 8·10e
Jagst, *river*.....49·14n 9·11e
Juist, *island*.....53·40n 7·00e
Kahler Asten, *mountain*.....51·11n 8·29e
Katzenbuckel, *mountain*.....49·28n 9·02e
Kieler Bucht (Kiel Bay), *bay*..54·35n 10·35e
Kocher, *river*....49·14n 9·12e
Königssee, *lake*..47·36n 12·59e
Lahn, *river*.....50·18n 7·37e
Langeoog, *island*.53·46n 7·32e
Lech, *river*.....48·44n 10·56e
Leine, *river*.....52·46n 9·35e
Lemberg, *mountain*.....48·09n 8·45e
Lippe, *river*.....51·39n 6·38e
Loreley, *scenic area*.....50·08n 7·44e
Lübecker Bucht, *bay*.....54·00n 10·55e
Lüneburger Heide, *physical region*..53·10n 10·20e

MAP INDEX (continued)

Lüneburger Heide, Naturpark, *park*. 51·37n 7·31e
Main, *river*. 50·00n 8·18e
Maria Laach, *cloister*. 50·24n 7·14e
Mecklenburger Bucht, *bay*. 54·20n 11·40e
Mitteland Kanal, *canal*. 52·25n 9·10e
Mosel, *river*. 50·22n 7·36e
Nagold, *river*. 48·52n 8·42e
Nahe, *river*. 49·58n 7·57e
Neckar, *river*. 49·31n 8·26e
Nordeifel, Naturpark, *park*. 50·40n 6·20e
Norderney, *island*. 53·42n 7·10e
Nordfriesland, *physical region*. 54·40n 8·55e
Nordpfälzer Bergland, *mountains*. 49·40n 7·40e
Nord-Ostsee Kanal, *canal*. 53·53n 9·08e
Nordstrand, *island*. 54·30n 8·53e
North Frisian Islands. 54·25n 8·30e
North Sea. 54·15n 6·00e
Nürburgring, *race course*. 50·12n 7·20e
Odenwald, *mountains*. 49·40n 9·00e
Oker, *river*. 51·26n 9·39e
Ostfriesische Inseln, *islands*. 53·44n 7·25e
Pellworm, *island*. 54·31n 8·38e
Pfalz, *forest*. 49·41n 12·25e
Regen, *river*. 49·02n 12·06e
Rhine, *river*. 51·52n 6·05e
Rothaargebirge, *mountains*. 51·05n 8·15e
Ruhr, *historic region*. 51·25n 7·00e
Saar, *river*. 49·42n 6·34e
Sauerland, *physical region*. 51·10n 8·00e
Schneeberg, *mountain*. 50·03n 11·51e
Schwaben, *historic region*. 48·20n 10·30e
Schwäbische Alb, *mountains*. 48·25n 9·30e
Schwarzwald, *mountains*. 48·00n 8·15e
Sieg, *river*. 50·46n 7·07e
Spiekeroog, *island*. 53·46n 7·42e
Starnberger See, *lake*. 47·55n 11·18e
Steinhuder Meer, *lake*. 52·28n 9·20e
Sylt, *island*. 54·54n 8·20e
Tauber, *river*. 49·46n 9·31e
Taunus, *mountains*. 50·10n 8·15e
Teutoburger Wald, *forest*. 52·10n 8·15e
Treene, *river*. 54·22n 9·05e
Vogelsberg, *mountains*. 50·30n 9·15e
Wagrien, *physical region*. 54·15n 10·45e
Walhalla, *monument*. 49·03n 12·14e
Wangerooge, *island*. 53·46n 7·55e
Wasserkuppe, *mountain*. 50·30n 9·56e
Watzmann, *mountain*. 47·33n 12·55e
Werra, *river*. 51·26n 9·39e
Weser, *river*. 53·32n 8·34e
Westerwald, *mountains*. 50·40n 7·55e
Westfalen, *historic region*. 51·50n 7·30e

ern extension of the Jura mountain system, which has its origins on the eastern bank of the Rhône Valley in southeastern France. Entering Germany at the juncture of the French and Swiss borders in the extreme southeast, the Jura branches into two ranges within Germany. South of the Jura Mountains and the basin of the Danube River, an undulating upland plateau forms the southernmost region of West Germany as it skirts the country's southern frontiers and terminates at the shores of Lake Constance, the mountainous Allgäu, and the precipitous rise of the Bavarian Alps south of Munich.

Topography and resources. *The North German Plain, or Lowland.* Embracing all of the *Länder* ("states") of Schleswig-Holstein, Hamburg, and Bremen, the greater portion of Niedersachsen and the northwestern areas of Nordrhein-Westfalen, the North German Plain has been graven by glacial action. It presents, in spite of some intimations of monotony, an area of considerable variety in topography—to say nothing of scenic beauty. Although great stretches of the coastal plain between the Weser River and the Dutch border are largely flat, the moraine-formed areas stretching southward down the Jutland Peninsula of Denmark include gentle, rolling terrain of low hills and broad valleys, woodland areas, lakes, marshland, and heaths. Of the last, the best known is the Lüneburger Heide (Lüneburg Heath), a plateau extending on a morainic belt between Hamburg and Hannover.

A northern morainic belt, only a small portion of which is in West German territory, extends along the Baltic from eastern Schleswig-Holstein. The coastline of Schleswig-Holstein and Niedersachsen along the North Sea is characterized by sandy dunes, tidal inlets, and marshland. Farther inland, the grassy dairylands are interspersed with the sudden rises of geest, oblong hills representing groups of old alluvial deposits.

Off the western peninsula, the Schleswig portion of the Jutland coast, lie the North Frisian Islands, popular as summer resorts; the best known of them is the isle of Sylt. Extending east to west off the coast of Niedersachsen lies the chain of the East Frisian Islands; their West Frisian extension belongs to The Netherlands. The tiny rocky island of Helgoland is located in the North Sea some 40 miles (65 kilometres) northwest of the mouth of the Elbe at Cuxhaven. The alluvial plain stretching from the eastern part of Holstein and the north portion of Nordrhein-Westfalen to the Dutch border is of great economic importance because of its rich soil and mineral wealth. Lignite and bituminous coal is found in areas west of the Elbe and in the Lower Rhine areas; iron ore, potash, and salt deposits are found in the plains north of the Harz Mountains, and, in addition, there are some oil and natural-gas deposits between the Ems, Weser, and Elbe rivers and up through Schleswig-Holstein.

Apart from its agricultural importance, particularly in grain, dairy products, and livestock, the plain's wealth of rivers and canals—largely a result of its glacial origin—is an important adjunct to Germany's intricate inland-waterway transportation system. The Elbe and Weser rivers flow through Germany's two great inland seaports, Hamburg and Bremen, respectively. The Ems, running parallel to the Dutch border, also empties into the North Sea. Farther to the south and moving to the northwest, the Rhine, Europe's greatest inland waterway, flows to the North Sea through The Netherlands after passing Cologne and Düsseldorf, just to the west of Europe's largest concentration of industrial and mineral wealth, the Ruhr. This district, taking its name from a river valley that constitutes only a portion of the area, represents a great industry–population complex roughly triangular in shape. At its southern tip lies Cologne, with Dortmund in the northeast and Duisburg in the northwest. Rich in coal and formerly in iron, it is Germany's greatest centre of heavy industry—steel, machinery, oil refining, chemicals, glass, ceramics, textiles, and many other manufactures.

Industrial concentration in the Ruhr

The Central Uplands. In general, the Central Uplands constitute an extremely intricate and geologically complex continuum of mountain ranges, separated in their larger divisions by river and stream valleys. They are rounded and of no great height, most less than 3,500 feet (1,100 metres) in elevation. At their upper levels they tend to be heavily forested; they are important sources of timber, and a number of these mountain districts are known as forestlands rather than mountain ranges. A number of areas, especially those formed by volcanic activity in relatively recent geological times, are important sources of minerals and centres of local industry. Valleys are utilized for crop raising and, where climatic conditions permit, for growing wine grapes. In other areas, a plentiful supply of hydroelectricity and easy access to water, rail, and road transport support mining, manufacturing, and processing industries.

This Central Upland district of West Germany—all of Hessen, Rheinland-Pfalz, and the Saarland and the southern portions of Nordrhein-Westfalen and Niedersachsen—is extremely variegated in character; sharing only its mountainous and hilly terrain as a common feature. It contains some of West Germany's most centrally located and wealthy industrial areas (in the Rhine–Main area of southern Hessen) and also some of the most remote and sparsely populated backwoods regions.

The high tier of the Central Uplands extends east of the Rhine and south of the Ruhr and the Lippe rivers and north of the Lahn River eastward to the upper reaches of the Weser. In this range, or bordering on it, are the Teutoburgerwald (Teutoburg Forest) in the north, the Siebengebirge and Sauerland in the west, and the rugged Westerwald in its south near the Lahn. Across the Weser and its tributary, the Werra, the range continues into the Harz region, south of Hannover and Braunschweig. The Harz Mountains, noted for their beauty and popular as a resort area, extend into East Germany. The Brocken, at 3,747 feet (1,142 metres) the highest peak in the Harz range, is known for its many associations with German heathen legend; its peak straddles the frontier of West and East Germany.

East of the Rhine and between the Lahn and Main rivers are the highland districts of the Taunus, also a favoured resort, and the Spessart, an extensive forest preserve in which the last of Germany's ancient oak trees

are found. To its south, between the Main and Neckar rivers, lies the fabled Odenwald, a remote forest near the populous Frankfurt am Main metropolitan area; in this wood the legendary Siegfried is said to have been slain.

To the northwest, west of the Rhine and north of the Mosel (Moselle), the vast tract of the Eifel Mountains continues into Belgium as the Ardennes. This range, together with the parallel mountain ridge south of the Mosel known as the Hunsrück, bears evidence of geologically recent volcanic action and is an important source of basalt. Semiprecious stones are mined in the vicinity of Idar-Oberstein. South of the Hunsrück, between the Franco-Luxembourg border, lie the highlands of the Pfälzerwald (Palatinate Forest). The Saar Basin, bordering French Lorraine in the extreme southwest of the Pfälzerwald, is, like the Ruhr, rich in coal and iron, though of an even higher quality than that of the northern mining regions.

The Rhine gorge cuts and winds through the plateau of lowland mountains that represents a continuum between the Eifel and the Hunsrück in the west and the Siebengebirge and Taunus in the east. This famed and most spectacular stretch of the Rhine, from about Andernach in the north to Bingen at its southern reach, is lined with hilltop castles, vineyards along its precipitous slopes, and a sequence of picturesque towns along its banks. The winding stretches near Sankt Goarshausen pass between all but vertical rocky cliffs, the most famous of which is the fabled Lorelei rock.

The Southern Highlands, or plateau region. The Main River, which, upstream from its mouth at the Rhine near Mainz, meanders in a westerly direction toward the juncture of the West German, East German, and Czechoslovakian borders, may be regarded as the dividing line between the Central Uplands and the mountainous and plateau region of southern Germany. The area lacks no one specific characteristic to distinguish it from the Central Uplands. In its topography, if not in its external appearance, it is rather similar to the central regions. Its mountains, however, are derived from different systems from those in central Germany, and its mean elevation above sea level, even in plateau regions, is considerably higher.

In an area of northern Baden-Württemberg and northeastern Bavaria framed by the basins of the Main and Neckar rivers and extending to the east is a region of low hills, plateaus, and valleys. It is both rich as an agricultural area and noted for its many ancient towns that, largely because of their remoteness from the main routes of transport, have remained in their outward appearance relatively untouched by modern development. This district includes the renowned "Romantic Road" from Donauwörth to Würzburg, passing through Rothenburg-ob-der-Tauber, a perfectly preserved walled city that has remained virtually unchanged since the 17th century.

The Jura range and the Black Forest

The northernmost reaches of the Jura range, entering Germany from Switzerland in the extreme southwest, branch into two forks. The western fork, moving in a northerly direction parallel with the Rhine River to a point just south of Karlsruhe, is commonly known as the Schwarzwald (Black Forest), popularly associated with the dark appearance of the area's giant fir trees. The Schwarzwald is rich in thermal water sources and abounds in resort areas, spas, and sanitoriums. West of the Schwarzwald is the wide, sloping eastern valley of the Rhine opposite Alsace, in which the lower and smoother Vosges Mountains on the French side run parallel to the Schwarzwald as part of the northern Jura.

The eastern fork of the Jura range branches off from the northern extension some 15 miles east of Freiburg im Breisgau, where the Danube River rises to meander 1,770 miles eastward past seven countries to the Black Sea. This branch extends due east some 320 miles to near the source of the Main, skirting to the south of Stuttgart and Nürnberg and following the northern basin of the Upper Danube as far as Donauwörth. It is known as the Schwäbische Alb (Swabian Jura), and portions of it evidence curious and often rugged geological formations.

From Donauwörth, the Jura swings in an arc almost due north to its terminal point near Bayreuth and just south of the mountain Schneeberg (3,455 feet). To the east, toward the Czechoslovakian border, lie the rolling wooded plains of the Böhmer Wald (Bohemian Forest); farther south, near Regensburg, the Bayerischer Wald (Bavarian Forest) lines the northern slopes of the Danube Valley.

Justly famed as they are for their dramatic beauty—and especially for the foothills in their immediate foreground—the Bavarian Alps represent, in effect, only the northern edge of two relatively minor extensions of the Austrian Alps, the Arlberg and Tirolean ranges. Germany's share in Alpine territory is, in fact, miniscule by comparison to that of Austria, Switzerland, Italy, and France.

Waters and soils. Because of Germany's steady upward incline from the coastal regions of the north to the Bavarian Plateau and Alpine regions of the southeast, the greater majority of its rivers and their tributaries flow north and empty into the North Sea via the Rhine, the Ems, the Weser, and the Elbe. While many of the tributaries to these major rivers meander east–west or west–east, only the Lahn flows in a southerly direction, emptying into the Rhine. The remainder of German waters—all of which are located in Bavaria—escape into the Danube on its eastward course through the Balkans; the main tributaries to the south of the Danube are the Iller, the Lech, the Isar, and the Inn; flowing southward from the north are the Altmühl, the Naab, and the Regen.

The lakes of West Germany

The Federal Republic has relatively few lakes. Most lakes across the German regions are the shallow lakes of the postglacial lowland of the north; the majority are in East Germany. In addition to the Dümmer See (Dümmer Lake) and the Steinhuder Meer in Niedersachsen, a few small lakes of glacial origin dot Schleswig-Holstein. The remainder of West Germany's lakes are concentrated at the extreme southeastern corner of Upper Bavaria, many of these in outstandingly beautiful surroundings. Germany shares Lake Constance (Bodensee), its largest lake (having the proportions of an inland sea), with Switzerland and Austria. Stretching 46 miles at its longest reach, Lake Constance is, in effect, an elongated basin filled by the Rhine River from its point of origin in eastern Switzerland; at its southwesternmost fork the lake reassumes the form of a river and flows eastward toward Basel.

Although practically all of Germany's arable land is under cultivation, comparatively few areas are highly fertile. The terrains thus unsuited for large-scale grain production—chiefly wheat, barley, rye, and oats, in addition to potatoes, sugar beet, and some tobacco—are given over to grazing land. In the lowlands of the north, the land overwhelmingly comprises sand and loam, though along the North Sea littoral up the western coast of Jutland the soil is predominantly sandy. The eastern half of the Jutland Peninsula and the tier of the Baltic coast extending southward to Hamburg is mostly the clay of the northern morainic belt, or Baltic Ridge.

An especially fertile area known as the *Börde* is covered with a siltlike loess; it extends in a wide belt from west of the Rhine near the Ruhr district eastward, following the southern edge of the Mittelland Kanal and extending into East Germany at the Elbe near Magdeburg. Other regions containing this richest of soil are found in scattered areas, particularly in the lowland regions west of the Rhine to the Dutch border and in three districts around the greater Frankfurt am Main area. Other loess-covered districts of considerable fertility are located in the north of Württemberg, in small portions of northern Bavaria, and in wide ranges to the southeast of Ulm and east of Munich. Some marshland is found both in the North German Plains and in the Alpine plateau.

The remainder of German soil types, because of the preponderance of mountainous and forested areas, range from sandy loam, from loam to clay, and from clay to rocky outcrops. Timber production thrives where the land is all but unarable, and viticulture in the southern regions flourishes in an otherwise inhospitable type of soil.

Climate. West Germany is favoured with a generally temperate climate, especially in view of its northerly latitudes and the distance of the larger portions of its territory from the warming influence of the Gulf Stream of the Atlantic. Excesses of extremely high temperatures in the summer or of deep, prolonged frost in the winter are rare. These conditions, together with ordinarily a more than abundant and well-distributed amount of rainfall, afford ideal conditions for crop raising. As throughout western Europe in general, however, Germany's climate is subject to quick variations when the warm westerly maritime climate emanating from the Gulf Stream collides with the more extreme climatic ranges moving in from northeastern Europe.

Progression of the seasons

The seasons, year by year, are subject to great variations: winters may be unusually cold or prolonged, particularly in the higher elevations in the south, or mild, with the temperatures hovering only two or three degrees above and below the freezing point. Spring may arrive early and extend through a hot, rainless, summer to a warm, dry autumn with the threat of drought. In other years, spring—invariably interrupted by a frosty lapse in May, popularly known as *die drei Eisheiligen* ("the three ice saints")—may arrive so late as to be imperceptible and be followed by a cool, rainy summer. One less agreeable feature of the German climate is an almost permanent overcast in the cool seasons, only infrequently accompanied by precipitation; it sets in toward the latter part of autumn and lifts as late as March or April. Thus for months on end virtually no sunshine may appear.

The intricate topography of the Central Uplands and Southern Highlands and their valleys causes infinite local variations in temperature, frequency of sunshine, humidity, and precipitation. For this reason Germany produces more vintages and varieties of wine than any other country, at least 10,000 registered for marketing purposes.

The northwestern and lowland portions of the republic are affected chiefly by the uniformly moist air, moderate in temperature, that is carried inland from the North Sea by the prevailing westerly winds. While this influence, on the whole, affords moderately warm summers and mild winters, it is accompanied by the disadvantages of high humidities, extended stretches of rainfall, and, in the cooler seasons, fog. The hilly districts of the central and southwestern regions and, to an even greater degree, the upland and plateau districts of the southeast are subject to the more pronounced ranges of hot and cold from the countervailing continental climate. Southeastern Germany may intermittently be the coldest area of the country in the winter; but the valleys of the Rhine, Main, Neckar, and Mosel rivers may also be the hottest in the summer. Winters in the North German Plains tend to be consistently colder, if only by a few degrees, than in the south, largely because of winds from Scandinavia.

One anomaly of the climate of Upper Bavaria in the warm seasons is the occasional appearance of warm, dry air passing over the northern Alps to the Bavarian Plateau. These mild winds, known as *Föhn*, not only create an optical phenomenon that makes the Alps visible from points where normally they would be out of sight but also are reputed to have a depressing psychological effect upon the inhabitants.

Annual mean precipitation varies according to region. In the North German Plain it ranges between 20 to 28 inches (510 to 710 millimetres); in the Central Uplands, from 28 to 59 inches; and in the Alpine districts, up to and exceeding 78 inches. The average temperature for January, the coldest month, varies from 27° to 34° F (−3° to 1° C) in the lowland areas of the north; in the upland regions, depending on altitude, temperatures in January will fall to an average of 21° F (−6° C). July temperatures average from 60° to 66° F (16° to 19° C) and run slightly higher in the sheltered river valleys.

Vegetation and animal life. In ancient times Germany was covered by preponderantly deciduous forests of oak, linden, beech, and birch. Only beech and birch remain evident in some abundance in the patchy forests of the North German Plains and Schleswig-Holstein; oak is found less frequently. After the felling of these primeval forests, Germany took the lead in systematic reforestation, mostly, however, by planting the faster-growing conifers. Thus, fir, spruce, and pine outnumber deciduous trees of all varieties by more than two to one. Because of a carefully regulated system of reforestation, the Federal Republic contains the greatest density of woodland of any larger region of western Europe exclusive of Scandinavia, and forestry and timber production are major facets of German agriculture. The heavily forested districts, nevertheless, are almost solely confined to the central and southern uplands.

Reforested areas

Fruit trees—apple, pear, peach, cherry, plum, apricot, and quince—and the walnut, chestnut, and hazelnut trees are widely distributed, but, like the vineyards, they thrive best in the western and southwestern regions. Strawberries, blueberries, blackberries, cranberries, and currants abound in the forests and meadows of all districts.

The vast tracts of forest and mountainous terrain, with only scattered habitation, contribute to a surprising variety of wildlife in so densely populated and highly developed a country. Game animals abound in all regions—several varieties of deer, quail, and pheasant and, in the Alpine regions, the chamois and ibex—and their numbers are protected by stringent game laws. The wild boar population, which soared after World War II because of restrictions on hunting, has now been reduced so that it no longer represents a danger to man and crops. The hare, a favoured game animal, is ubiquitous. Although the bear and the wolf are now extinct in the republic, the wildcat has had a resurgence in the postwar years, especially in the Eifel and Hunsrück regions and in the Harz. The lynx has reappeared in the areas near the Czechoslovakian border. The polecat, marten, weasel, beaver, and badger are found in the central and southern uplands. Among the reptiles are salamanders, blindworms, various lizards and snakes, of which only the adder is poisonous.

Threatened and reappearing animal species

Although the numbers of the white stork, the cormorant, the horned owl, the fishhawk, and the osprey have decreased to near the point of extinction, the golden eagle has gained in numbers in the Alpine districts, and the heron still holds forth in scattered areas over various parts of the country. The republic has established upwards of 50 natural parks and endowed five major resort areas, such as portions of the Schwarzwald, with the status of natural park. In addition, it has declared some 1,000 localities as wildlife preserves and 7,200 sites as protected scenic areas. A further 40,000 locations of unusual historical interest or natural beauty have been set aside as natural monuments.

CHARACTER OF THE TRADITIONAL REGIONS

The boundaries of the modern *Länder* of Germany were established in comparatively modern times. The more than 2,000 political units, large and small, that existed under the Holy Roman Empire gradually were consolidated through war, treaties, dynastic marriages, large-scale amalgamations effected under Napoleonic rule, further consolidations brought about by the establishment of the German Empire, or *Reich*, in 1871, and even, in the wake of World War II, by plebiscite. Transcending and often defying these modern political borders, however, are the traditional regions of Germany, the characters of which are chiefly ethnic in nature and often prehistoric in origin. These regions have persisted with remarkable tenacity throughout the turbulent history of the German territories and the radical social upheavals of recent times. In addition to what might be described as the "tribal" cohesion of the regions, the affinities of linguistic dialect, traditional political allegiances, and, often as not, religion will in different measure serve not only to set off one region from another but also to account for the various subdivisions within the major regions themselves.

Ingrained feelings for the regional homelands

Regional feelings are still very strong in Germany. Just as a Briton of Celtic stock may call himself a Scotsman or Welshman, even though his family has resided for years in England, or as an eighth-generation American will often describe himself as Scots-Irish or Dutch, a German will claim descent from the region of his ancestors no matter how remote the ties may be. To this per-

sistent consciousness of one's ancestral origins must be affixed the German's deeply felt sense of *Heimat*, his "homeland" (or often the homeland of his forebears), which connotes a much narrower sense of native region, viz., the valley or forestland, plain, mountain, or even village or town where one's ancestral ties run deep. These feelings prevail less strongly among the postwar generation, however, and have been dissipated by the enormous influx after World War II of German expellees and refugees from East Germany and the former provinces now incorporated into Poland.

Many of the regional names are no longer reflected in any formal political designation, but they are still used in everyday parlance. In many instances, political consolidations, especially in the south, have not only cut across the conventional boundaries of the ancient regions but also have brought together under a single political administration regions that were, at best, disparate and, at worst, disdainful or even hostile to one another.

Southern regions. In the south the major regions are largely within the *Länder* of Baden-Württemberg, Bavaria, and Rheinland-Pfalz. Baden extends from an area east of the Rhine River and north of Karlsruhe southward and along the Rhine and through the Schwarzwald to the Swiss border. The natives of Baden have close ties and affinities with the northwest area of German-speaking Switzerland and with Alsace, the former German province across the Rhine; like the Swabians, they are of ancient Alemannic stock. North of the Baden region, beginning south of Heidelberg and Mannheim and reaching across the Rhine is the region traditionally known as the Pfalz (Palatinate). The Saarland is a region unto itself, but it has characteristic aspects of both the surrounding Pfalz and the neighbouring French region of Lorraine, which used to be a province of Germany.

Schwaben (Swabia) is a notable example of a region the name of which no longer is affixed to any political entity (or component of a political division), yet its extent transcends *Länder* boundaries. It embraces most of the former states of Württemberg and Hohenzollern, but its eastern portion between the Iller and Lech rivers, roughly between the cities of Ulm and Augsburg, was ceded to Bavaria in 1806. The Allgäu district, in a pocket in the extreme southwestern corner of Bavaria, not only is Swabian in character but also has close historic ties with the Vorarlberg district of Austria, itself settled by Alemannic tribes. Schwaben may be roughly divided into Oberschwaben (Upper Swabia), the district south of Stuttgart leading to Lake Constance, and Niederschwaben (Lower Swabia), from Stuttgart northward to the area around Heilbronn.

The region popularly thought of as Bavaria—a land famed for its Alpine costumes, the *Dirndl* for women and *Lederhosen* for men, and fabled for gargantuan consumption of beer to the accompaniment of music from a brass band—is more correctly Upper Bavaria. It comprises only the southeastern quadrant of West Germany's largest *Land*—east of the Lech and south of the Danube eastward from Donauwörth. The northern and geographically larger portion of the *Land*, Bavaria is almost completely occupied by the region of Franken (Franconia)—its name deriving from its East Frankish settlers. Oberfranken (Upper Franconia) takes in the highland district near the East German and Czechoslovakian borders near Bayreuth; Mittelfranken (Central Franconia) is located in the areas surrounding Nürnberg, while Unterfranken (Lower Franconia) centres around Würzburg. Portions of northeast Baden-Württemberg are also Franconian in origin and character.

Northern regions. The regions north of the Main River, with certain exceptions, are larger and more homogeneous, except for the inroads of modern internal migration in such large urban districts as the Rhine–Main and Rhine–Ruhr metropolitan areas. Only the Rheinland, consisting of the areas of the west bank of the Rhine from Mainz to Cologne, differs notably from most of the regions north of the Main. As a region it enjoys an international fame surpassed only by that of Bavaria. The Mosel and Eifel districts to its west are of a perceptibly different cast. On the opposite bank of the Rhine, the state of Hessen, in spite of its relatively large size, is one of the few states that can claim an almost complete identity of *Land*, region, and ethnic origin; the Hessians are a branch of the Franks in the ancient land of the Chatti.

The Rheinland, by contrast, is politically wedded today to the very different regions of the Sauerland, east of the Rhine opposite Cologne, and to Westphalen, with its close affinities in character and outward appearance to the eastern provinces of The Netherlands. The Ruhr itself is sometimes called the *Kohlenpott* ("coal pot"), a derisive reference to the modern migration of workers from the German eastern provinces and Poland to tend its mines and factories.

Niedersachsen, apart from incidental subregions demarcated by landscape, such as the Lüneburger Heide (Lüneburg Heath) and the Harz, divides roughly into the regions of Hannover and Oldenburg, except for its coastal area, East Friesland, which has close ties to West Friesland in The Netherlands and North Friesland off the coast of Jutland. The slip of coastline extending from the Dutch border up around the shore of the North Sea to Schleswig-Holstein is also sometimes referred to by the Low German term *Waterkant*, "water's edge." The ancient Hanseatic cities of Bremen, Hamburg, and Lübeck may be ranked as separate regions in their own right; all have in common a sense of self-sufficiency, independence, and aloofness from the remainder of Germany. The amalgamated *Land* created from the provinces of Schleswig and Holstein, once duchies under the Danish crown, although not settled by tribes of Danish stock, retains obvious ties of kinship and culture that extend up the peninsula to its Jutlandic cousins across the border in Denmark.

Varieties of the people. Although such generalizations fail to account for the exceptions, it can be stated that the people of the regions north of the Main River tend more toward the popular notion of the blond, blue-eyed, sobersided, hard-working, and often dour German of legend. A conspicuous exception to this stereotype would be that of the Rhinelander, noted for his carefree disposition and almost Gallic affinities with Latin Europe. The Franconians in the south also share in the north German profile of being rather sturdily given to work and no nonsense, and they are rather conservative in the preservation of their traditional ways. A great contrast is found between the easy-going natives of Baden and their fellow Alemannians the Swabians, who among other Germans are thought to be the hardest working, the thriftiest, the most inclined to introspection, and one of the most self-sufficient of German peoples. No greater regional contrast could exist than that between the Swabians and the Bavarians, the latter noted for their garrulous and sometimes coarse temperament, rough humour, and vigorous pursuit of fleshly pleasures. But, if Schwaben has furnished Germany with its poets and great intellects, Bavaria is the home of its artists.

PATTERNS OF HUMAN SETTLEMENT

Rural West Germany. The manner in which land under cultivation in West Germany is divided and the major patterns that have prevailed in the utilization of crop-growing and grazing lands and timberland have been determined not only by the somewhat severe limitations of the soil but also by regional convention and historical circumstance and, especially in modern times, by economic expediency. The disposition of arable land and the structure of rural settlement have undergone steady change over the centuries, as circumstances have dictated, until the major land reforms and the onset of industrialization in the 19th century. But the most radical and abrupt adjustments in agriculture and the rural population have occurred since World War II.

Under the traditional patterns that prevailed throughout medieval and early modern times, to some degree even until the mid-20th century, the simplest structural form of farmland was that of the *Einödflur*, or *Einzelhof*, the individual farm unattached to a community or village. This type of farming in detachment has been rather

Historic methods of farming and land use

the exception in Germany, and it is widespread chiefly in the grazing lands of the northwest. Numerous variations are encountered in the more complex patterns, usually centred around a village, often in conjunction with a large estate. To these variations must be added the nature of tenancy—whether the land was the property of a *Gutsherr*, a large landholder, usually of the nobility; a *Grossbauer*, a peasant or farmer with large holdings; a *Kleinbauer*, a small farmer-proprietor; or a *Pächter*, a tenant farmer, large or small.

The ancient division of land, the *Hufe*, comprising 60 to 100 acres (25 to 40 hectares), was conceived of as the amount of land, arranged into one or several strips, that was necessary for growing enough crops or for maintaining sufficient livestock to sustain one farming operation. In the larger village enterprises and especially common in most of the regions now forming the Federal Republic, the *Hufen* were arranged into *Gewannfluren*, or the acreage surrounding a village to which the farmers went from their villages to tend their crops on the strips of land allotted to them or, in some instances, on the communal property of the village. Of this latter form, large tracts of forestland are still under the ownership of the *Gemeinde*, or "community."

The patterns of the *Hufenfluren*, or "acreage," differed widely, depending on the type of soil, the terrain, and the region. As often as not, a farmer's land would consist of several plots or strips frequently separated from one another. The divisions could also be quite irregular and uneconomically small. For this reason, a major reform in German agriculture in the postwar era has been the attempt to bring about a more efficient arrangement in the division of agricultural lands through a series of consolidations and redistribution of acreage. Along with this realignment of agricultural properties, largely accomplished by land sales and exchanges, has come the gradual disappearance of the small farmer and of multi-crop or self-sustaining farms in favour of the large-scale farmer concentrating on one or a few selected crops, the nature of which is likely to be determined by planning requirements at the *Land* or *Bund* (federal) level and adjusted to fit the crop quotas and price structures indicated by West Germany's membership in the European Economic Community. As in other industrial countries, the migration from the land continues: whereas in 1961, about 13 percent of the working population was engaged in agricultural pursuits, including fishing, by 1970 this figure had decreased to 8.9.

Village patterns

German rural settlement follows certain set patterns—from the relatively infrequent individual farm to the *Weiler*, a loose collection of farms found most frequently in the southwest and the Alpine forelands, to the many variations of the *Dorf*, or "village." By far the most common form of village in the Federal Republic, representing as it does the historically older portion of the German regions, is the *Haufendorf*, a conglomerate village of irregular pattern. Typically, it comprises a main thoroughfare and a square or village centre with such customary fixtures as a church, local administrative offices, marketing facilities, inns, shops, and the like and an assymetrical patchwork of streets and roads leading from the centre.

Other patterns of village structure include the *Rundling*, a round village (the form of which was determined by economic function and not, as was commonly supposed, by considerations of defense), and the *Reihendorf*, a linear village, the simplest form of which is the *Strassendorf*, in which the houses, buildings, and farm structures simply line the highway. These latter two forms are much more common in present-day East Germany, the lands of which were settled relatively later than those in West Germany. Among the minor variants found in all areas of West Germany and determined by terrain are villages formed in forest clearings, mountainous areas, drained marshland, or moors. The transition of a village to the rank of a town historically was betokened by its being granted the privilege of the *Marktfleckenrecht*, the privilege of holding a market. Today, however, a village may lose its essentially agricultural character by the sudden appearance of a factory or an urban housing tract on its outskirts, or, especially in the southern regions, through a reorientation of its economy toward tourism.

Urban West Germany. In the Federal Republic, as in other highly industrialized nations, well over one-half of the population reside in urban areas of 10,000 or more, while almost one-third of the remainder still live in communities of fewer than 5,000 inhabitants—many of which, nevertheless, are satellite or dormitory adjuncts to larger municipalities. Germany is essentially a country of small- to medium-size towns and cities, which, apart from the divided metropolis of Berlin, include only two cities, Hamburg and Munich, of more than 1,000,000 inhabitants. Nevertheless, almost 47 percent of the population live within the spheres of commercial and urban influence of only 10 major metropolitan centres: in approximate order, Rhine–Ruhr, Rhine–Main, Hannover–Braunschweig, West Berlin, Stuttgart, Hamburg, Munich, Bremen–Oldenburg, Mannheim–Ludwigshafen, and Saarland.

Metropolitan areas

The Rhine–Ruhr area, by far the greatest aggregate of heavy and light industry of almost every description, is Europe's greatest industrial and commercial complex. The Hamburg and Bremen metropolitan areas are primarily centres of shipping, trade, maritime industries, and manufacturing. The Hannover–Braunschweig region is largely a manufacturing and trading centre, as is West Berlin. The Rhine–Main area is important not only for certain heavy industries but also as a general commercial centre; because of its central location, it is also a crossroads of transport. The greater Stuttgart area, apart from a few key heavy-industrial firms, has West Germany's major concentration of small- to medium-size industries, mainly in the processing fields. Since World War II, Munich has grown considerably as a region of small manufacturers and processors, as well as being the urban focal point of the southeast. The Saarland is important primarily for its coal, iron, and steel industries. West Germany's two great centres of finance are Frankfurt am Main and Düsseldorf, and, with Munich and Hamburg, are favoured as headquarter cities for the German or European subsidiaries of the large international corporations.

Germany, Federal Republic of, Area and Population

	area		population	
	sq mi	sq km	1961 census	1970 census
States (*Länder*)				
Baden-Württemburg	13,803	35,750	7,759,000	8,895,000
Bayern	27,238	70,547	9,515,000	10,479,000
Bremen	156	404	706,000	723,000
Hamburg	291	753	1,832,000	1,794,000
Hessen	8,151	21,111	4,814,000	5,382,000
Niedersachsen	18,304	47,408	6,641,000	7,082,000
Nordrhein-Westfalen	13,145	34,044	15,912,000	16,914,000
Rheinland-Pfalz	7,659	19,838	3,417,000	3,645,000
Saarland	991	2,568	1,073,000	1,120,000
Schleswig-Holstein	6,052	15,676	2,317,000	2,494,000
Berlin (West)*	185	480	2,197,000	2,122,000
Total West Germany	96,094†‡	248,882†‡	56,185,000†	60,651,000†

*Berlin is under tripartite (France, United Kingdom, United States) jurisdiction and is only administratively a part of the Federal Republic of Germany. †Figures do not add to total given because of rounding. ‡Includes 118 sq mi (305 sq km) of West German portion of Lake Constance.
Source: Official government figures.

II. The people of West Germany

DIVERSE GROUPINGS WITHIN THE POPULATION

The German-speaking peoples—to whom must be accounted not only the inhabitants of the Federal Republic and East Germany but also those of Austria, the major parts of Switzerland and Luxembourg, small portions of France and Italy, and the remnants of German communities in eastern Europe—are extremely heterogeneous in their ethnic origins, in their dialectal divisions, and in their political and cultural heritage, in which the split between Protestant and Catholic has played a significant role since the Protestant Reformation and Catholic Counter-Reformation in the 16th century.

The areas of central Europe called Germania by the Ro-

Interacting factors in cultural patterns

mans (for only one of numerous tribes in the regions north of the Alps) were known as Germany (Deutschland) only in an imprecise geographical sense until the founding of the *Reich* in 1871. Until that recent time, Germany had only the loosest bonds of a multidialectal language, certain cultural affinities, disjointed political affiliations, and, to an even lesser degree, ties of kinship upon which to form any sense of nationhood or national consciousness. Certainly Germany was far short of being a nation-state in the manner of France or England. The ethnic and linguistic subdivisions of today's Federal Republic of necessity transcend political boundaries, even though West Germany, by and large, embraces most of the territories that comprised the nucleus of the German nation—of both the Roman Germania and the Carolingian East Frankish kingdom.

The Germanic, or Teutonic, element represents only a portion of the ethnic profile of what are known as the German peoples. Account must be taken also of a considerable substratum of the pre-German Celtic inhabitants in the west and of numerous Slavonic strains in the east. Whatever internal cohesion existed in the past within the borders of present-day West Germany has been obscured to some extent by the absorption of some 13,000,000 expellees and refugees from the territories to the east. The Federal Republic has recently acquired a substantial non-German minority of over 2,300,000 *Gastarbeiter* ("guest workers"), migrant labourers most of whom are from six nations of southern Europe.

Ethnic divisions. The traditional ethnic divisions of West Germany, coinciding in their larger dimensions with the major regions, go back to tribes and branches of tribes that had settled mainly there by the time of the Roman occupation. Their ultimate origin is still a matter of speculation if not dispute, but it is commonly held that they moved into central Europe from the Baltic regions and either drove the resident Celtic tribes farther west or partly absorbed them. The Celtic strain is notably marked by one of the most enduring forms of evidence: the place-names of rivers, mountains, settlements, and other landmarks that have come into the German nomenclature most often from the Latin rendition of the original Celtic term. The names of the Rhine, Danube, and Neckar rivers, to cite only the best known, are derived from the Celtic.

Six major German ethnic groups have persisted in their historic regions from Roman and early medieval times to the present, and their numerous subdivisions are of no small historic, linguistic, and cultural importance. The southwest of the modern Federal Republic was settled by the Alemannic tribes, who occupy not only the German regions of Baden and Schwaben, but also the French province of Alsace, the German regions of Switzerland, and the Vorarlberg province of western Austria. The southeast, roughly south of the Danube River and east of the Lech River, was settled by the Bavarians, who share ties of kinship and linguistic dialect with the greater portions of modern Austria.

The central districts of West Germany were occupied by the Upper Franks—the Main, Rhine, Mosel, and Hessian Franks—as well as the Thuringians. The Lower Franks find their present-day descendants in portions of Westphalen, Flanders, and The Netherlands. The Thuringians, moving eastward, colonized the areas now in East Germany that formerly were known as Saxony.

The greater portion of the North German Plain, inhabited by the "original" Saxons, is preserved in the name of the *Land* of Niedersachsen. The descendants of these ancient Saxons, however, have few kinship ties with the people of the later and now defunct region of Saxony in East Germany. They comprised one of the several west Germanic tribes of the North Sea littoral, including the Angles, who invaded Britain in the 5th century—hence, "Anglo-Saxon."

The sixth major group, still intact if now small in number, is that of the Frisians in the islands off the coasts of Jutland and Niedersachsen. They are closely related to the West Frisians in the northern portions of The Netherlands.

The German ethnic groups of the east had their origin in later colonizations from the west, beginning in the 9th century; they are known generally as the Mecklenburgers, the Upper Saxons, the Brandenburgers, the Silesians, the Pommeranians, the Prussians, and the Baltic Germans. During these eastern colonizations, the indigenous Slavonic and Baltic peoples were absorbed and Germanized in great measure. To the south, Austria, in German Österreich (Eastern Realm, as both the Latin and German versions of its name indicate), represents an eastward colonization by the Bavarians, of whom the Tiroleans form an ethnic subgroup.

Minority groups in the population

Chief among the historical minority groups in Germany were the Jews, present in western Germany since early medieval times; they numbered about 800,000 in all of Germany prior to the genocide of World War II. Small cultural pockets of the original Slavic inhabitants of the regions colonized east of the Elbe still exist in East Germany, and their language, Wendish or Sorbian, still survives. The persecutions of the French Protestants in the late 17th century brought large numbers of Huguenots to the German lands, especially to the cities of the northwest and to Berlin. The provinces of Schleswig and Holstein, duchies of the Danish crown until the 1860s, were inhabited largely by peoples of west Germanic as opposed to north Germanic or Scandinavian stock; thus, only a small minority of Danes, estimated at 3 percent of the population, still live there.

Various other minorities, large and small, that may be noted in German territory in relatively modern times include the Polish workers who migrated to the Ruhr in the late 19th century; small groups of Russian émigrés who settled in Germany, especially in Berlin, after the October Revolution in Russia; the countless numbers of "displaced" or "stateless" persons from eastern Europe and the Balkans who were stranded in Germany at the end of World War II, and a minority that may well become permanent—the *Gastarbeiter*, their families, and descendants.

Linguistic diversities. The dialectal divisions of Germany, once of conspicuous significance for the ethnic and cultural distinctions they implied, persist in spite of such levelling and standardizing influences as mass education and communication and in spite of internal migration and the trend among the younger, better trained, better educated, and more mobile ranks of society to speak a standard, "accentless" German. The repository of dialectal differences thus lies more with the rural populace of the regions and the long-time native inhabitants of the cities.

Standard German itself is something of a hybrid language in origin, drawn from elements of the dialects spoken in the central and southern districts but with the phonetic characteristics of the north predominating. The pronunciation of standard German is, in fact, an arbitrary compromise that gained universal currency only in the late 19th century. Even today, the most "accent conscious" of the well educated will speak with the coloration of his native district's dialect and, if from the southern regions, all the more intently so.

The dialects of German

The three major dialectical divisions of Germany—as with the ethnic divisions—coincide almost identically with the major topographical regions of the North German Plains, the Central Uplands, and the southern Jura, Danube Basin, and Alpine districts. In reverse order, these dialectical divisions are known as Upper German, Central German, and Low German.

Of the Upper German dialects, the Alemannic branch in the southwest is subdivided into Swabian, Low Alemannic, and High Allemannic. The first, the most widespread and still-ascending form, is spoken to the west and north of Stuttgart and as far east as Augsburg. Low Alemannic is spoken in Baden, southern Württemberg, and Alsace, while High Alemannic is the dialect of German-speaking Switzerland. The Bavarian dialect, with its many local variations, is spoken in the areas south of the Danube River and east of the Lech River and throughout all of Austria, except in the Vorarlberg district, which is Swabian in origin.

The Central German, or Franconian, dialects and the Thuringian helped to form the basis of modern standard German. The present-day influence of Thuringian touches only partly upon West German territory, but it is of great significance in areas now in East Germany. East Franconian is spoken in northern Bavaria, South Franconian in northern Württemberg. The Rhenish Franconian dialect extends from approximately Metz, in French Lorraine, northwest through the Pfalz and Hessen. Mosel Franconian extends from Luxembourg through the Mosel Valley districts and across the Rhine into the Westerwald. Ripuarian Franconian begins roughly near Aachen, at the Dutch–Belgian border, and spreads across the Rhine between Düsseldorf and Koblenz into the Sauerland.

The dialect known as Low German, Plattdeutsch, historically covered all regions occupied by the Saxons and spread across the whole of the North German Plain. Although largely displaced in common usage by standard German, it is still widely spoken, especially among the elderly and rural inhabitants in the areas near the North Sea and the Baltic. Tiny dialectal pockets of Frisian, linguistically the most closely related of German dialects to English, still persist.

Religious profiles. Germany, ever since the Protestant Reformation and the ensuing Thirty Years' War, has notably been divided into the sectarian lines of Catholic and Protestant, though in ways that far transcend considerations of religious doctrine and practice. Following the precept formulated in the Peace of Augsburg (1555), whereby the inhabitants of a given region remained Catholic or became Protestant according to the choice of their temporal ruler, the matter of whether a person or a region is Catholic or Protestant has had the greatest effect not only on such subjective factors as culture and personal attitudes but also on the entire social and political profile of Germany. The very name of the numerically largest political party, the Christian Democratic Union, attests to the fact that it is a federation of an older middle class Catholic party with its Protestant counterparts. The Social Democratic Party, by contrast, owed much of its early orientation not only to Marxism but to antisectarianism.

Origin and political impact of religious division

In today's Federal Republic, without the overwhelmingly Protestant areas now comprising East Germany, the balance between Protestant and Catholic is almost equal. By a 1970 count, Protestants constituted about 49 percent of the population; and Roman Catholics, about 45 percent. A minute fraction of the German public belong to what are known as the free churches, such as independent Protestants (Lutherans or Calvinists), Old Catholics, Methodists, Jehovah's Witnesses, and Eastern Orthodox. The remainder, those people professing no religion (*Konfessionslose*), have been sharply on the increase in recent years, not only because of a slackening of religious loyalties but also as a means of avoiding the "church tax." That relic of Bismarckian Germany requires a citizen to pay a certain percentage of his net income tax to whatever church he acknowledges, despite the fact that there is no established state church. To become eligible for this tax exemption, a citizen must formally renounce his religious affiliation before a civil authority.

The predominantly Roman Catholic districts are found in the Rheinland and Upper Bavaria. The Rhine–Main, Pfalz, and southwestern districts are mixed, while north Germany is overwhelmingly Protestant. Germany now has 32,000 Jews, some 27,000 of whom are formally attached to Jewish communities. The greater number are concentrated in the cities of Berlin, Cologne, Düsseldorf, and Frankfurt am Main, as well as in Bavaria.

CONTEMPORARY DEMOGRAPHY

Vital statistics. The Federal Republic, like most comparable advanced industrial countries of western Europe, manifests a relatively low birth rate coupled with the expectancy of increasing longevity, with births barely outnumbering deaths each year. While the death rate has remained relatively constant in the years 1950–71, averaging around 11.5 deaths per 1,000 of population, the birth rate has varied in the same period from 16.2 in 1950 to a high of 18.3 in 1963 to a low in 1971 of 12.8. Between 1950 and 1971 the average net gain has been 5.2 per 1,000 of population, or an average annual gain of births over deaths of about 291,000 over the two decades.

Only in Hamburg and, to a much greater extent, in West Berlin does a net population loss occur through a preponderance of deaths over births. The reason is a disproportionately high percentage of elderly persons in both of those cities. West Berlin, because of its isolated situation, constantly is confronted with the problem of keeping or attracting younger residents. At the census of 1970, almost 47.5 percent of its population was aged 45 or more, and about 45 percent of that group was 65 or over.

Demographers expect that the population growth rate will be retarded in the immediate decades ahead. In 1972 the net gain in population dropped to an all-time postwar low to −0.2 net births per 1,000. Although the average age for first marriages has become steadily lower, the peaks of child-bearing years for German women were exceptionally late: around ages 29–30. One forecast for the population of the Federal Republic for the year 2000 is given at 67,000,000, as against 61,500,000 in 1970.

Migratory patterns. An additional factor of the population's composition is the presence of about 2,500,000 aliens, excluding foreign military personnel and their families. Of these resident aliens, two-thirds to three-quarters are migrant workers, the *Gastarbeiter* from southern Europe whose residence is mostly of limited duration. Because of West Germany's chronic labour shortage, it can be expected that, even with fluctuating rates of employment, they will constitute a steady if constantly changing minority. Relatively few of the *Gastarbeiter* bring their families to West Germany or marry during their stay in the country, and most will ultimately return to their home countries. It is apparent, however, that many will have settled permanently and will constitute a consistent minority. In descending order of numbers, these immigrant workers are drawn from Yugoslavia, Italy, Turkey, Greece, Spain, and Portugal.

Influx of the "guest worker"

The large numbers in this intrinsically mobile category somewhat distort the profile of internal migration in the Federal Republic, which during 1970 showed some 3,661,500 changes of residence. Among the native German population, however, two trends are discernible, both following the pattern of other comparable industrial nations: first, young Germans are more mobile than their elders and are more willing to move from one part of the country to another as their work requires. Second, the greater amounts of leisure time and disposable income tend to direct young Germans toward the southern regions, in which more attractive leisure-time amenities can be found and which are closer to the favoured holiday spots within the Federal Republic itself and in Austria, Switzerland, and the Mediterranean countries.

Since Germany's economic recovery in the 1950s, the net loss of population through emigration to such nations as the United States, Canada, Australia, New Zealand, South Africa, and, to a lesser extent, those of South America has shrunk to negligible proportions: about 15,000 to 25,000 a year. Declining overseas job opportunities as well as a growing reluctance to leave the excellent living conditions of West Germany and western Europe in general have contributed to this lowered rate of emigration.

Population distribution. Although almost one-half of West Germany's population live in 10 major metropolitan regions, the country still enjoys a highly favourable distribution of population. Other European countries are often dominated by a single metropolis or have only a few metropolitan regions that constantly grow at the further expense of the less populous or less prosperous districts. This rather even population spread is one beneficial dividend of Germany's traditional regionalism, a factor that has militated against the dominance of any one city or district. Nor is economic opportunity in the Federal Republic essentially confined to the great industrial or metropolitan regions. Upwards of one-third of the population still reside in villages of fewer than 5,000

persons and one-fourth in small- and medium-sized towns of from 10,000 to 100,000. The spread of light industry to previously sparsely populated and underdeveloped areas, along with the Federal Republic's continued improvement of its already excellent transport facilities, will probably maintain West Germany's comparatively even distribution of an otherwise dense population of 640 persons per square mile (247 per square kilometre).

(G.H.K.)

III. The national economy

West Germany is western Europe's greatest economic power in terms of total gross national product (GNP), ahead of both France and the United Kingdom. In terms of GNP per capita, however, West Germany trailed Sweden, Switzerland, and Denmark in the early 1970s. In the world, it stands fourth in terms of total GNP after the United States, the Soviet Union, and, since 1968, Japan. West German industrial production grew by 74 percent between 1960 and 1970, an average annual GNP growth, discounting inflation, of about 4.7 percent—below the growth rate of Japan, Italy, and France, above that of the United States and the United Kingdom.

Comparative standing in the international marketplace

If the post-World War II "economic miracles" of West Germany and Japan are compared, it appears that West Germany's economic growth has been the less spectacular and that its miracle ended at about the time of the first upward revaluation of the Deutsche Mark in 1961. The great immigration into West Germany from the German Democratic Republic and other parts of eastern Europe between the end of the war and 1961 was probably largely responsible for West Germany's extraordinary economic upsurge in the 1950s, whereas forces of a more dynamic nature produced Japan's rapid rate of growth.

Japan is a particularly troublesome rival for West Germany, for the two countries deal in many of the same industrial products. West Germany's export performance in the early 1970s, however, remained impressive. In 1970 it was the world's second-largest exporter by a wide margin. In 1971 West Germany was also the world's second-largest importer as well, followed by the United Kingdom, Japan, and France.

West Germany is a member of the European Economic Community (EEC). With increasing European economic and monetary integration, its economic strength will be affected increasingly by that of the other EEC members.

EXTENT AND DISTRIBUTION OF RESOURCES

Minerals. West Germany's most important mineral resources are coal, petroleum, natural gas, and iron ore. Deposits of hard coal are concentrated largely in the Aachen and Ruhr areas (Nordrhein-Westfalen) and in the Saar. The Ruhr deposits alone are estimated at about 65,000,000,000 tons; the seams are deep, often descending to more than 3,000 feet. Uneconomic mining methods, however, as well as the shift of consumer demand to oil and natural gas, caused a serious decline in the coal industry in the 1960s. West of Cologne and, to a lesser extent, in Hessen and Bavaria, deposits of lignite, a fuel intermediate between peat and soft coal, reach to a depth of 330 feet; they are usually surface mined. Total accessible reserves are estimated to be nearly 5,000,000,000 tons.

Coal, oil, and natural gas reserves

Petroleum deposits are located mainly in Emsland in Niedersachsen and on the coast of Schleswig-Holstein. Output in 1971 met about 8 percent of domestic demand, and reserves were estimated to be 89,000,000 tons. Deposits of natural gas occur mainly in the northwest, in the area between the Ems and Weser rivers and around the mouth of the Ems in Niedersachsen. Reserves of natural gas have been put at 507,000,000,000 cubic yards (388,000,000,000 cubic metres). According to unofficial estimates, however, as much as 520,000,000,000 cubic metres might become proved.

West Germany's best iron-ore deposits are located along the Lahn, Sieg, and Dill rivers in Nordrhein-Westfalen; low-grade ore is found in the Harz district near Salzgitter in Niedersachsen. Total iron reserves are estimated to be about 3,000,000,000 tons, low grade. In the 1960s output declined; total iron ore mined in 1971 amounted to 6,400,000 tons, with an iron content of 1,800,000 tons. Of great importance are West Germany's deposits of potash, its annual output accounting for about one-fifth of the world's production. Among other minerals mined in West Germany, in relatively small quantities, are copper, lead, zinc, and bauxite.

Biological and hydroelectrical resources. In 1970, more than one-half of West Germany's area was devoted to farming, although the cultivated area decreased slowly but steadily in the 1960s with continuing industrialization and urbanization. Arable land suited for temperate crops accounted for about 60 percent of this agricultural land, meadows and pastures for the remainder. The important dairying districts are in the northern coastal lowlands and in Bavaria. Orchards are mainly in the west; vineyards, in the Mosel, Rhine, Neckar, and Main valleys.

Between 1949 and 1965, reafforestation increased the woodland area by 4 percent, and, by the end of the decade, forests represented almost 30 percent of the nation's land area. Two-thirds of the forests are coniferous. The *Länder* with most woodland are Baden-Württemberg, Bavaria, Hessen, and Rheinland-Pfalz. About 30 percent of the total forest area is owned by the federal government and the *Länder*, another 30 percent by municipalities and corporations, with the remainder privately owned.

West Germany's hydroelectric stations are located mainly in the Bavarian Alps. In the late 1960s, output of hydroelectricity decreased slightly, but rose 21 percent in 1970. Long term trends favour increased production of nuclear energy, although output of hydroelectricity could be stepped up by inundating larger valleys.

SOURCES OF NATIONAL INCOME

Agriculture, forestry, and fishing. During the 1950s and 1960s the agricultural sector of West Germany's economy underwent a dramatic decline. One-third of all farms disappeared, and two-thirds of all persons employed in agriculture had to give up their jobs. During the 1960s the number of people employed full-time in all of these fields fell from 3,600,000 to 2,400,000. In all likelihood, this development will continue throughout the 1970s.

Decline in farming

In spite of the continuing drift of farmers into other sectors of economic activity, however, West Germany in the early 1970s was still a country of peasant-type farming. As a result of traditional partitioning among heirs, the land is divided up into a large number of relatively small units, though concentration took place in the 1960s. By 1971, farms totalled 1,200,000, more than 500,000 of which were less than 12 acres (five hectares) in size.

Since 1955, the federal government has drawn up an annual plan providing for structural improvements and for farm income subsidies to make up for low prices. The entire farm price system is governed by the EEC common agricultural policy. Decisions for promoting land reform were taken by the EEC in 1971; subsequently, the West German government laid down a program of supplementary national measures for structural reform and social aid.

From 1960 to 1969 the index of farm output rose by more than 18 percent. As a result of the surplus of labour and the increasing mechanization and improvement of farm structures, productivity also showed some increase during the 1960s. The agricultural sector was unable, however, to narrow the gap between its productivity level and that of the economy as a whole; farms, with less than 10 percent of the republic's workers, were producing less than 4 percent of the GNP. Such measures as retraining agricultural employees, offering incentives for retirement of farmers, and the merging of farming with other pursuits to reduce the number of full-time farmers might increase productivity in the long run.

West Germany must rely on food imports, but it is approaching self-sufficiency in a number of important commodities such as wheat, oats, and potatoes. The country is self-sufficient in rye. The main crops are rye, wheat, barley, oats, potatoes and sugar beets. Livestock farming, however, plays a more important role in the agricul-

tural sector than crop farming, and in most dairy products the nation is self-sufficient.

About one-third of West Germany's timber requirements are met by imports. In a European comparison, West Germany takes fourth place after Sweden, Finland, and France in timber production. Although the fishing catch is important, imports are still needed to satisfy consumption. The principal fishing grounds are in the North Sea, which supplies more than one-quarter of the catch, mainly herrings, and the North Atlantic, especially off Norway, Iceland, and Greenland. West German fishing vessels also ply waters off Labrador, Nova Scotia, and New England.

Mining. Mining and quarrying is a relatively unimportant sector of the West German economy, representing 4 percent or less of domestic product. The coal industry declined steadily throughout the 1960s, and a 1968 law provided for the closing of uneconomic pits. Thereupon all major coal companies joined in a giant holding company, Ruhrkohle AG, to facilitate, with financial help from the government, the orderly slowing down of coal output and diversification of industry in coal-producing areas. In all likelihood, EEC coal production in the future will be concentrated in the Aachen and Saar regions.

Contrary to developments in the coal sector, output of natural gas rose considerably in the second half of the 1960s and was likely to continue through the 1970s in view of the extent of proved reserves. Reserves of crude oil are more limited, and production has stagnated for years. Iron-ore mining is on the decline, and reserves of nonferrous metals are small, though potash mining remains important.

Manufacturing. Manufacturing accounted for about 43 percent of the domestic product in the early 1970s, virtually the same proportion as a decade earlier. With this contribution it represented the most important sector of the economy. The next most important, services, increased its share slightly during the decade. Between 1960 and 1970, manufacturing output increased by more than 75 percent, with production of capital goods rising faster than that of consumer goods. Growth leaders among important capital goods were electrical engineering, transport equipment, and machinery construction. Productivity increased during the 1960s in terms of both output per worker and output per man-hour.

Economic dominance of industry

This excellent industrial performance was to a large extent fuelled by a sustained high level of capital investment, even during periods of economic recession. Like Japan (but at a considerably lower level), West Germany devoted a considerable share of its national income to investment—in 1971, nearly 27 percent. During boom periods the labour market became extremely tight in spite of a continued shift of manpower resources from agriculture to industry and a strong influx of foreign workers. In 1971, almost 13,220,000 persons, nearly 50 percent of the work force, were engaged in the industrial sector, 9,800,000 of them in manufacturing. The largest industry in terms of employment and monetary income and expenditure was mechanical engineering, followed by electrical engineering, chemicals, textiles and clothing, and transport equipment. These industries—apart from textiles and clothing, which are declining—are also West Germany's leading export industries.

In the 1960s, a trend developed toward concentration of enterprises in a number of industrial sectors, especially in steel and petrochemicals. In spite of the government's plans in the early 1970s to frame a new cartel law prohibiting certain large-scale mergers, concentration will probably continue because of the force of world competition. In 1971, West Germany's largest enterprise, the partly state-owned automobile producer Volkswagenwerk, ranked as the fourth-largest company in Europe. With a few such exceptions of partial governmental ownership, West Germany's industry is mainly in private hands.

Foreign participation, especially investment from the United States, was of particular importance in the oil-refining, chemical, electronics, and motor industries. Of the total foreign investments in West Germany, about 90 percent was in manufacturing industries. The major part of West Germany's investment abroad was in other European countries, and to a lesser extent, in the United States and Canada.

Energy. Dramatic changes in the source of energy took place in the 1960s, when the share of hard coal dropped from 61 to 31 percent and that of lignite from 15 to 9 percent, while the share of oil rose from 21 to 52 percent. At the same time, energy consumption increased by 50 percent. With the Coal Adjustment Law of 1968, the government became committed to the gradual elimination of high-cost coal production. Thus, with the increasing importance of oil and natural gas, West Germany has become largely dependent on energy imports. Although domestic oil production seems stagnated, output of natural gas rose rapidly in the 1960s. It is estimated that by 1975 natural gas will supply 10 percent of energy, with imports from The Netherlands and the Soviet Union providing part of the supply.

Over 90 percent of electricity is produced by conventional thermal generators; small additional sources are waterpower and nuclear energy. It was planned to expand the capacity of nuclear-power plants about threefold by 1980. Most of the electric-power and gas plants are publicly owned, but a considerable part of the total volume of electricity and gas is produced by privately owned companies. Since the 1960s, the gas industry has been shifting to natural-gas sources, closing down local gasworks and constructing a pipeline network for the transport of natural gas.

Financial services. From 1948 to 1957, bank notes were issued solely by the Bank Deutscher Länder (Bank of German States), which was at the same time the bank of the federal government. Since then, the functions of central bank and bank of issue have been taken over by the Deutsche Bundesbank (German Federal Bank), which by law has the additional task of watching the stability of the currency. In exercising these tasks, the Bundesbank is not subject to any control by the federal government; as a result, repeated conflicts have taken place between the government and the Bundesbank over fundamental issues concerning the management of the national economy.

Banking and securities institutions

At the end of 1971 there were 311 commercial banks, including the "Big Three"—the Deutsche Bank AG, the Dresdner Bank AG, and the Commerzbank AG, which have a broad network of branches throughout the country. The fourth-largest bank, the Bank für Gemeinwirtschaft AG, is run on a cooperative basis by West Germany's major trade-union confederation. The major German banks not only dominate the financial sector but also have a long history of direct participation in the industrial sector.

There are also more than 100 *Land* and local banks, over 172 private banks, and 23 foreign banks. Of the credit institutions, the state-owned Kreditanstalt für Wiederaufbau (Reconstruction Loan Corporation) is the channel for public aid to developing countries. The banking system also includes over 800 savings banks, for which 12 central *Giro* institutions act as clearinghouses. There are several thousand small industrial and agricultural credit co-operatives and allied institutions; and in addition, private and public mortgage banks, installment-credit institutions, the postal-check and postal-savings schemes.

A characteristic feature of banking business in West Germany is the large amount of long-term finance provided to industry, the government, and local authorities. The public sector's long-term borrowing is controlled, however, while industrial and business borrowing is regulated by the Konjunkturrat (Business Cycle Council).

Although the government had encouraged the purchase of securities by the public, by the early 1970s the results were relatively modest. Investments in savings deposits and similar instruments were growing faster than in securities. The largest part of the securities held were shares of government-owned concerns that had been offered to the small investor at attractive terms; saving and treasury certificates were less popular.

Stock exchanges are located in Bremen, Düsseldorf, Frankfurt am Main, Hamburg, Hanover, Munich, Stuttgart, and West Berlin. In the 1960s share prices were subject to heavy fluctuations, and as a result of the general economic recession a serious stock-market slump occurred in 1966. Share prices recovered, to reach a record level in 1969, but they dropped again in 1970.

Foreign trade. In the 1960s, foreign trade played an increasingly important role in West Germany's economy. Imports of goods and services accounted for 20 percent of the GNP in 1971, exports of goods and services for 21.5 percent. West Germany's competitiveness on foreign markets remained strong throughout the 1960s; even after the upward revaluation of the Deutsche Mark in October 1969, it was hardly impaired, partly because of the readiness of exporters to accept lower profits.

The average annual export growth rate between 1960 and 1970 was 11.1 percent. The outstanding feature of the export pattern is the large proportion of manufactured goods, which in a recent year made up over 85 percent of all exports. The most important export industries are machinery construction, transport equipment, chemicals, electrical engineering, iron and steel, iron and metal products, textiles, precision and optical instruments, and shipbuilding. Although West Germany remains Europe's leading exporter, in the 1960s it lost ground to Italy in such exports as refrigerators and clothing. Japanese competition is most dangerous for key products such as automobiles, cameras, and electronic equipment.

Balance of trade

About one-half of West German exports go to EEC countries, with France the largest national market, followed by The Netherlands, the United States, Italy, and Belgium–Luxembourg. More than 10 percent of its exports go to developing countries, about 5 percent to eastern European nations. The average annual growth rate of imports in the 1960s was of the same order as that of exports. The commodity composition of imports changed to give greater weight to manufactures, less to raw materials. West Germany's best customers were also often its largest suppliers. Of total imports, about 45 percent originated from EEC countries, 16 percent from developing countries, and 4 percent from the Eastern bloc.

Trade with the German Democratic Republic is classified as "interzonal" trade and is not included in the external trade accounts. West German exports to the German Democratic Republic usually exceed imports.

MANAGEMENT OF THE ECONOMY

The public and private sectors. West Germany's system of economic management has undergone a number of shifts since the end of World War II. In 1948, the system of a planned economy, inherited from wartime, was abolished, and until 1965 the government pursued the principle of "social market economy." That principle gave virtually unlimited scope to free enterprise, except to restrict certain practices that restrained competition (though the growth of monopoly power as such was not forbidden). After 1965 a trend toward a more interventionist economic policy developed. Plans for stricter legislation to regulate competition, however, were only a minor part of that trend, when the government began consideration of a stability law the primary aim of which would be to coordinate the fiscal policies of the federal government and those of the *Länder*. The resulting Stabilization and Growth Law of 1967 not only gave the federal government far-reaching control of overall public spending, including that of the *Länder* and municipalities, but also some new weapons with which to influence private demand without introducing supplementary budgets. Under this law, the government in 1970 temporarily suspended investment depreciation allowances and imposed a temporary and repayable surcharge on income and corporation taxes. It also instituted a procedure by which the minister of economics meets at regular intervals with employer and union representatives to persuade them to exercise restraint in price increases and wage demands. In 1970 the government reaffirmed, however, that it would not impose price or income controls. In view of the income gap between persons employed in agriculture and industry, the government is involved in regulating private economic activity in the farm sector, justifying its actions as an instrument of adjustment to Common Market conditions.

Taxation. Responsibility for raising tax revenues is divided among the federal government, the *Länder*, and the municipalities. The federal government draws its revenue mainly from duties on customs and excise (except for the tax on beer) and sales taxes; it also had a share in income and corporation tax revenues, which were at the disposal of both the federal government and the *Länder*. Besides their share in those tax revenues, the *Länder* receive income from property, vehicle, and beer taxes. Municipal revenues are derived mainly from taxes on tradesmen's profits and taxes on real property. Legislation in 1969 provided that revenue from income and corporate, as well as sales, taxes was to be shared by the federal government and the *Länder* in proportions to be fixed annually; in addition, income taxes were to be shared with the municipalities. Special grants were made to the poorer *Länder*. A further reform, placing stronger emphasis on indirect taxation, was scheduled to be implemented at the beginning of 1974.

Divisions of tax sources and revenues

The financial structure is undergoing other changes as a result of EEC policies: the government's revenue from agricultural levies has been flowing into the EEC budget since 1971; its customs duties gradually will be made over to the EEC budget by 1975; and up to 1 percent of its sales-tax revenue also may have to be paid into the EEC budget as of 1975. In accordance with EEC decisions, West Germany switched to a tax-on-value-added system in 1968.

Tax receipts more than doubled during the 1960s, with the most significant development the growing share of revenue from income taxes. During the same period, the share from corporation taxes decreased, while sales taxes changed little in their contribution.

Trade unions and employer associations. Labour relations in West Germany have a strong legal bias. At the federal level, the Federal Labour Court and the Federal Constitutional Court have had considerable influence by defining the limits of the unions' and employer associations' freedom of collective bargaining and by settling important legal and industrial disputes. At the *Länder* level, the labour ministries are in charge of implementing federal labour and social legislation; they also supervise labour courts and social-security institutions. The trade unions and employer associations are represented, however, in the labour courts, labour exchanges, and social-security institutions.

In the early 1970s, somewhat less than one-third of the 27,200,000-man work force belonged to West German trade unions. The major union, the Deutsche Gewerkschaftsbund (DGB; German Labour Federation) is one of the three largest and most powerful union federations in the non-Communist world. Its several affiliated single-industry unions combined in a rigid organizational structure in 1949. The DGB executive board represents the interests of the affiliates in social, economic, educational, and cultural matters. The most influential member within the DGB is the metal-workers' union, IG Metall.

At the federal level, the employers' interests are represented by the Bundesvereinigung der Deutschen Arbeitgeberverbände (BDA; Confederation of German Employers' Associations), grouping 43 large federations and some 800 local federations in West Germany, representing about 90 percent of all privately owned enterprises. For the defense of their interests in fields other than labour policy, employers are grouped in federal organizations, of which the most important one is the Bundesverband der Deutschen Industrie (BDI; Federation of German Industries).

Contract and conciliation procedures between unions and employer associations are complicated. Postwar legal decisions have tended to restrict the right to strike. Before a strike can be called legally, the industry must be without a collective agreement because of the expiration of the old one and a breakdown of talks on a new one;

Restrictions on strikes

both sides must have observed a compulsory three-week period of mediation; and the union must ballot its members and obtain a 75 percent majority in favour of a strike. Serious strikes occurred legally in 1956–57 and 1963, while a series of illegal strikes alarmed the government and the unions alike in 1969. On the whole, the number of working days lost was considerably smaller in the 1960s than in most other industrialized countries. In spite of the 1969 strikes, statistics convey a picture of relatively peaceful relations between labour and management, with working days lost by strikes well below 1 percent of the figure in the United States and Italy and only a fraction of those in the United Kingdom, Japan, and France.

The DGB is a strong advocate of workers' participation in company management. In 1951 it obtained a law on co-determination of policy, which applied to major firms in the coal and steel industries; in the early 1970s it still was seeking an extension of this law to other important industries.

CONTEMPORARY ECONOMIC POLICIES

The Stabilization and Growth Law of 1967 marked a turning point in West Germany's economic policy making. The earlier uncompromising application of the principle of a social market economy created conditions under which market processes ran their course. The federal government had control over only part of public spending, since the budgets of the *Länder* and municipalities were operated completely independently, with the result that all budgets became completely out of control in 1965.

The Stabilization and Growth Law was designed to provide a framework of economic decision making at federal level, especially the coordination and federal supervision of public spending, with a view to achieving price stability, full employment, external equilibrium, and growth. Under the psychological impact of the 1966–67 economic recession, the *Länder* and municipalities accepted the curtailment of their financial autonomy. One of the most important features of the law is that it obliges the government to draw up a medium-term finance plan, and since 1969 an official finance council composed of representatives of local government has played a part in forming this plan.

Contacts with local-government representatives are also cultivated within the framework of the Business Cycle Council, which advises the federal government on bond issues but, in turn, is held to support the government in its attempts to control economic cycles. Almost for the first time since World War II, the government's fiscal policy in 1967 and 1968 was intentionally countercyclical. The periodic meetings with trade unionists and employers initially looked successful, when the unions, under the shock of the 1966–67 recession, kept their wage claims at a very modest level. During the boom years of 1968–69, however, company profits soared at a rate disproportionate to that of wages, leading to the illegal strikes of 1969. Subsequently, the pace of wage increases began to accelerate, accompanied by substantial price increases.

In 1970 both the ministry of economics and the Bundesbank took a series of measures to check inflationary pressures. The ministry of economics pursued a policy of restraint on public expenditure; in addition, in 1970 it introduced a temporary and repayable surcharge on income and corporation taxes and temporarily suspended investment depreciation allowances. In the same year, the Bundesbank raised the discount rate to the then record level of 7.5 percent and introduced measures to keep credit as tight as possible. Subsequently, however, the Bundesbank had to cut the discount rate, in several stages, in view of cuts in interest rates abroad and a heavy influx of foreign money, especially United States dollars, into West Germany.

ECONOMIC PROBLEMS AND PROSPECTS

In international comparisons, West Germany's economic situation in the early 1970s seemed less worrisome than it was viewed at home. The inflation it had to combat in the early 1970s was not an isolated, domestic inflation but a worldwide one. The weakness of the economy, however, is that all of its booms up to 1970 were led by exports, and for that reason the nation's economic development has depended largely on that of its major markets. The most obvious solution, a shift of resources from exports into the improvement of the domestic infrastructure, is difficult to put into practice, however. The government has increased spending to realize part of its promised reform program to raise the standard of schools, hospitals, and social services, but it seems doubtful that a far-reaching reform can be carried out without substantial tax increases. According to estimates, annual budget deficits will occur until 1975. It is also most doubtful whether public spending, which is slower to take effect than private investment spending, will be a fast-acting stimulus to the economy.

At the same time, it is evident that a real need for social reform exists and the call for a more equitable distribution of the country's wealth is likely to become more intensive. In particular, more measures will have to be taken to narrow the income gap between the farm sector and other economic sectors, which is often part of complex regional problems. Thus, in the framework of a regional program drawn up by the government in the late 1960s, official efforts are to be undertaken to attract industry to the large farming areas. In view of these domestic problems, it is almost certain that West Germany's economic growth, at least in the first half of the 1970s, will continue to hinge upon the general development of world trade. (E.I.U.)

IV. Transportation

The Federal Republic not only must serve its own extremely complex needs as a densely populated and highly industrialized nation, but, because of its central location in Europe, it must also function as Europe's principal nation of transit for the passenger and freight traffic. In consequence, its system of overland transport—highways, railways, and waterways—and, to a lesser but adequate degree, of air transport—are among the most intricate and highly developed of any nation. Elaborate plans, made in coordination with neighbouring countries and partners in the EEC, call for further improvements and ambitious long-range programs to expand and improve existing facilities and to incorporate the most advanced concepts in transport engineering to handle the growing volume of traffic anticipated in the decades ahead. More limited-access motorways and widened highways will lessen the present immense burden on Germany's labyrinthine road system, while the continuing modernization of existing rail facilities will be augmented by the construction of new lines to relieve the most heavily travelled routes.

PATTERNS OF TRAFFIC MOVEMENT

Predominance of north–south routes

Although throughout history Germany has served as a crossroads for European traffic, both from north to south and from east to west, the geographic contours of West Germany and its inherited overland routes, as well as existing political conditions, make the north–south and northwest–southeast routes of greater significance than those running from east to west. By its location, it is the one Western country with the greatest number of points of access to the Communist nations of the eastern European bloc, but the volume of passenger and goods traffic to and from these nations, including Germany's estranged sister-state, the German Democratic Republic, amounts to only a small fraction of the volume hauled across its frontiers.

The notable exceptions are the freight and passenger routes to West Berlin, a political and economic adjunct to the Federal Republic: access to the city, except for two of the three air corridors and the Elbe River, is direct east–west. Until the partition of the *Reich* in 1945, Berlin was the focal point of the nation's railway system, but today this function is all but vestigial, even within East Germany.

National routes. West Germany's major overland routes follow parallel patterns in both the highway and rail systems. From north to south, traffic gathering from the direction of Scandinavia focusses at Hamburg and moves largely in two directions. It may follow a course roughly due south via Hannover and Würzburg to Stuttgart and the southwest or from Würzburg eastward toward Nürnberg or southeastward toward Munich. Alternatively, it may go from Hamburg southwestward toward Bremen, whence it veers south and then west to the Rhine–Ruhr district, from which area the major routes of transport follow the Rhine Valley, past the Rhine–Main district to all points south.

Major intermediary routes are the rail and highway links from Hamburg and Hannover to Kassel and Frankfurt am Main; from the Rhine–Ruhr district to Hannover; from Frankfurt am Main to Würzburg, Nürnberg, and Munich; from the Mannheim–Ludwigshafen area westward to the Saar and northern France; and from the Mannheim–Heidelberg area to Karlsruhe or, further eastward, to Stuttgart, Ulm, Augsburg, and Munich. Munich is the great highway and rail terminal point for traffic to and from Austria, Italy, and the Balkans. The major routes to southwestern Europe run south past Karlsruhe and up the Rhine Valley—with a major access to eastern France at Strasbourg—or south past Freiburg, im Breisgau, to the juncture of the German, Swiss, and French borders at Basel.

Germany's unexcelled system of overland transport is complemented by its great endowment of navigable rivers and their interconnections by means of an intricate system of canals. Apart from the Danube River and its tributaries, which flow eastward toward the Balkans but eventually will be linked by canal to the northern water routes, these rivers have access to the North Sea and, via a link with the Elbe, to the Baltic. Although three of West Germany's major rivers—the Elbe, the Weser, and the Ems—empty into the North Sea, a major portion of freight to or from overseas passes through The Netherlands and Europe's largest port, Rotterdam, for the Rhine route via Rotterdam is closer by almost half the distance to the North Sea and the Atlantic than even the Ems. As a corollary, The Netherlands' extensive motor-freight fleet has been developed to service goods passing through the port of Rotterdam, and a major portion of this fleet is deployed to and in transit through the Federal Republic via Nordrhein-Westfalen.

Local and interurban services. West Germany's systems of passenger transit are among the most thoroughly developed and efficient in the world. A dense network of mainline railways not only provides frequent and speedy service in long-distance passenger traffic but also serves simultaneously for local commuter traffic.

Local urban transport is no less intensely structured: local bus systems operate in even the smallest towns, and all communities have access to the regional interurban bus lines. German cities, far from abandoning streetcar and tram services in the post-1945 era in favour of municipal bus lines, continue to rely heavily upon urban rail service; and in the larger cities, municipal buses are usually ancillary to the streetcar and tram lines. Although Berlin and Hamburg have maintained both underground and elevated train services for decades and Wuppertal constructed an 8.3-mile-long elevated monorail in 1900, several other West German cities have begun to construct similar rapid-transit facilities. Most of these rapid-transit systems, in such cities as Frankfurt am Main, Cologne, Stuttgart, and Hannover, represent essentially a submerging of the surface rail lines in the central portions of the city. Only Munich has built an underground rapid-transit system independent of the surface lines, complemented by an independent fast-rail service linking communities within a 25-mile radius of the city centre.

Problems and prospects of municipal services

Interurban bus transport, most of it operated by the federal postal service as a relic of the ancient postal coaches, is limited in most instances to localized suburban and nearby interurban services. Because of the superior train service available, long-distance coach runs play a relatively minor role in passenger transport.

COMPONENTS OF THE SYSTEMS

Of the major forms of long-distance freight carriage of all kinds, including transshipments, the railways haul well over one-third of all goods; inland shipping accounts for almost one-quarter; overseas shipping, for upwards of 15 percent; and motor-freight carriers, for slightly over 15 percent. The remainder of freight transport is conveyed by pipeline (chiefly oil and natural gas), now at some 8 percent, plus a relatively minute fraction of goods shipped by airfreight. The latter two forms of haulage showed the greatest growth (virtually fourfold) during the 1960s. At the same time, water and road transport increased slightly in the proportion of goods carried over that of the railways.

Highways. Approximately 75,000 miles of roads and highways of all categories remained in West Germany after World War II. Even the best of these were antiquated in terms of early postwar traffic demands, and few exceeded more than two lanes in width. Superimposed on the older highway system was the badly damaged system of limited-access motorways, then some 1,330 miles in length, known as the *Autobahn*, begun under the Nazi regime in the mid-1930s and considered the marvel of the world in its time. It was in operation well before comparable superhighways in the United States. But though the *Autobahn* complex linked the major industrial and population centres, many important segments had not been completed by the war's end, and its relatively narrow dual lanes, impeded by war-damaged or incomplete sections, were soon overtaxed by the mounting volume of traffic. Improvements and extensions of the *Autobahn* network were begun in the early 1950s, and by the early 1970s it covered over 3,000 miles. The Federal Republic's present highway system comprises some 100,000 miles (exclusive of municipal street systems), but only about 23,000 miles of these are major arteries. By 1985, the long-distance highway system will have been expanded to approximately 36,500 miles.

The *Autobahn*

Since neighbouring countries have developed most of their own limited-access highway systems to complement the *Autobahn* routes, it is possible to cross directly into the motorways of The Netherlands, Belgium, and Austria. Similar links are in preparation for expressway transit into Denmark, France, and Switzerland. Plans for overhauls that will run into the 1980s involve a trans-European highway system coordinated by the federal and *Länder* governments as well as those of neighbouring countries.

Railways. Germany's first train service operated over a five-mile line between Nürnberg and Fürth in 1835. The privately owned railway corporations formed throughout the late 19th century were transferred to public ownership through the individual states, and by 1920 the entire network was owned and operated by one national corporation. Some minor lines—for the most part narrow-gauge, resort, or excursion lines—survived under private ownership. Prior to territorial annexations from Austria and Czechoslovakia before the outbreak of World War II, Germany had some 35,000 miles of railway lines. About 19,000 miles of this network, much of it heavily damaged during the war, devolved to the Federal Republic, which by 1951 had taken over operation of service from the occupying powers. (Its counterpart in East Germany lost almost all of its dual-track routes to the Soviet Union, as reparation, and only minor portions have so far been restored to multitrack operation.)

As with so many facets of West Germany's infrastructure, widespread wartime destruction called for an almost complete overhaul or replacement of capital stock and equipment, so that the new rail network was rebuilt with the benefit of the latest technology and stocked with new equipment. By 1970 the network had increased its total length to 18,500 miles, over 5,300 miles of which were electrified.

A major innovation since 1967 has been the rapid addition of containerized-freight facilities, facilitating the speedy transfer of cargo between ship, rail, and motor transport. In addition to rapid long-distance passenger

service, the Deutsche Bundesbahn (German Federal Railway), in conjunction with other countries of western Europe, operates luxurious high-speed trains known as the TEE's (Trans-European Expresses) and the Inter-City routes. Attaining cruising speeds of about 90 miles per hour and higher, these express services enable passengers to travel, for example, from Hamburg to Munich, a rail distance of 522 miles, in seven hours. Considerable progress also has already been made toward perfecting revolutionary forms of rail transport in superspeed monorail motor trains, as well as other still experimental innovations in rail transport.

Waterways. Germany's intricate system of inland waterways, of which the Rhine is the very symbol, assumes a vital economic importance in the inexpensive haulage of bulk goods, raw materials, and fuels. Of the 3,700 miles of navigable waterways, about 2,500 are represented by rivers, most of the rest by canals, and a few by lake routes, especially on Lake Constance. About 75 percent of internal water transport is on the Rhine, whose density of ships, barges, freighters, and other vessels from the fleets of five nations is comparable to that of a heavily travelled highway.

The river and canal systems

The most important of Germany's canals is the 200-mile-long Mittelland Kanal, which connects Berlin with the Elbe within East Germany, extends westward across the West German frontier into Niedersachsen and the Greater Hannover industrial region, leads to the Weser at Minden, and continues westward with the Dortmund-Ems-Kanal leading to the North Sea port of Emden. The Kiel Canal (also known as the Nord-Ostsee-Kanal), 61 miles long, has linked the Baltic and the North seas since 1895. Intensified improvements to inland waterways are actively in progress. Both rivers and canals are constantly being deepened and channels widened to increase capacities; lock facilities are undergoing widespread remodelling and enlargement. The most important development in inland shipping both for Germany and all Europe will be the linking of the northern river and canal system with that of the Danube and its tributaries and canals, a project contemplated since the reign of Charlemagne in the 9th century AD. This trans-European link between the North Sea and the Black Sea is expected to be opened by the 1980s.

Although having only limited access to the open sea, Germany long has ranked high as a maritime power: its merchant marine was the fifth largest in the world at the outbreak of World War II. As with other leading shipping nations, however, the Federal Republic's position among the merchant fleets of the world tends to be obscured by the intricacies of postwar patterns of ownership and through ship registration under flags of convenience. Discounting, thus, the ships under, for example, Liberian and Panamanian registration, the Federal Republic held eighth position in the early 1970s. Of its major seaports, Hamburg, noted for its excellent modern and swift facilities for docking and the transfer of cargo, accounted for the greatest number of arrivals and departures. As in other comparable maritime nations, however, passenger traffic to and from overseas dwindled to insignificant proportions because of the ascendancy of airline transportation. On inland waterways, passenger carriers are limited mainly to ferry transport and excursion traffic.

Airways. Germany, apart from its many technical contributions to aeronautics, was one of the great pioneers of commercial aviation in the years between World Wars I and II. The original Deutsche Lufthansa AG, founded in 1926, whence the Federal Republic's state-controlled airline Lufthansa takes its name, was an early leader in operating regular passenger service throughout Europe and by 1936 was flying the North Atlantic route. In 1938, this airline offered nonstop passenger service between Berlin and New York. Since the re-establishment of German commercial aviation in the mid-1950s, Lufthansa has become one of the leading international carriers, with routes around the world.

It is, in fact, in international rather than in domestic aviation that West Germany's role is important, not the least reason for which is the nation's central location and, thus, convenience as a point of arrival and departure in Europe. Because of the relatively short distances between the cities of Germany and the availability of fast and less expensive train service, domestic passenger service represents less than 1 percent of the traffic on public air carriers of all types. Only on the traffic to and from West Berlin do most persons prefer to fly—to avoid passing through East German controls. Air freight represents only a tiny fraction of the total of long-distance goods hauled in the Federal Republic, though its volume grows significantly each year.

Eleven commercial airports serve the major cities of West Germany, of which Frankfurt's Rhine-Main—Europe's third largest in volume after the airports serving London and Paris—far exceeds all others in numbers of passengers and total air freight handled. Because upwards of 1,000,000 visitors fly in to Frankfurt am Main each year to attend its numerous international trade fairs, an underground railway connection links Rhine-Main airport with the fairgrounds. Major new airports are being opened for Munich and Hamburg.

V. Administration and social conditions

The structure and authority of West Germany's government is derived from the *Grundgesetz*, or Basic Law, which was signed and proclaimed on May 23, 1949, after formal consent to the formation of the republic had been given by the military governments of the Western occupying powers and on the assent of the parliaments of the *Länder* to form the *Bund*. The extension of full sovereignty was achieved only gradually: many powers and prerogatives, including those of direct intervention, were retained by the Western powers and devolved to the Federal Republic only as it was able to grow in economic and political stability and to be integrated into the Western community of nations. The tripartite offices of military governor were replaced upon the creation of the Federal Republic by those of high commissioners, and upon its achievement of full sovereignty on May 5, 1955, the high commissioners became ambassadors, accredited to the president of the republic. Since its founding, the West German government, representing the older and more populous of the traditional Germanic territories, has looked upon itself as the legitimate heir to the Wilhelmian (or Bismarckian) *Reich* that united the German states from 1871 until 1945.

Overview of the political and social environments

In the days of Imperial Germany, the society of the nation was among the most intricately hierarchical in all Europe. With certain notable exceptions, the social upheaval of two major wars and economic change, though loosening this rigidity, left the basic class structure intact. West German society is far from plagued by class consciousness, but a sense of one's station in life is implicitly understood. Education still commands a greater awe in Germany than elsewhere; and the professor is held in an esteem incomprehensible to foreigners, and the title "doctor" is an all but essential credential for advancement not only in the upper echelons of the professions and the civil service, where a certain erudition is not inappropriate, but also—even more baffling to outsiders—in the ranks of business. The greatest loss of prestige has been suffered by the military; the German officer, who in the imperial era could snub bank directors with impunity, now commands mere apathy from the public. The older authoritarianism still lurks, but it has been enfeebled beyond repair. As in other advanced nations, the basis of true power has passed to the younger technical and managerial meritocracy.

STRUCTURE OF GOVERNMENT

The Basic Law has many affinities with the Anglo-American democracies and the predecessor Weimar Constitution (upon which it drew heavily). The parliamentary form of government incorporated many features of the British system, but since West Germany, unlike Great Britain, is a federation, many features were incorporated from the models of the United States and other federative governments. Because of Germany's heritage of par-

The constitutional framework

ticularism, however, the individual *Länder* were given and rigorously maintain greater autonomy than, for example, state governments in the United States. The Basic Law also exhibits two features similar to the Constitution of the United States: (1) its formal declaration of the principles of human rights and of bases for the government of men and (2) the strongly independent position of the courts, especially in the right of the Federal Constitutional Court to declare a law unconstitutional and void.

Executive and legislative power. The formal chief of state is the president. Intended to be an elder statesman of stature, he is chosen for a term of five years by an assembly specially convened. His functions are far from being honorific. Apart from representing West Germany among other nations and signing all federal legislation and treaties, he nominates the federal chancellor and the chancellor's cabinet appointments, whom he may dismiss upon the chancellor's recommendation. He cannot, however, dismiss the federal chancellor or the Bundestag, the federal parliament. Among his other important functions are those of appointing federal judges and certain other officials and the right of pardon and reprieve.

The government in power is headed by the chancellor, who is elected by a majority vote of the Bundestag upon nomination by the president; in practice, the chancellor is always the chairman of his party. He is vested with considerable independent powers and initiates government policy. His cabinet and its ministries also enjoy extensive autonomy and powers of initiative. The chancellor can be deposed by an absolute majority of the Bundestag but only after a majority has been assured for the election of his successor. Thus it is unlikely for a chancellor and his government to be unseated, however much his working majority may have dwindled. The cabinet may not be unseated by a vote of no confidence by the Bundestag. The president may not dismiss a government or, in a crisis, call upon a political leader at his discretion to form a new government, the latter constitutional provision being based on the experience of the manner whereby Adolf Hitler—against the better judgment of the *Reichspräsident*—became chancellor in 1933.

Ministries of the cabinet

The number of cabinet ministers may vary. Under the government formed in 1969, there were 14, whereas under the preceding coalition government there had been 19. Most cabinet members are delegates to the Bundestag and are drawn from the majority party or proportionally from the parties forming a coalition, but the chancellor may appoint persons without party affiliation from a certain area of technical competence. These nondelegate members speak or answer questions during parliamentary debates.

The chancellor is immediately assisted by his secretaries of state, who administer various aspects of foreign and internal affairs or the conduct of press and information services; they may exercise wide powers of discretion in carrying out the instructions of the chancellor. The cabinet ministries—apart from the major areas of foreign policy, finance, defense, internal affairs, justice, and commerce—are responsible for such technical and social functions as transport and telecommunications (including the post office, the telephone system, and certain aspects of broadcasting); youth, family, and health; economic cooperation; education; nutrition; labour; and housing and urban planning. A ministry peculiar to the German situation is that devoted to All-German Affairs, which deals with matters pertaining to the Federal Republic's tenuous relations with the German Democratic Republic.

Certain individual organs of government administer such areas as internal security, intelligence, press and information, and statistics; they operate under the direct authority of the chancellor. The Federal Audit Office, independent of both chancellor and Bundestag, is charged with the accounting and budgetary control of all governmental functions.

The Bundestag, the cornerstone of the West German system of government, consists of 518 members (subject to slight variation), including 22 nonvoting delegates from West Berlin. Its delegates are chosen either in general elections held every four years or in special by-elections. In addition to the members elected by each district, a set of *Länder* delegates at large are elected simultaneously both as a means of ensuring stability and continuity of representation by the major parties in the lower chamber and as a corollary to look to the interests of nation, *Land*, party, or bloc of voters at large. In this latter function the delegates at large serve as a counterbalance to the parochial tendencies inherent in strict representation by district constituency mandate.

Methods of constituting the parliamentary chambers

The Bundestag exercises much wider powers than does the upper chamber, known as the Bundesrat, or Federal Council. In the Bundesrat, the *Länder* themselves exercise authority to protect their rights and prerogatives. Its 45 members are appointed by the governments of the *Länder*, each *Land* sending from three to five members, depending on size and population; West Berlin sends four nonvoting members. The delegations are bound by the instructions of their provincial governments. All legislation originates in the Bundestag, and the Bundesrat's consent is necessary only on certain matters directly affecting the interests of the *Länder*, especially in the area of finance and administration, and for legislation in which questions affecting the Basic Law are involved. It may exercise a restraint on the Bundestag by rejecting certain routine legislation passed by the lower chamber, but unless the bills fall within certain categories, its vote may be overridden by a simple majority in the Bundestag. Should the president be absent abroad for long periods or withdraw from office, the speaker of the Bundesrat deputizes for him.

The powers of the Bundestag are kept in careful balance with those of the Landtage, the provincial parliaments. Certain powers are specifically reserved to the republic—foreign affairs, defense, currency and minting, post and telecommunications, customs and problems of international trade, and matters affecting citizenship. The Bundestag and the *Länder* may pass concurrent legislation in such matters when it is necessary and desirable, or the Bundestag may set out certain guidelines for legislation; drawing from these, each individual Landtag will pass appropriate legislation in keeping with the particular needs and circumstances of its own *Land*. In principle, the Bundestag initiates or approves legislation in matters in which uniformity is essential, but the Landtage otherwise are free to act in areas in which not expressly restrained by the Basic Law.

Provincial and municipal government. Certain functions are specifically the province of the *Länder*, notably education and law enforcement; yet even here an attempt is made to maintain a degree of uniformity among the 10 *Länder* and West Berlin through joint consultative bodies. The governments of the *Länder* are generally parallel in structure to that of the Bund but need not be. In eight of the *Länder* the head of government has his own cabinet and ministers; each has its own parliamentary body, but in the city-state *Länder* of Hamburg, Bremen, and Berlin, the mayor is simultaneously the head of government of the *Land*. The municipal senates serve also as provincial parliaments, and the municipal offices assume the nature of provincial ministries.

The administrative subdivisions of the *Länder* (exclusive of the city-states) are known as *Kreise* or, in parts of North Germany, as *Grafschaften*, roughly equivalent to counties. Larger communities enjoy the status of what in Great Britain would be a county borough. The *Kreise* themselves are further subdivided into the *Gemeinden*, roughly "boroughs" or "parishes," which through long German tradition have considerable local autonomy and responsibility in the administration of schools, hospitals, housing and construction, social welfare, public services and utilities, and cultural amenities.

Justice. German law is based on two ancient systems: German law, deriving from the Germanic tribal conventions of antiquity (of which the Salic Law of the Franks, dating from Merovingian times, was the pre-eminent form), and Roman law, the principles of which gained in favour in Germany from the Renaissance to during and after the Napoleonic era. Laws were codified, unlike

Anglo-Saxon common law, in three great civil, criminal, and commercial codes after the founding of the *Reich* in 1871.

Levels and operations of the judicial system

The Federal Constitutional Court, located in Karlsruhe, is the highest court of the land. It enjoys complete independence in its principal tasks of determining the constitutionality of all laws and monitoring the administration of justice and the legality and propriety of administrative and political procedures. The court of highest instance, it admits for deliberation only cases of major import. Immediately below it are 19 high *Land* courts, then the *Land* courts of the first and second instance. The lowest court is the equivalent to a county or district court, which hears cases involving minor offenses or smaller claims. When acting as criminal courts or in cases involving manslaughter, both the *Land* and *Kreis* employ both professional judges and jurors. Certain serious cases such as high treason or conspiracy against the state will be heard in the first instance by the high *Land* court. The Federal Republic has also established a system of labour, social, administrative, and finance courts, appeals from which go to the appropriate federal courts.

While the structures and conventions of Anglo-American and German law, court systems, and procedures are so very different in their provenience and practical workings both systems seem, nevertheless, to arrive at roughly the same conclusions by entirely different routes. By comparison to the general tenor of Anglo-American law, German law often seems to combine a curious mixture of assets—painstaking care, a tendency toward leniency and lighter sentences, and certainly greater consistency and uniformity in the administration of justice—with the faults of a more conservative outlook, especially in the assertion of individual grievances against larger and more powerful organizations and authorities. Claims for damages, substantive or physical, are very difficult to assert. Though German law may fall short of the common law on these counts, the German legal system is far less bound by tradition than the English and American systems and more open to reform and the incorporation of modern procedures.

In the late 1960s a sweeping law reform was begun, some of whose components were the introduction of modern concepts of penology and criminal rehabilitation, the removal of certain matters of moral conduct (such as homosexuality, abortion, and pornography) from the realm of criminal offenses, and the beginning of improvements in the treatment of juvenile offenders. The death penalty is expressly forbidden in the Basic Law.

POLITICAL INSTITUTIONS

The voting age was reduced from 21 to 18 in 1970. Both the quadriennial general and provincial elections as well as local elections are attended with the greatest interest and involvement on the part of the electorate. The electorate is kept informed on political issues through saturation coverage in the press, television, and radio, and political affairs provide a topic for frequent debate among German citizens. The pride with which many Germans once asserted that politics held no interest for them is an attitude of the past. Although voting is not compulsory, an extremely high percentage of citizens participate: in the 1972 general election, 91.1 percent of registered voters went to the polls. Since elections in the *Länder* are staggered throughout the life of each Bundestag, they act as weather vanes of public response to the policies of the incumbent federal government.

Germany's political parties, the sheer proliferation of which contributed to the downfall of the Weimar Republic in 1933, have shown an increasing tendency toward consolidation since the early days of the Federal Republic. Smaller parties have either allied themselves to the larger ones, have shrunk into insignificance, or simply have vanished. Germany in the 1970s has, in effect, only two major parties, though neither can easily attain a parliamentary majority. Most governments, therefore, to date, have held power in coalition with a third party.

The party of the middle and the right—the "bourgeois" party in older European terms—is the Christian Democratic Union (CDU), which headed every Bonn government, often in coalition, from 1949 to 1969, when the Social Democratic Party (SPD) succeeded in forming a government. It functions in all *Länder* except Bayern, where the more conservative Christian Socialist Union (CSU) functions as its counterpart in quasi-permanent coalition.

Political parties and their constituencies

In its origins, the CDU represents a merger of the old Catholic Centre Party with kindred bourgeois parties, either Protestant or nonsectarian. In a nation in which one's religion had also often been one's politics, its strongest constituencies are still in the Catholic districts, although the sectarian Christian aspect is of only incidental emphasis, chiefly among older voters. Its policies emphasize Germany's place in the Western community and NATO, its free-market economy, skepticism if not hostility toward negotiations and a rapprochement with the Communist bloc, and the reunification of Germany on the basis of free elections. Historically the successor to the Marxist parties dating from the 19th century, the SPD under its early postwar leadership maintained a rigorous policy of adherence to classic Marxist doctrine, vehement opposition to the Communist movement from which it had split in the early 20th century, and rejection of rearmament for West Germany and its integration into the Western military defense system. In 1959, however, the SPD discarded is doctrinaire approach: the call for nationalizing large industries was foresaken in preference to gradualist reform, and appeals to class warfare were abandoned. The SPD was able, thereby, to broaden its base to attract greater segments of the middle class, whether these persons were liberal or intellectually Marxist or merely seeking an alternative to the tenure of CDU leadership. Current SPD policies call for social and industrial reform, a rapprochement with the nations of the Eastern bloc (a coming-of-age for the nation in foreign affairs that caused more consternation at home than abroad) while remaining politically and militarily allied with the West, and western European unity through EEC; and it was not hostile to the free-market system.

The third party, minute in comparison but wielding power because it can and will join in coalition with either of the two major parties when no clear parliamentary majority has been obtained, is the Free Democratic Party (FDP). Existing since the 19th century, it was formerly of a conservative, pro-business, and anti-sectarian cast.

Many smaller parties and some older regional parties have disappeared, their members realigned with the larger parties. Both the radical right and left have been reduced to small, dissident groups. Few have been able to get the 5 percent of the votes required by the Basic Law before a party may send delegates to legislative bodies of government. A resurgence of alleged Nazi sentiment under the National Democratic Party (NDP) in the mid-1960s was short-lived, before being fragmented into a shattered collection of elderly unteachables and youthful romantics.

THE ARMED FORCES

A wave of shock spread through war-weary West Germany when in 1950, in view of the likelihood of a long-term Soviet threat to western Europe and the fact that the Soviet Zone of Occupation was rearming, the Western allies reversed the policy—taken at the end of World War II—that Germany should and must never fight again in a European war. A special task force scrupulously strove to create the foundations for an army that would be all but completely purged of the former German military traditions in matters of discipline and of outer appearance—the end effect of which was a positively liberal and democratic cast. To combat the previous blind obedience to orders, members of the Bundeswehr (Federal Army) are indoctrinated in their role as citizen-soldiers and in the range of discretion allowed in conflicts between orders and personal conscience.

Conscription

The Bundeswehr at its inception sought to raise 12 divisions, a goal achieved not without difficulty. Conscription is uniform for males between 18 and 45 years of age; conscientious objectors must perform a term of non-

combatant service in socially useful work, and the sole surviving sons of fathers killed in military service in World War II are exempt. The period of service for conscripts is presently set at 18 months, but plans are underway for a widely based reserve army and a long-term reserve obligation to offset this relatively short period of active duty. Approximately three-quarters of the Bundeswehr's officer corps but less than one-third of its noncommissioned officers are career soldiers. A smaller air force and navy are maintained by the Bundeswehr, along with such other military or quasi-military services as the territorial army, an independent border patrol, and a coast guard.

The German armed forces operate in close coordination with their partners in the North Atlantic Treaty Organization (NATO), of which West Germany became a full member in 1955. The Federal Republic helps to support the presence of some 360,000 Allied troops on its soil, the majority of which (over 200,000) are from the United States. The continued presence of these troops is a frequently reaffirmed objective of German national policy. In 1969 the nation signed the nuclear nonproliferation pact.

THE SOCIAL MILIEU

Education. Schooling is free and compulsory for all Germans from the sixth through the 18th year of age; through age 15 the time must be spent in classroom study. Although each *Land* is absolutely sovereign in matters of education, a permanent conference strives for uniformity in curriculum, requirements, and standards. Primary and secondary education, although presently subject to far-reaching reform, still largely follows the traditional pattern of a common elementary school for the first through fourth years, after completion of which a determination is made as to whether a pupil will terminate his education after a further five to six years of elementary schooling and subsequently be apprenticed to a trade or trained in a special industrial vocation at age 15 or 16; will attend a school essentially geared to clerical, administrative, and commercial careers; or will receive a lengthy academic training in preparation for entry to university or the higher ranks of business and the civil service.

Thus, all children attend the *Grundschule* ("basic school") until about age 10. The majority of children then go on to the *Hauptschule* (roughly, "main school"), usually for five more years, after which they will be assigned to a *Berufsschule* ("vocational school"), usually attended part-time in conjunction with practical apprenticeship. Children to be given training with a practical commercial emphasis attend the *Mittelschule* ("middle school") or *Realschule* (*Real-* implying "practical"). About 14 percent of all children will be chosen for study at the *Gymnasium*, in which a rigorous program lasting from seven to nine years will prepare them—with emphasis variously on the classics, modern languages, mathematics and natural science, or the fine arts—for the *Abitur* degree, the holding of which until recently forthwith entitled a student to matriculation at any German university.

Tradition and change in German universities

The German universities, famed in history and noted for their enormous contributions to learning, especially in the 19th and early 20th centuries, are undergoing profound upheavals. Since their inception in the late Middle Ages and until recently, the traditions of *Lehrfreiheit* and *Lernfreiheit*—the freedom of what to teach as well as the freedom of how a student could go about his studies and preparations for examinations—were sacred. But the expansion of higher learning, as in other Western countries, and the burgeoning number of students and changing social conditions have taxed the traditional structures of the universities beyond their capacities or their accustomed functions. Today it has become all but impossible for a student to take as long as he wished, often eight to 10 years, to complete his studies with the privilege of moving from university to university as he pleased. Lecture rooms, seminars, and libraries are disastrously overburdened, and the higher education explosion and the knowledge explosion in the past generation have undermined the usefulness of original research as well as the solitary authoritarian position of the one professor who stood at the pinnacle of each discipline in each university.

A general reform of higher education long has been under discussion, but little progress has been made. An attempt to limit enrollments—despite the opening of new universities and enlargement of staffs—has been received with violent hostility by the student population. Ironically, the introduction of patterns inherent in the ancient English model and common throughout the English-speaking world runs so contrary to German academic tradition as to have provoked riots. The student unrests of the late 1960s apparently had subsided by the early 1970s, but how the universities are to be reformed remains not only an open but also a volatile question.

By the late 1970s West Germany will have some 40 institutions of university rank, of which over a dozen are or will be institutes of technology or specialized in medicine, economics, or agriculture. Little or no difference in prestige and no social distinction attach to whether a student studies at Heidelberg, founded in 1386, or at Saarbrücken, founded in 1948. The doctorate is the only degree offered as such (although the *Magister*, or Master of Arts, abandoned in the 17th century, has been partially revived); the state examination, roughly equivalent to the master of arts or a good honours degree, is the level at which most students complete their studies.

An extensive range of possibilities exists for extended education or extramural studies. Upward of 1,500 *Volkshochschulen* ("people's universities") enroll almost 2,000,000 adults for complete courses or individual subjects, whether in preparation for or furtherance of a career or out of personal interest.

Welfare and health. West Germany's system of social benefits is one of the most elaborate and all-embracing in the world. The country's pioneer work in social legislation initiated in the 1880s to cover health and accident insurance, workers' and employees' benefits and pensions, miner's insurance, and the like long anticipated and served as a model for similar programs in other countries.

Compulsory national and private insurance plans

Health and retirement insurance are compulsory for all workers and employees earning below a certain level of income geared to the cost of living. Under German labour law, a categorical distinction is made between workers and clerical and managerial employees, and differing rules and rates apply to each group. Employees above a certain salary level or self-employed persons are generally exempt from most obligatory systems; however, although the former will usually participate in a firm's retirement plan, almost all persons in the upper salary brackets or the self-employed will be covered by private insurance as comprehensive as the government-sponsored plans. With increasing prosperity, an increasing percentage of the population is subscribing to these somewhat more expensive but also more generous private plans. Nearly 90 percent of the population is covered by compulsory health insurance, and West Germany ranks highest among comparable nations of continental western Europe in the proportion of money allowed toward health costs—about 90 percent of medical costs incurred. Contributions range from about 8 to 13 percent of wages or salaries.

Medical personnel and facilities

Medical care is of a high standard, and rural areas are well served. Hospitals, no less plagued by runaway costs and staff shortages than in other industrial nations, are usually operated by the municipalities, religious bodies, or as proprietary operations by one or more physicians. There was one physician in 1970 for every 612 inhabitants, of whom very slightly more than one-half were in private practice; of almost 32,000 dentists, nearly 30,000 were in private practice. The nearly 3,600 hospitals hold upward of 680,000 beds. Public health standards are high. Great inroads have been made in the control of tuberculosis, formerly a disease especially endemic to Germany, and free and quasi-obligatory X-rays are offered by local public health authorities. Compulsory medical checks are made on all immigrants, the majority

of whom are drawn from the Mediterranean countries. The standards of public sanitation are among the highest of any nation.

Accident and retirement insurance are tied in with the health- and medical-care plans. The rate of compulsory accident insurance rises with the risks involved in one's job. The three major pension plans cover miners (the oldest, dating from Bismarck's introductory social legislation), the workers, and employees. In general, men are eligible for retirement at 65 and women at 60, but both sexes may be allowed to retire earlier under special circumstances.

In addition, several special systems of coverage are available for such special groups as war widows, orphans, and farmers. Unemployment insurance is provided through deductions from wages and salaries. Allowances are made for families having more than two children, though a grant in some cases may be made for the second child born. Additional public allowances are granted to persons suffering disabilities from wartime injury, whether as military personnel or as civilians. Some small indemnification has been made to property owners whose holdings lay in the territories now under Polish authority, in the Sudetenland (which reverted to Czechoslovakia after 1945), and (by the early 1970s) in East Germany.

Reparations to the Jewish people and Israel

The Federal Republic, consistent with the role it asserted as the one legitimate successor to the defunct *Reich*, has assumed the immense financial responsibility of making restitution to the Jewish and non-Jewish victims of National Socialism. Claims for properties confiscated under the Nazi regime have been honoured, and Jewish refugees and expellees, the vast majority of whom reside abroad, have been paid indemnifications and pensions. Massive reparations have been paid to the State of Israel in the name of the Jewish people at large. (In contrast, East Germany has failed to make any restitution whatsoever to Jewish victims of Nazi persecution and has made no contributions to the State of Israel, with which it holds no ties.)

Housing. In World War II some 20 percent of all dwelling units in what is now the Federal Republic were either destroyed or rendered uninhabitable; almost half of all housing suffered some damage. The immense job of providing replacements had been largely overcome by the mid-1950s, and by 1965 8,800,000 dwelling units had been built to accommodate some 25,000,000 persons. In the early postwar years, the greater amount of funds for housing construction came from government sources, including foreign aid, and relatively little money could be raised in the capital market. By 1965, however, over two-thirds of the funds emanated from private capital, the remainder from funds set aside by the *Bund*.

Assistance is provided by the *Bund* to every citizen wishing to avail himself of a building savings policy. After a certain minimum contribution by the indivdual has been deposited over a determined period of years, the government provides a housing loan on generous terms for those wishing to build their own house or buy a flat.

West Germany still suffers, nonetheless, from a serious housing shortage. Rents and house prices on the private market are high; the great majority of city dwellers live in blocks of flats or multiple housing units, and until recently private rental property was subject to relatively little control over rents or arbitrary imposition of conditions of rental by the landlords.

Law enforcement. No nationwide police force exists, and law enforcement remains a function reserved to the *Länder*. Each *Land* maintains its own force, the *Landespolizei*, which is charged with all phases of enforcement throughout each *Land* except where their function is assumed by a municipal police force. In a state of national emergency the *Bund* may commandeer the services of the various Landespolizei units, together with the standby police reserve that is trained and equipped by each *Land* for action during civil emergencies. Federal offices investigate certain actions, however, notably those inimical to the security of the state or criminal actions that transcend the confines of the *Länder*, and one assists the provincial and municipal units as a clearing agency, the Bundeskriminalamt (Federal Crime Agency), on criminals and criminal actions.

Standards of living. Despite recent sharp rises in the cost of living, the German wage earner still enjoys a relatively high and stable standard of living. As in most countries of EEC, food prices are high in relation to income. The average German consumer, if not the large exporter or farmer, benefitted to a small extent when the value of his mark increased during and after 1969—by some 10 percent in relation to the currencies of neighbouring countries and by almost 25 percent in relation to the United States dollar.

The distribution of wealth in Germany compares favourably with that of other advanced nations. Wages in the lower paid occupations are adequate if not high, while salaries in the medium range are only moderately higher. Thus, no exaggerated discrepancy exists between the rewards of blue-collar versus white-collar workers. The salary range may rise precipitously, however, above a certain level of management, and since, as in other EEC countries, a major portion of tax revenue derives from excise levies and valued-added tax, the low- and medium-income earners bear the greatest burden.

VI. Cultural life and institutions

The Federal Republic of Germany, as heir to the older regions of the *Reich*, is custodian to the greater portion of its rich cultural legacy. The major wealth of Germany's architectural monuments—of Roman Germany, of medieval Romanesque, of south German Baroque—fell within the borders of West Germany, as did many of the great libraries and archives, art treasures, and facilities for the performing arts. Much of incalculable value in the tangible heritage of the past lies in the trusteeship of the German Democratic Republic, which contains the Wartburg of Luther, the Weimar of Goethe, the Leipzig of Bach; a large share of prewar Berlin's art treasures now rests in the Eastern Sector, notably the Pergamon Museum and its archaeological holdings; and the Baroque monuments of Dresden savagely destroyed in World War II have undergone restoration. It would be untrue and unfair to assert that East Germany has been negligent in sustaining, within its resources, its share of the older Germany's cultural monuments and institutions. The archives, buildings, and museums pertaining to Germany's golden age of literature have been carefully restored at Weimar; the choir of St. Thomas' Church at Leipzig still continues in unbroken continuity since the days of Bach and before; Leipzig's Gewandhaus Orchestra, Dresden's Kreuzchor, and the German State Opera in East Berlin all maintain the high standards of the past.

But after the division of Germany, many of the cultural assets originally in the East were removed to the West. Many of its artists and writers and also entire institutions, such as a number of illustrious publishing houses, transplanted themselves or set up successor organizations in the Federal Republic. Although the older cultural traditions have been well served in East Germany, it has become virtually sterile ground for any great contributions to the arts—with the notable exception of Berliner Ensemble, the theatrical company founded and once directed by Germany's pre-eminent modern playwright, Bertolt Brecht, and his wife, Helene Weigel. In the Federal Republic, the German cultural tradition has been unimpeded by the strictures of political doctrine and flourishes with vigour and in abundance.

THE CULTURAL MILIEU

Governmental and audience support. For four centuries, Germany has enjoyed a tradition of governmental support of the arts. Before the founding of the *Reich*, the many small kingdoms, principalities, duchies, bishoprics, and free cities that preceded modern Germany—as well as Austria and German-speaking Switzerland—supported the arts; established theatres, museums, and libraries; and acted as patrons to poets, writers, painters, and performers. The institutions thus founded and the convention of generous public support has continued uninterrupted.

The quantitative dimensions of Germany's cultural life

astound foreigners. In the Federal Republic alone some 195 theatres are subsidized, about 85 by the Bund and the *Länder* and about 110 by the cities, in addition to upwards of 70 privately financed theatres. Within this system of theatres are over 50 opera companies. Unlike the United States, Britain, and France, in which theatre is more often than not centred in one city, no one city in Germany dominates over the others. Productions in Vienna or Zürich are significant to the artistic life of the Federal Republic, and artists and resources move easily and freely among the theatrical and operatic companies within the German-speaking regions. Only in Vienna, in which capital the arts arouse passions far more intense than those of politics, does theatre have a broader audience base than in Germany. Audiences in Germany are not limited to a small intellectual or social elite but are drawn from all ranks of society. Season tickets, group arrangements, bloc tickets bought by companies, and theatre clubs constitute the major part of the regular patronage of such production companies as the Theater der Freien Volksbühne (Free People's Theatre), dating from 1890 in Berlin. Going to the theatre or opera in Germany is about as inexpensive and as commonly done as attending the films would be elsewhere.

Theatrical and musical traditions

The same is true of concert music. Every major city has one or more symphony and chamber orchestras offering many concerts and recitals each week; the music and concert fare is less well provided for in the smaller cities and towns only in terms of quantity and, perhaps, professional quality. Musical festivals are a prominent feature in the calendar of the performing arts in Germany. The most renowned, the annual Wagner Festival held each summer in Bayreuth, is still under the direction of Wagner's descendants. Others of note are the autumn Berlin Festival, the May Festival in Wiesbaden, the Beethoven Festival in Bonn, the Mozart Festival in Würzburg, and the Festival of Contemporary Musical Art in Donaueschingen.

Music festivals

The greatest annual event in German music, the Salzburg Festival, is held across the Bavarian border in Austria. But in German, the adjective *deutsch* ("German") implies no strict nationality when referring to the arts, whether German, East or West, Austrian, or Swiss. Unlike the usage of the English-speaking world, the German-speaking nations will refer to Friedrich Dürrenmatt, a playwright of Swiss nationality, as a German writer or to the Austrian Mozart as a German composer, for they are all in the German cultural tradition.

Museums, libraries, and publishing. West Germany has more than 500 museums of all descriptions, ranging from some of the world's great collections of paintings and sculpture or of archaeological and scientific displays to exhibitions of minutiae such as those in the playing-card museum in Stuttgart. The western portions of Germany contain the greater amount of the nation's prewar art treasures. In addition to West Berlin's museum, museums and art galleries of great note are the Alte Pinakothek in Munich, with its unrivalled Rubens collection, the German Museum in Munich, the Roman-Germanic Museum in Mainz, the Senckenberg Museum of Natural Science in Frankfurt am Main, and the Alexander König Museum of Zoology in Bonn. Important individual art treasures are scattered in the scores of smaller museums, libraries and archives, cathedrals, churches and monasteries, and castles throughout the country. Special exhibitions are frequently held, one of the most outstanding of which was the 500th-anniversary celebration in Nürnberg in 1971 of Albrecht Dürer, which displayed almost every accessible work of the great Renaissance master.

Among the great libraries in Germany are the Bavarian State Library in Munich, the nation's largest, and portions of the former Prussian State Library that were removed to Marburg and are known as the West German Library. The German Library at Frankfurt am Main is the country's bibliographical centre; the Technical Library at Hannover is the most important for science, technology, and translations of works in the sciences and the engineering field. Apart from the great university libraries at Heidelberg, Cologne, Göttingen, Tübingen, and Munich, a wealth of ancient manuscripts, early printed works, manuscripts, and documents are in diverse collections. The great research libraries are complemented by an extensive system of lending libraries operated by the *Länder*, the municipalities, the library associations of the Roman Catholic and Evangelical Churches, and other public associations and institutes; virtually no citizen is out of easy reach of a free lending library.

Germany is a major publishing nation, with nearly 2,500 publishing houses. Few of these are on the scale of the giant houses known in British and American publishing, and most quality titles are published by prestige houses of small to moderate size. Academic and educational publishing is undertaken by scores of smaller houses. Over 900 publishing enterprises serve also as retail booksellers. In 1971, over 47,000 books were published in the Federal Republic, almost 39,000 of which were new titles.

CURRENT STATE OF ARTISTIC PRODUCTION

In no country in the world are the arts so lavishly cultivated as in the Federal Republic in terms of the proliferation of cultural amenities, the funds allotted to them, and the attendance upon them. Notwithstanding this abundance and generous support, relatively few major talents of international renown have emerged in the major fields of the fine arts in the postwar years, especially in comparison to the era of the Weimar Republic, when Germany (especially Berlin) experienced a resurgence in the arts and a proliferation of innovative talents unparalleled since the ages of German classicism and romanticism in the late 18th and early 19th centuries.

Sparsity of contemporary artistic contributions

Literature and theatre. German literature holds less than its deserved status in world literature in part because the lyrical qualities of its poetry and the nuances of its prose are ill served by translation. Even the most sublime of figures in German literary history such as Goethe and Schiller are doomed to remain known to the world largely by reputation. In the 20th century perhaps four German poets and writers have won a permanent niche in world literature—Franz Kafka, Thomas Mann, Rainer Maria Rilke, and Bertolt Brecht, all of whose works date from the early decades. Of German novelists since the war, only two have been published widely in translation abroad—Heinrich Böll and Günter Grass; among the playwrights (apart from Brecht and Carl Zuckmayer, whose major plays extend from the 1920s to the mid-1950s), the works of Siegfried Lenz and Peter Weiss have been well received abroad.

The German theatre today is thus presented with the dilemma of either relying on the rich repertory of German classics from the 18th and 19th centuries, along with a restricted number of truly great modern dramatic works, or of producing plays, in not always the most successful of translations, of leading British, American, and French authors. Small experimental theatres enjoy a lively and hazardous existence in the major cities and supplement and often closely resemble Germany's long-lived and still lively convention of political cabaret.

A major innovation to the German stage has been its adoption of the American musical. Ever since a game and gingerly attempt was made at Cole Porter's *Kiss Me Kate* (*Küss mich Kätchen*) in Berlin in 1955, this most indigenously Broadway of genres is now widely enjoyed throughout Germany; four separate productions of *Hair* (*Haar*) played to fully booked audiences concurrently from 1969 until 1971.

Music and dance. The performance of both operatic and concert music in the Federal Republic, as elsewhere, is faced with the dilemma of the public's preference for 18th- and 19th-century composers as opposed to the producers' and artists' wish to give contemporary works a better hearing. The works of Hans Werner Henze, whose opera *The Young Lord* is widely performed abroad, the symphonic works of Gottfried von Einem, and the avant-garde electronic music of Karlheinz Stockhausen have established these men as the most widely known of contemporary German composers. The Berlin Philharmonic, the Bamberg Symphonic Orchestra, and the Stuttgart

Chamber Orchestra continue to rank with the leading world ensembles. Innumerable German operatic voices have gained world renown, but an interesting phenomenon in recent decades has been the many North American and British voices who, lacking suitable outlets for their careers at home because of the few opera companies there, have trained in Germany and Austria and stayed on to perform on their stages. A refreshing trend in contemporary opera is the fading of the old insistence that the librettos be sung in German translation.

A conspicuous international success in the performing arts has been that of the Stuttgart Ballet under the direction of the South African-born John Cranko. German ballet, so often overshadowed by the Russian and English conventions, has been thrust into pre-eminence by the Württemberg State Theatre's company, which performs frequently in the major cities of the West and also has won acclaim by Russian audiences—the world's most critical—on its visits to the Soviet Union.

The visual arts. Contemporary painting has moved from under the dominance of the Expressionist mode dating back to the Dresden school known as Die Brücke (The Bridge) of the early 20th century and such successor movements as Die neue Sachlichkeit (Neo-Objectivity) of the 1920s, in which an emotive or subjective suggestion of reality was achieved by contorting or exaggerating natural forms. The greater number of younger painters have turned to the problems of abstraction and the eclectic experiments preoccupying artists in other Western countries; many concern themselves principally with graphic design or drawings. In sculpture, a similar nonparochially German trend is apparent in the drift from the conservatism of the older national sculpture to experiments in abstraction and the use of unconventional materials.

Increasing internationalism of painting and architecture

German architecture today—indeed world architecture today—is still very much the creature of the Bauhaus school, originating in Dessau in the 1920s. The basic ideals of its leading exponents, such as Walter Gropius and Mies van der Rohe, a spartan harmonizing of function with design, still prevails. This has been profoundly mitigated, however, by a less puritan decorativeness, by the social dictates of harmonizing the new with the old in rebuilding the cities and restoring old monuments, and by the exigencies of soaring building costs. Contemporary architecture—especially as evident in major commercial and public projects or housing estates on the outskirts of the larger cities, which are freed from an obligatory conformity with the existing architecture of the inner city—reflect a marked cosmopolitanism no longer identifiably German or even northern European in cast.

Traditional arts and crafts. The incursions of modern patterns of life have done much to weaken the traditional arts, entertainments, and customs of regional and rural Germany. In southern Germany the older arts and usages have persisted concurrently with a gradual adaptation to a modern, urban pattern of life; the old and the new coexist in an incongruous compatibility. The young still dance around the village maypole, but they also dance to rock. The woodcarvers, violin makers, and gunsmiths of Upper Bavaria continue, with increasing economic hardship, to follow their trades, not because it is quaint but because they still believe in them; the rural women in the Black Forest still wear elaborate costumes known as *Tracht* on festival days, not to amaze tourists but because they have always done so—yet these are the areas in which the tourist industry is most highly developed. Some usages have all but disappeared in the villages: older women now seldom wear black dresses and scarves, and the village men no longer appear in top hat and cutaway for a funeral procession.

Popular festivals still abound. Near-heathen usages such as the elaborate wooden masks that surround the pre-Lenten celebrations in southwestern Germany remain unaffected in spite of being televised; hundreds of smaller towns and larger villages in the south still commemorate an anniversary from the Thirty Years' War by a parade in 17th-century costume or, in Catholic areas, march in full procession on Corpus Christi Day. What is remarkable is not merely that these usages survive but also that the homelier and less celebrated of them remain truly genuine and naïve in the observance.

Popular culture. In contrast to the situation after World War I, when Germany helped set the pace in many of the popular forms of art and entertainment, most conspicuously in developing the film as a genre in its own right, modern West Germany has never been able to regain a métier of its own—with the exception of the political satires presented in scores of cabarets. In the cinema, the important pioneer films of Georg Wilhelm Pabst, Fritz Lang, and others were succeeded in the postwar years by farces and romances of so meagre a substance as to defy explanation in a nation so given to the performing arts. Television entertainment is of mixed quality, relieved in part by the occasional original play written for the medium or adapted from the stage. Many foreign imports are carried, mostly American and British television series and older films, the sound tracks dubbed into German.

Films and television

Styles in popular music have found no indigenously German vehicle of expression—nothing to compare, for example, with the world-weary and vaguely wicked Berlin chanson of the 1920s. Only slavish (if technically competent) copies of contemporary styles from abroad are heard along with the never-changing fund of conventional dance music. The tango still lives; no revival of the so-called big band sound of the 1930s was required in West Germany, since it never had waned.

CULTURAL INSTITUTIONS

Great importance has been attached in the Federal Republic to the support of the nation's cultural, educational, and scientific resources through institutions supported in whole or part by public funds. A prodigious complex of organizations are devoted to acquainting the public with the culture, life, and language of the German peoples and of making the culture and life of other nations more familiar. Cultural representation abroad and cultural exchanges are maintained in abundance with the advanced industrial nations of the West and, increasingly, with eastern Europe, but special emphasis is laid upon the development of cultural ties with the developing nations of the world. West Germany not only has assumed a major role in lending reserves of technological skill and capital to developing the resources of the lesser or non-industrialized nations; it also has become a major centre for the education and training of students from these countries in the professions, the sciences, and technology.

International cultural activities

Prominent among these groups is Inter Nationes, a nationwide organization devoted to the higher levels of cultural representation abroad through the media of books and films and to fostering a better understanding of the arts and education in the Federal Republic by sponsoring the visits of leading persons in these fields from abroad. A similar task is performed by the Institute for Foreign Relations and several bodies devoted to academic exchanges.

The massive task of making the German language and literature accessible to peoples abroad and of introducing them to German life and civilization falls in largest measure to the Goethe Institute in Munich. Through its 21 schools within the Federal Republic and more than 120 branches abroad, it sponsors intensified courses in German, with a practical emphasis on its use in daily life, and operates libraries and German cultural centres in the host countries. Its work is complemented by more than 70 German cultural institutes of various sponsorship abroad, prominent among them the jointly German–American-subsidized Goethe House in New York City.

Learning, the arts, the social sciences, and the natural sciences and technology are fostered by a proliferation of foundations, academies, and institutes. A series of learned journals published in Tübingen records the progress made in German research in all branches of learning. A great variety of major social projects and research, both in the Federal Republic and abroad, is sponsored by foundations established by the larger commercial and industrial corporations.

THE COMMUNICATIONS MEDIA

Newspapers and magazines. West Germany is very much a land of newspaper and periodical readers. Freedom of the press is guaranteed under law, and the economic state of its some 430 daily and 60 weekly newspapers and more than 870 magazines and journals, by comparison to the situation in similar nations, is enviably healthy. Most major cities support two or more major newspapers in addition to community periodicals, and few towns of any size are without their own daily. Although rising costs are endangering this abundance and consolidations and takeovers are narrowing the basis of ownership, few nations are so prodigiously served by the press. A national press similar to that in England has developed through such newspapers not identified with any one city as *Die Welt* and the weeklies *Die Zeit* and *Deutsche Zeitung*, in addition to certain municipal newspapers read throughout the nation—some of ranking international stature—such as the *Frankfurter Allgemeine Zeitung*, the *Süddeutsche Zeitung* of Munich, the *Stuttgarter Zeitung*, and Koblenz's *Rheinischer Merkur*. An important and enormously influential segment of the press is under the proprietorship of Axel Springer Verlag. Its empire includes five newspapers complemented by a wide spectrum of popular magazines. The Springer newspapers account for over one-fourth of all newspaper sales in the Federal Republic.

The genre of the *Illustrierte* dominates the German magazine market. Some few of these popular glossies, on a par with *Paris Match*—*Stern* foremost among them —feature investigative reporting of a high calibre; most, however, cater to the public's appetite for the escapades of celebrities, especially members of royal families and entertainers, and for bizarre crime, gracious living, and other trivia. Apart from a wealth of specialized journals and quality business-oriented magazines, the Federal Republic lacks prestigious magazines of opinion, a function served in some measure by the weighty weekend editions of the quality press.

A solitary power in political journalism, the weekly news magazine *Der Spiegel* has performed a unique function in the shaping of West German public opinion through its posture of the skeptical, nonaligned observer and guardian of the public conscience. Exhaustive in its coverage and invariably polemic in tone, it features in-depth muckraking investigations of the distant past and of contemporary affairs.

Broadcasting. Radio and television operate as public corporations under the authority of the Federal Ministry of Post and Telecommunications, which tenders broadcast licenses, assigns frequencies, and collects public revenues from fees levied on owners of radio and television sets. The individual corporations are otherwise free to establish their own broadcasting policies. Attempts to control these policies, which often are hostile to incumbent governments, have been rebuffed repeatedly, so that German television, in practice more than radio, is able to assert remarkable latitude and independence in the tenor and content of its broadcasts.

Radio is arranged along regional lines, with some 10 corporations engaged in local broadcasting. Each may operate two or more separate programs on various frequencies. Each corporation is independent of the other, but programs and facilities are shared. In addition, the Deutsche Welle broadcasts around the world in 30 languages. The *Deutschlandfunk* is a national program beamed at all Germany and Europe.

Two television networks, operating through two separate channels, transmit nationwide programs incorporating the full range of news, documentaries, and entertainment through the facilities of the regional broadcasting corporations, each of which appends its own television coverage and has full responsibility for programming on a third channel intended to function region by region as an educational and fine-arts medium. The uneven quality of the entertainment fare in both radio and television is offset by high-quality news coverage and political and social reporting.

VII. Prospects

Over the centuries, Germany has been a name in search of a nation, a nation in search of a state, a state in search of a role. Even Bismarck's creation of the *Reich* in 1871 only lasted a generous lifespan of 74 years of the two millennia of recorded German history, and it was accomplished only by excluding Austria, the once-dominant state among the lands once known as Germany. The German peoples, who in recent generations have had to come to grips more than once with what had been the unthinkable, may never succeed in realizing a precisely defined, geographically tidy concept of Germany. But if by analogy to the Anglo-Saxon nations and the Hispanic lands there may be several German states—not only the German Democratic Republic but also Austria, much of Switzerland, even Luxembourg and Liechtenstein—the Federal Republic of Germany, as the mainspring of the German peoples' language and culture, as heir to the oldest portions of the country, and as custodian of the title, seems best entitled to bear the ancient name.

Reunification, a hope that had already dimmed by the mid-1950s, later appeared impossible. The Socialist government's 1969 submission that there were two German states within one German nation appeared unlikely to satisfy either portion of a country that had been fully divided since 1949. The Federal Republic is firmly integrated into the Western political system, the German Democratic Republic into the Eastern. Their economic, political, and legal structures appear to have grown too far apart to be mutually compatible; notwithstanding the strongest ties of kinship and the bitter personal hardships of separation, the two societies have moved so far apart with the passage of time that the estrangement scarcely can be reversed.

From its beginnings, the Federal Republic has implicitly proceeded on these assumptions of a harsh political reality. Whatever its people may hold of the past, they cherish no fond hopes that it will be recouped, assuredly not in terms presently imaginable. Instead, they have devoted the considerable human resources of the nation into building not merely what many believe to be one of the sturdiest of the world's democracies but also one of the mightiest of the world's national economies. In so doing, the Federal Republic has seemingly laid a firmer and more lasting basis for a rightful German role in the world than its predecessor state could have sustained by military might. In 1830 the Prussian military theorist Karl von Clausewitz wrote that war is a continuation of diplomacy by other means. In its reaffirmed goals of working toward a united Europe, of using its offices to strengthen the cause of peace, and of lending its formidable assets to relieve imbalances in the potentials of the weaker nations, the Federal Republic of Germany, many believe, may prove itself an example to the world—a paradoxical turn of von Clausewitz's dictum of how to extend a nation's policy through truly "other means."

BIBLIOGRAPHY. The best general geography of Germany in English is THOMAS H. ELKINS, *Germany* (1968), while MICHAEL WINSCH, *Introducing Germany*, new ed. (1967); MARGARET WIGHTMAN, *The Faces of Germany* (1971); and the *Nagel Travel Guide to Germany*, 3rd ed. (1968), are readable and informative for their detailed descriptions region by region, city by city, and for the treatments of such ancillary topics as German history, geography, cultural life and folklore, the news media, and the economy. The definitive source for figures and details on population, infrastructure, and the economy is the annual edition of the *Statistisches Jahrbuch* published by the Federal Republic's *Statistiches Bundesamt* in Wiesbaden; details of all major public institutions are reflected in ALBERT OECKL (ed.), *Taschenbuch des öffentlichen Lebens* (annual). West Germany's standard reference book is *Der Grosse Brockhaus*, a revised edition of which has been appearing serially since 1966.

Of the prolific body of works on the Federal Republic, an abundance of titles dating from the late 1960s serves well to complement or supplant treatments available in English that rapid changes have in large measure superseded. For informative general accounts of the country, the people, the regions, and the historical, political, economic, and cultural background, see J.P. PAYNE (ed.), *Germany Today: Introductory Studies* (1971); and, with scholarly emphasis on the current

state of institutions and the arts, MALCOLM PALSEY (ed.), *Germany: A Companion to German Studies* (1972), intended as a successor volume to JETHRO BITHELL'S classic of the same title (1932), never revised but ever valuable for its historical perspectives. RUDOLF LEONHARDT'S perennially successful collection of observations and speculations about his country, *X-Mal Deutschland*, 11th ed. (1971), is available in English as *This Germany* (1964). TERENCE PRITTIE, long a senior foreign correspondent in Bonn, has contributed the volume *Germany* (1965) in the "Time-Life Series."

An intimate view of modern Germany and its people as recalled from the author's own career, together with insights into the historic and philosophic background seldom emphasized by foreign commentators, may be found in GEORGE BAILEY, *Germans: The Biography of an Obsession*, new ed. (1972). A further authoritative summation of the history and institutions of the Federal Republic since its founding, with comparisons to developments in East Germany, is offered in ALFRED GROSSER, *L'Allemagne de notre temps, 1945–1970* (1970; Eng. trans., *Germany in Our Time*, 1971). The U.S. journalist ADOLPH SCHALK gives a comprehensive view of contemporary German life in *The Germans* (1971).

A trenchant critique of German democracy, marred by its neglect of accomplished political fact, is notable for the reputation of its author, the philosopher-theologian KARL JASPERS, in *Wohin treibt die Bundesrepublik*, pt. 3 (1966; Eng. trans., *The Future of Germany*, 1967); RALF DAHRENDORF, the German sociologist and statesman, has contributed an account of especial interest to foreigners in *Gesellschaft und Demokratie in Deutschland* (1968; Eng. trans., *Society and Democracy in Germany*, 1969). Other works with particular attention to West German government and politics are R.B. TILFORD and R.J. PREECE, *Federal Germany: Political and Social Order* (1969); OTTO SIEGNER, *Germany* (1970); GWENDOLEN M. CARTER and JOHN H. HERZ, *The Government of Germany*, 5th ed. (1967); GERHARD LOEWENBERG, *Parliament in the German Political System* (1967); JOHN C. LANE and JAMES K. POLLACK (comps.), *Source Materials on the Government and Politics of Germany* (1964); ELMER PLISCHKE, *Contemporary Governments of Germany* (1969); and KARL KAISER and ROGER MORGAN (eds.), *Britain and West Germany* (1971).

German popular culture and folkways are discussed in MARION ADAMS, *The German Tradition* (1971); and LAVERN RIPPLEY, *Of German Ways* (1970). One major aspect of German popular arts is treated in ROGER MANVELL and HEINRICH FRAENKEL, *The German Cinema* (1971).

(G.H.K.)

Germany, History of

This article traces the history of Germany from the breakup of the Carolingian Empire in the mid-9th century to the present. For the early history of the Germanic peoples, see GERMANS, ANCIENT; for the history of Germany from the 6th to the 9th centuries, see MEROVINGIAN AND CAROLINGIAN AGE.

The article is divided into the following sections:

I. Germany until *c.* 1250
  The disintegration of the Carolingian Empire
    The kingdom of Louis the German
    Rise of the duchies
  The 10th and 11th centuries
    Conrad I (911–918)
    The accession of the Saxons
    The eastern policy of the Saxons
    Dukes, counts, and advocates
    The promotion of the German Church
    The Ottonian conquest of Italy and the imperial crown
    The Salians, the papacy, and the princes, 1024–1125
  Germany and the Hohenstaufen, 1125–1250
    Dynastic competition, 1125–52
    Colonization of the east
    Hohenstaufen policy in Italy
    The fall of Henry the Lion and the estate of princes
    The Hohenstaufen conflict with the papacy, 1159–1215
    Frederick II and the princes
    The *Reich* after the Hohenstaufen catastrophe
II. Germany from 1250 to 1493
  1250 to 1378
    The extinction of the Hohenstaufen dynasty and the rise of the Habsburgs and Luxembourgs
    The Great Interregnum
    The rise of the Habsburgs and Luxembourgs
    The growth of territorialism under the princes
    Constitutional conflicts in the 14th century
    The continued ascendancy of the princes
  1378 to 1493
    Internal strife among cities and princes
    The Hussite controversy
    The Habsburgs and the imperial office
    Developments in the individual states to *c.* 1500
    German society, economy, and culture in the 14th and 15th centuries
III. Germany from 1493 to *c.* 1760
  The Reformation, to 1555
    Maximilian I
    Beginning of the Reformation
    The election of Charles V
    The Diet and Edict of Worms (1521)
    Lutheran church organization and the Peasants' Revolt (1524–25)
    The Diets of Speyer (1526 and 1529) and Augsburg (1530)
    Protestant and imperial politics after Augsburg
    The Schmalkaldic War
    The Augsburg Interim
    Maurice of Saxony's war (1552)
    Charles V's abdication and the Peace of Augsburg (1555)
  The Counter-Reformation and the Thirty Years' War
    Ferdinand I and Maximilian II
    Rudolf II, the Cologne War, and Matthias
    Ferdinand II and Bohemia
    The Thirty Years' War
    The Peace of Westphalia (1648)
  The empire in decay, 1648–1721
    Emperor and empire after 1648
    The territorial powers in the 17th century
    The War of the Spanish Succession (1701–14)
    The Great Northern War (1700–21)
  Austro-Prussian rivalry in the 18th century
    Charles VI and the Pragmatic Sanction
    The War of the Polish Succession (1733–35)
    The War of the Austrian Succession (1740–48)
    The Seven Years' War (1756–63)
IV. Circa 1760–1871
  Germany from *c.* 1760 to 1815
    Further rise of Prussia and the Hohenzollerns
    The cultural scene
    Enlightened reform and benevolent despotism
    The French Revolutionary and Napoleonic era
    Results of the Congress of Vienna
  The age of Metternich and the era of unification: 1815–71
    Reform and reaction
    Evolution of parties and ideologies
    Economic changes and the Zollverein
    The revolutions of 1848–49
    The 1850s: years of political reaction and economic growth
    The 1860s and the triumphs of Bismarck
V. Germany from 1871
  The Second Reich, 1871–1918
    Bismarck as imperial chancellor, 1871–90
    Chancellors Caprivi and Hohenlohe, 1890–1900
    Chancellorship of Bülow, 1900–09
    The prewar years, 1909–14
    World War I
    Defeat of revolutionaries, 1918–19
  The German Republic, 1919–33
    The Weimar constitution
    The Treaty of Versailles
    Years of crisis, 1920–23
    Attempts to stabilize the republic, 1923–30
    The end of the republic
  The Third Reich, 1933–45
    The Nazi revolution
    The totalitarian police state
    Foreign expansion and defeat
  Germany after World War II, 1945–49
  The Federal Republic of Germany
    Chancellorship of Adenauer
    Chancellorship of Erhard, 1963–66
    Chancellorship of Kiesinger, 1966–69
    Chancellorship of Brandt, 1969
  The German Democratic Republic

I. Germany until c. 1250

THE DISINTEGRATION OF THE CAROLINGIAN EMPIRE

The kingdom of Louis the German. Louis I the Pious, Charlemagne's successor as emperor, was not unpopular with his German subjects; on two occasions he owed his restoration to power largely to their support. In 825 one of his sons, Louis the German, was entrusted with the government of Bavaria, whence he was gradually to extend his power over all Carolingian Germany. This was

the first time that the German nations had had a ruler whose authority was confined to their own lands and whose time was largely taken up with defending them from Slav penetration; but this was by no means the sum of ambition for Louis the German, who tended, like all his house, to regard the whole of Franconia as a divisible family inheritance of which each member in each generation should take for himself and his followers what he could get. Louis was thus satisfied neither with the partition treaty of Verdun (843), by which he obtained the bulk of the lands east of the Rhine together with the districts around Mainz, Worms, and Speyer on the left bank, nor with that of Mersen (870), by which he and his half brother the West Frankish king Charles the Bald came to terms over Lotharingia, the middle kingdom of their nephew Lothair that generally comprised the Low Countries, northwest Germany, and northeast France. Under the Treaty of Mersen Louis's dominions reached almost the proportions of medieval Germany. On the east they were bounded by the Elbe and the Bohemian mountains; on the west, beyond the Rhine, they included the districts afterward known as Alsace and Lorraine. Ecclesiastically, they included the provinces of Mainz, Trier, Cologne, Salzburg, and Bremen. But Charlemagne's capital at Aachen and the rich family estates in Lotharingia were never finally abandoned by either branch of the Carolingian dynasty, although the bulk of the lands that they controlled increasingly assumed the separate outlooks of France and of Germany. An example of this increasing separation is provided by the oaths sworn at Strasbourg in 842 by Louis and Charles, the former swearing in his brother's language, Romance, the latter, conversely, in German; but this drifting apart is in some ways less significant than the ties that still held France and Germany together.

Partition treaty of Verdun

The ceaseless external blows from Danes, Saracens, and Magyars that fell upon the Carolingian world in the 9th century did not have the effect of uniting it in resistance. Not only the Carolingians themselves but their followers also were prepared to take advantage of each other, to compromise with the enemy and to carve out even more dominions from each other's lands. The motives that led them to behave that way, however, were not so simple as they may now at first appear. For instance, in 887, Arnulf, an illegitimate son of Louis the German's son Carloman, led an army of Bavarians against Charles the Fat, in part because Charles was not defending the Rhineland from the ravages of the Danes, in part because his aim was the full Carolingian inheritance. But Arnulf was not equally successful in defending his eastern possessions. After his death in 899, the German kingdom came under the nominal rule of his young son, Louis the Child, and in the absence of strong military leadership became the prey of the Magyar horsemen and other invaders from the east.

Barbarian invasions

Rise of the duchies. The rise of the German duchies was a direct outcome of the Carolingian decision, avoidable or not, to leave defense in the hands of those who were attacked; in other words, to decentralize military command and with it, inevitably, something else of the royal authority. The new *duces* were neither, as was once thought, appointed by the peoples concerned, nor were they the descendants of the tribal chieftains of the post-migration period. It seems more likely that they were Carolingian counts who took the initiative in organizing defense on a local basis, without thereby seeking to shake men's loyalty to the house of Charlemagne, of which the German church was a natural champion. All the same, their initial success established them in the hearts of those whom they protected. This was particularly true of the Saxons, whose dukes, the Liudolfings, were descended from the military commanders first sent by Louis the German to defend eastern Saxony. Similarly, the Swabian dukes began with a military title (*duces Raetianorum*), as did the Bavarian ducal family of Liutpold. The origins of the short-lived duchy of Thuringia are less easy to determine. Franconia naturally remained the German duchy most intimately associated with the East Frankish kingship. How, in practice, the lives of the German land-owning or land-renting freemen were affected by these changes is a matter largely of guesswork. Perhaps it is true that political insecurity and its economic consequences tightened the lord–vassal relationship, as in France and elsewhere (see MIDDLE AGES). Nothing is more striking, however, than the region-to-region variety of German social organization. Perhaps much more of tribal ways of local government survived the Carolingians and the Magyars than had once been thought possible.

Development of lord–vassal relationship

(Ed.)

THE 10TH AND 11TH CENTURIES

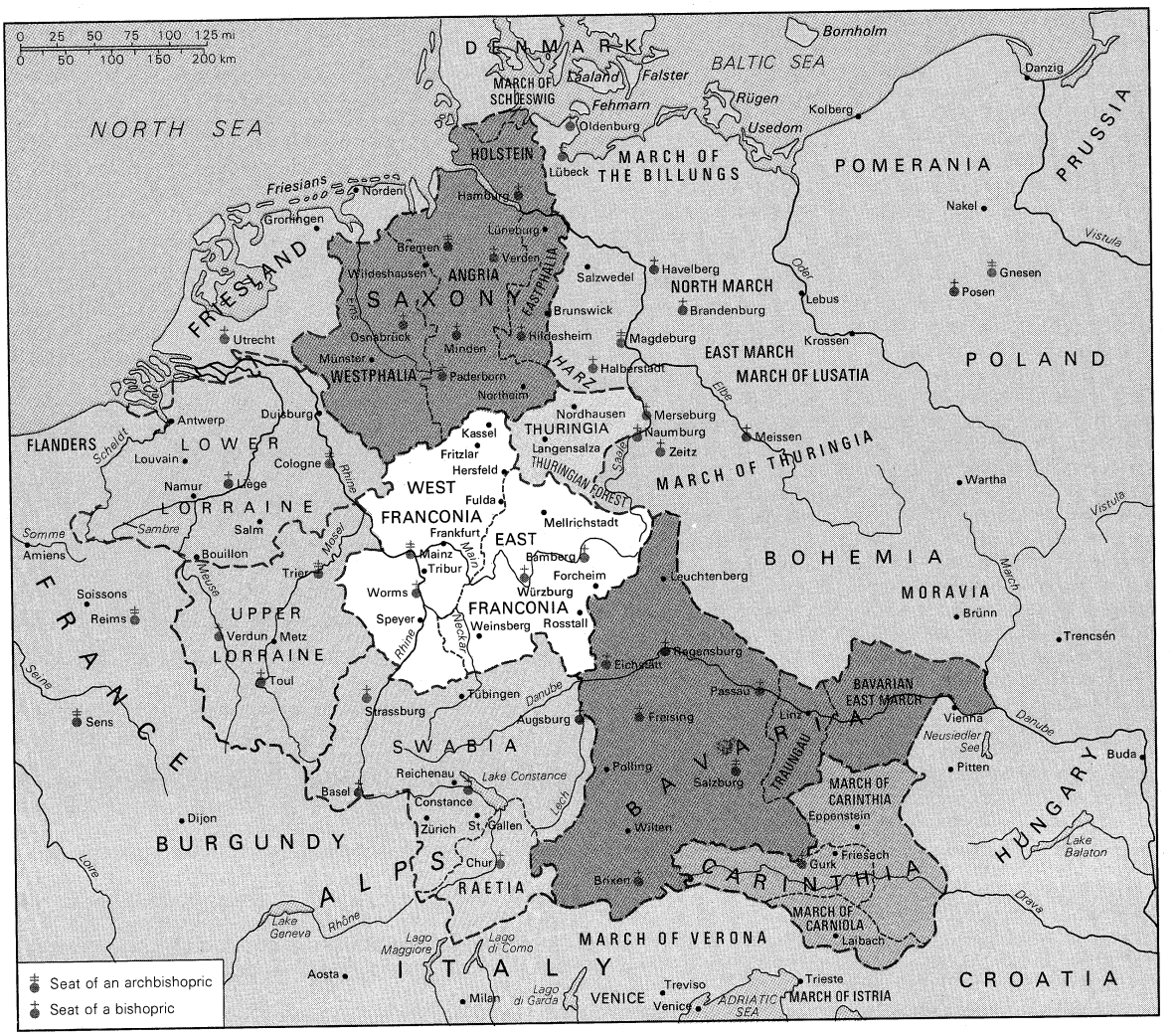
Conrad I (911–918). When in 911 Louis the Child, last of the East Frankish Carolingians, died without leaving a male heir, it seemed quite possible that his kingdom would break into pieces. In at least three of the four stem lands, Bavaria, Saxony, and Franconia, the ducal families were established in the leadership of their tribes; in Swabia (Alemannia) two houses were still fighting for hegemony. Only the church, fearing for its endowments, had an obvious interest in the future of the monarchy, its ancient protector. Against the growing authority of the dukes and the deep differences in dialect, in customs, and in social structure between the tribes there stood only the Carolingian tradition of kingship; but, with Charles the Simple as holder of the West Frankish kingdom, its future was certain and not very hopeful. Only the Lotharingians put their faith in the ancient line and did homage to Charles, its sole reigning representative. The other component parts of the East Frankish kingdom did not follow suit.

One can only guess at the motives of the Saxon and Frankish tribal hosts who on November 10, 911, elected Conrad, duke of the Franks, as their king at Forchheim in Franconia. At the opening of the 10th century the Germanic peoples in the lands east of the Rhine and west of the Elbe, the Saale, and the Bohemian forest—as rude and as thinly spread as their settlements were—had to face even more savage and pagan races pressing in from farther east, especially the Magyars. The Saxons, headed by their duke Otto of the house of the Liudolfings, were threatened by more enemies on their frontiers than any other tribe; Danes, Slavs, and Magyars simultaneously harassed their homeland. A king who commanded resources farther west, in Franconia, might therefore prove to be of help to Saxony. The Rhenish Franks, on the other hand, did not wish to abdicate from their position as the leading and kingmaking people, which gave them many material advantages.

Struggle against the Magyars

Conrad of Franconia, elected by Franks and Saxons, was soon recognized also by Arnulf, duke of Bavaria, and by the Swabian clans. In descent, honours, and wealth, however, Conrad was no more than the equal of the dukes who had accepted him as king. To gain a lead over them, to found a new royal house, and to acquire those wonder-working attributes that the Germans venerated in their rulers long after they had been converted to Christianity, he had yet to prove himself able, lucky, and successful. In this period, political affairs became the monopoly of the German kings and a few score families of great magnates. The reason for this concentration of power was that, at the very foundation of the German kingdom, circumstances had long favoured those men whom birth, wealth, and military success had raised well above the ranks of the ordinary free members of their tribe. Their estates were cultivated in the main by half-free peasants—slaves who had risen or freemen who had sunk. The holdings of these dependents fell under the power of the lord to whom they owed service and obedience. Already they were tied to the lands on which they laboured and were dependent on their protectors for justice. For many reasons ordinary freemen tended generally to lose their independence and had to seek aid from their more fortunate and powerful neighbours; thus, they lost their standing in the assemblies of their tribe. Everywhere, except in Friesland and parts of Saxony, the nobles wedged themselves between king or duke and the rank and file. They alone could become prelates of the church, and they alone could compete for the possession

Development of serfdom



Germany in the 10th and 11th centuries.

and enjoyment of governmental rights. At the level below the dukes, the bulk of administrative authority, jurisdiction, and command in war lay with the margraves and counts, whose hold on their charges developed gradually into hereditary right. The commended men and the half free disappeared from the important functions of public life. In the local assemblies they came only to pay dues and to receive orders, justice, and penalties. Their political role was passive. Those lords whose protection was most worth having also had the largest throng of dependents and thus became more formidable to their enemies and to the remaining freemen. Lordship and submission to it were hereditary, and thus the horizon of the dependent classes narrowed until eventually the lord and his officials filled the place of all secular authority and power in their lives. Military strength, the possession of arms and horses, and tactical training in their use were decisive. Most dependent men were disarmed; that became part of their degradation.

The accession of the Saxons. Conrad I was quite unequal to the situation in Germany. According to the beliefs of contemporaries, his failure meant that his house was luckless and lacked the prosperity-bringing virtues that belonged to true kingship. On his deathbed in 918, he therefore proposed that the crown, which in 911 had remained with the Franks, should now pass to the leading man in Saxony, the Liudolfing Henry (later called the Fowler). Henry I was elected by the Saxons and Franks at Fritzlar, their ancient meeting place, in 919. With a monarch of their own race, the Saxons now took over the burden and the rewards of being the kingmaking people. The centre of gravity shifted to eastern Saxony, where the Liudolfing lands lay.

The election of Henry I the Fowler

The transition of the crown from the Franks to the Saxons for a time enhanced the self-sufficiency of the south German tribes. The Swabians had kept away from the Fritzlar election. The Bavarians believed that they had a better right to the Carolingian inheritance than the Saxons (who had been remote outsiders in the 9th century) and in 919 elected their own duke Arnulf as king. They, too, wanted to be the royal and kingmaking people. Henry I's regime rested in the main on his own position and family demesne in Saxony and on certain ancient royal seats in Franconia. His kingship was purely military. He hoped to gather authority by waging successful frontier wars and to gain recognition in the first place by concessions rather than to insist on the sacred and priest-like status of the royal office that the church had built up in the 9th century. At his election he refused to be anointed and consecrated by the Archbishop of Mainz. In settling with the Bavarians, he abandoned the policy of supporting the internal opposition that the clergy offered to Duke Arnulf, a plank to which Conrad had clung. To end Arnulf's rival kingship, he formally surrendered to him the most characteristic privilege and honour of the crown: the right to dispose of the region's bishoprics and abbeys. Arnulf's homage and friendship entailed no positive obligations toward Henry, and the Bavarian duke pursued his own tribal interests—peace with the Hungarians and expansion across the Alps—as long as he lived.

From these unpromising beginnings the Saxon dynasty not only found its way back to Carolingian traditions of government but soon got far better terms in its relations with the autonomous powers of the duchies, which had gained such a start on it. But the constitution that it bequeathed to its Salian successors was self-contradictory; while seeking to overcome the princely aristocracies of

Reliance on the clergy

the stem lands by leaving them to themselves, the Saxon kings came to rely more and more, both for the inspiration and for the practice of government, on the prelates of the church, who were themselves recruited from the ranks of the same great families. They loaded bishoprics and abbeys with endowments and privileges and thus gradually turned the bishops and abbots into princes with interests not unlike those of their lay kinsmen. These weaknesses, however, lay concealed behind the personal ascendancy of an exceptionally tough and commanding set of rulers up to the middle of the 11th century. Thereafter, the ambiguous system could not take the strain of the changes fermenting within German society and even less the attack on its values that came from without—from the reformed papacy.

The Liudolfing kings won military success, and with it they gained that respect for their personal authority that counted for so much at a time when the great followed only those whose star they trusted and who could reward services with the spoils of victory. In 925 Henry I brought Lotharingia back to the East Frankish connection. Whoever had authority in that half-French-speaking, half-German-speaking region could treat the neighbouring kingdom of the West Franks as a dependent. The young Saxon dynasty thus won for itself and its successors a hegemony over the west and the southwest that lasted at least until the mid-11th century. The Carolingian kings of France, as well as the great feudatories who sought to dominate if not to ruin them, became, in turn, petitioners and even vassals of the German court during the reign of the Ottos. The kings of Burgundy—whose suzerainty lay over the valleys of the Saône and the Rhône, the western Alps, and Provence—likewise fell under the tutelage of the masters of Lotharingia. Rich in ancient towns, this region, once the homeland of the Carolingians, was more thickly populated and wealthier than the lands east of the Rhine. Lotharingian merchants controlled the slave trade from the Saxon marches to Córdoba.

The eastern policy of the Saxons. Greater prestige still and a claim to imperial hegemony fell to the Saxon rulers when they broke the impetus of the Hungarian invasions, against which the military resources and methods of western European society had almost wholly failed for several decades. In 933, after long preparations, Henry routed a Hungarian attack on Saxony and Thuringia. In 955 Otto I (reigned 936–973), at the head of a force to which nearly all the tribes had sent mounted contingents, annihilated a great Hungarian army on the Lech River near Augsburg. The battle again vindicated the efficiency of the heavily armed man skilled in fighting on horseback.

Defeat of the Hungarians at the Lech

With a Saxon dynasty on the throne, Saxon nobles gained office and power, with opportunities for conquest along the eastern river frontiers and marches of their homeland. Otto I indeed had an eastern policy that aimed at getting more than slaves, loot, and tribute. Between 955 and 972 he founded and richly endowed an archbishopric at Magdeburg, which he intended to be the metropolis of a large missionary province among the heathen Slavs beyond the Elbe. This would have brought their tribes under German control and exploitation in the long run; but the ruthless methods of the Saxon lay lords clashed with the church's efforts at more peaceful penetration.

In the 10th century there was little or no German agricultural settlement beyond the Elbe. Far too much forest clearing remained to be done in all the regions of western and southern Germany. The Saxon conquests up to the Oder were secured by military strongholds, called burgwards, and were held only as long as their garrisons had the upper hand. Beyond the Slav peoples of Brandenburg and Lusatia, moreover, new powers rose: the Poles under Mieszko I and, to the south, the Czechs under the Přemyslids received missionaries from Passau and Magdeburg without falling permanently under the political and ecclesiastical domination of Bavarians and Saxons. The heathen Elbe Slavs, subjugated by the Saxon margraves, rose in 983 when the military occupation collapsed along with the missionary bishoprics that had been founded at Oldenburg, Brandenburg, and Havelberg. Farther south the defenses of the Thuringian marches between the Saale and the middle Elbe remained in German hands, but only after a long and fierce struggle against Polish invaders early in the 11th century. The northern part of the frontier reverted to what it had been before Otto's trustees, Hermann Billung and Gero, opened their wars. Missionary enterprises directed from Bremen and Magdeburg achieved little before the 12th century. The Saxon ruling class, bishops, and margraves must bear the responsibility for the fiasco of eastward expansion in the 10th century. The prelates, too, saw their missions as means to found ecclesiastical empires with subject dioceses and tithes on Slav soil. The tribes across the Elbe therefore remained unconverted and implacable foes, a standing menace to the nearby churches. The wars also left a legacy of savagery on both sides so that from *c.* 1140 onward the substitution of German settlers for the native Slavs became the common policy of both the church and the princes.

Legacy of savagery

Dukes, counts, and advocates. Conrad I's and Henry I's kingships rested on the will of the tribes—or rather on that of their leaders and of the higher aristocracy. It was in the first place an arrangement between the Franks and the Saxons that the Bavarian and Swabian dukes recognized at a price by acts of personal homage. But the German kings, of whatever dynasty, had to live under Frankish law. After the death of Conrad I's brother Eberhard in 939, Otto I kept the Franconian dukedom vacant and the Franconian counts henceforth stood under the immediate authority of the crown. In Saxony, too, Otto kept in his hands the dukedom of his ancestors. The marchduchy of the Billungs, a bulwark raised against the Danes and the northern Slav tribes, did not give that family authority over all the other Saxon princes.

In the south the Ottonians sought to turn the dukedoms of the stem lands into royal fiefdoms and to supplant native dynasties by aliens and members of their own clan. When even that policy did not stop rebellions under the banner of tribal self-interest, they began to break up the ancient Bavarian stem land by creating a duchy in Carinthia to cut off the spearhead of Bavarian expansion southward. The first two Salians, Conrad II (reigned 1024–39) and Henry III (reigned alone 1039–56), also bestowed vacant duchies quite freely on their own kin and on men from outside the stem boundaries. They competed against ducal power but could neither abolish nor replace it. In the 11th century as before, the dukes held assemblies of their folk, led the tribal host in war, and enforced peace.

The counts, who were the ordinary officers of justice in serious, criminal cases, obeyed the ducal summons; but, for the most part, they received their "ban," the power to do blood justice, from the king himself. The fiefs and the customary rights attached to their office, and indeed the office itself, not only became hereditary but also came to be treated more and more as a patrimony to which they had an inherent right against all men, king and duke included. Even so, however, a good many lines died out and their counties fell back into the king's hands. From Otto III's reign (983–1002) onward, it became not at all unusual to bestow these counties on bishoprics and certain great abbeys rather than to grant them out again to other lay magnates. The bishops, however, could not perform all the functions of the counts; in particular, their holy orders forbade them to pass judgments of blood. They needed officials called advocates (*Vögte;* singular *Vogt*) to take charge of the higher jurisdiction in the counties and franchises that their churches possessed by royal grant. In the 10th and 11th centuries these advocates had to be recruited from the aristocracy, the very class whose greed for hereditary office was to be checked, because ordinary freemen could not enforce severe sentences or defend the privileges of the church against armed intrusion. Dangerous neighbours of bishoprics and abbeys in any case, the nobles as advocates and protectors of ecclesiastical possessions were anything but reliable servants of their ecclesiastical overlords.

The rise of the advocates

Thus, there arose in nearly all German lands, whether

the ducal office survived or not, powerful lines of margraves, counts, and hereditary advocates who enriched themselves at the expense of the church (which meant also the crown) and in competition with one another. From the abler, more fortunate, and long-lived races among these dynasts sprang the territorial princes of the later 12th and 13th centuries, absorbing and finally inheriting most of the rights of government.

Power of the kings

The king was the personal overlord of all the great. His court was the seat of government, and it went with him on his long journeys. The German kings, even more than other medieval rulers, could only make their authority respected in the far-flung regions of their kingdom by travelling ceaselessly from duchy to duchy, from frontier to frontier. Wherever they stayed, their jurisdiction superseded the standing power of dukes, counts, and advocates, and they could collect the profits of local justice and wield some control over it. As they came into each region, they summoned its leaders to attend their solemn crown wearings, deliberated with them on the affairs of the *Reich* and the locality, presided over pleas, granted privileges, and made war against peacebreakers at home and on enemies abroad.

The promotion of the German Church. The royal revenues came from the king's demesne lands and from his share of the tributes that Poles, Czechs, heathen Slavs, and Danes had to pay whenever he could enforce his claims of overlordship. The king's demesne was his working capital. He and his household lived on its produce during their wanderings through the *Reich*, and it also served to provide for his family, to found churches, and to reward faithful services done to him, especially in war. To swell the hosts, vassals had to be enfeoffed, and alienations were inevitable. The Salians, though they inherited the remains of Ottonian wealth as imperial demesne, brought little of their own to make up for its diminution. Already the last Saxon, Henry II (1002–24), and after him Conrad II accordingly took to enfeoffing vassals with lands commandeered from the monasteries. But the beneficiaries often enough were already powerful and wealthy men in their own right, so that no class of freeborn, mounted warriors linked permanently with the crown sprang from the loyalties and rewards of one or two reigns. In any case, the lion's share of grants went to the German Church.

The growth of central government

From the Carolingians, the German kings inherited their one and only institution of central government: the royal chapel, with the chancery that does not seem to have been distinct from it. Service there became a recognized avenue of promotion to the episcopate for highborn clerks. In the 11th century, bishops and abbots conducted the affairs of the *Reich* much more than the lay lords, even in war. They were its habitual diplomats and ambassadors. Unlike Henry I, Otto I and his successors sought to free the prelates from all forms of subjection to the dukes. The king appointed them, and to him alone, as to one sent by God, they owed obedience.

Thus, there arose besides the loose association of stem lands in the German kingdom a more compact and uniform body with a far greater vested interest in the *Reich:* the German Church. By ancient Germanic custom, moreover, the founder of a church did not lose his estate in the endowment that he had made; he remained its proprietor and protecting lord. The bishoprics, it is true, and certain ancient abbeys, such as Sankt Gallen, Reichenau, Fulda, and Hersfeld, did not belong to the king; they were members of the kingdom but under his guardianship. The greater churches therefore had to serve the rulers with mounted men, money, and free quarters. Gifts of royal demesne to found or to enrich bishoprics and convents were not really alienations but pious reinvestments, as long as the crown controlled the appointments of bishops and abbots. But the church did not merely receive grants of land, often waste, to settle, develop, and make profitable: it was also given, as has been shown, powers of jurisdiction over its dependents. Nor did the kings stint the prelates in other regalian rights, such as mints, markets, and tolls. These grants broke up counties and to some extent even duchies, and that was their purpose: to disrupt the secular lords' jurisdictions that had escaped royal control.

This policy of fastening the church, a universal institution, into the *Reich*, with its well-defined frontiers, is usually associated with the name of Otto I. But it gathered momentum only in the reigns of his successors. It reached a climax under Henry II, the founder of the see of Bamberg in the Upper Main Valley; but Conrad II, though less generous with his grants, and his son Henry III continued it. Bishops and abbots became the competitors of lay princes in the formation of territories, a rivalry that more than any other was the fuel and substance of the ceaseless feuds, the smoldering internal wars in all the regions of Germany for centuries. The welter and the confused mosaic of the political map of Germany until 1803 is the not-so-remote outcome of these 10th- and 11th-century grants and of the incompatible ambitions that they aroused.

The Ottonian conquest of Italy and the imperial crown. Otto I's marriage with Adelaide (Adelheid), daughter of Rudolph II of Burgundy, and the Italian rivalries between his brother Henry, duke of Bavaria, and Otto's son Liudolf, duke of Swabia, drew him southward. After 951, expeditions into Italy were a matter for the whole *Reich* under the leadership of its ruler and no longer just an outlet for the expansion of the south German tribes. For the Saxon military class, too, the south was more tempting than the primeval forests and swamps beyond the Elbe. With superior forces at their back, the German kings gained possession of the Lombard kingdom in Italy. There, too, their overlordship in the 10th and the 11th centuries came to rest on the bishoprics and a handful of great abbeys.

After his victory over the Magyars in 955, Otto I's hegemony in the west was indisputable. By the standards of one chronicler, the Saxon Widukind, he had already become emperor because he had subjected other peoples and enjoyed authority in more than one kingdom. But the right to confer the imperial crown, to raise a king to the higher rank of emperor, had fallen to the papacy, which had crowned Charlemagne and most of his successors. The Carolingian order in the west was still the model and something like a political ideal for all Western ruling families in the 10th century. Otto had measured himself against the political tasks that had faced his East Frankish predecessors and more or less mastered them. To be like Charlemagne, therefore, and to clothe his newly won position in a traditional and time-honoured dignity, he accepted the imperial crown and anointment from Pope John XII in Rome in 962. The substance of his empire was military power and success in war; but Christian and Roman ideas were woven round the Saxon's throne by the writers of his own and the next generation. Although the German kings as emperors did not give the law to the Roman Church in matters of doctrine and ritual, they became its political masters for nearly a century. The imperial crown enhanced their standing even among the nobles and knights who followed them to Italy and can hardly have understood or wanted all its outlandish associations. Not only the king but also the German bishops and lay lords thus entered into a permanent connection with an empire won on the way to Rome and bestowed by the papacy.

The formation of the Holy Roman Empire

Otto II (reigned alone 973–983) and above all Otto III (983–1002) were strongly drawn toward their new Mediterranean sphere of action, but Henry II returned to a sober regime centred on Germany and contented himself with three brief Italian expeditions.

The Salians, the papacy, and the princes, 1024–1125. Under Conrad II (reigned 1024–39), the first member of the Rhine–Frankish house known as the Salians, the kingdom of Burgundy fell finally under the overlordship of the German crown, and this tough and formidable emperor also renewed German authority in Italy. His son and successor, Henry III (reigned 1039–56), treated the empire as a mission that imposed on him the tasks of reforming the papacy and of preaching peace to his lay vassals. Without possessing any very significant new resources of power, he gave to his authority an exalted and

strained theocratic complexion. Yet, under him, the last German ruler to maintain his hegemony in western Europe, the popes themselves seemed to become mere imperial bishops. He deposed three of them, and four Germans held the Holy See at his command; but lay opposition to the Emperor in Germany and criticism of his regime over the church were on the increase during the last years of his reign.

The papal reforms and the German church. More than any other feudal society in early medieval Europe, Germany was divided and torn by the revolutionary ideas and measures of the reformed papacy. Beginning with the pontificate of Leo IX—one of Henry III's nominees—the most determined and inspired spokesmen of ecclesiastical reform placed themselves at the service of the Holy See. Only a few years after Henry III's death (1056), they agitated against lay authority in the church, founded on proprietary rights. They regarded the laity as passive partakers of the sacraments and denied the supernatural status of kingship. Priests, including bishops and abbots, who accepted their dignities from lay lords and emperors at a price, according to the reformers, committed a sin; for these earthly powers could not rightly confer churches at all, nor could they own them. They believed, moreover, that thorough reforms could only be brought about by the exaltation of the papacy so that it commanded the obedience of all provincial metropolitans and was out of the emperor's and the local aristocracy's reach.

Effects of the reformed papacy

The endless repetition of the reformers' teachings in brilliant pamphlets and at clerical synods spread agitation in Italy, Burgundy, and Lotharingia—all parts of the empire. Their new program committed the leaders of the movement to a struggle for power because it struck at the very roots of the regime to which the German church had grown accustomed and on which the German kings relied. The vast wealth that Henry IV's predecessors had showered on the bishoprics and abbeys would, if the new teaching prevailed, escape his control and remain at the free disposal of prelates whom he no longer appointed. Under Roman authority the churches were to be freed from most of the burdens of royal protection without losing any of its benefits. The most fiery spirits in Rome did not flinch from the consequences of their convictions. Their leader Hildebrand, later Pope Gregory VII (reigned 1073–85), was ready to risk a collision with the empire.

Henry IV was not yet six years old when his father died in 1056. The full impact of the Gregorian demands—coming shortly after a royal minority, a Saxon rising, and a conspiracy of the south German princes—has often been regarded as the most disastrous moment in Germany's history during the Middle Ages. In fact, the German church proved thoroughly unreliable as an inner bastion of the empire even before Rome struck. Its leaders, Anno and Adalbert, archibishops of Cologne and of Hamburg–Bremen, respectively, shamelessly exploited their hold over the young king by hunting for spoils out of the imperial demesne. In 1074 Gregory proceeded against simony (the buying and selling of church office) in Germany, humiliated the aristocratic episcopate by summonses to Rome and sentences of suspension, and, a year later, forbade rulers to invest bishops and abbots with their churches. These papal actions demoralized and shook the German hierarchy. The prelates' return to their customary support of the crown was not disinterested, nor whole hearted, nor unanimous.

Decrees of Gregory VII

The discontent of the lay princes. Henry IV's minority also gave elbowroom to the ambitions and hatreds of the lay magnates. The feeble regency of his mother, Agnes of Poitou, faltered before the throng of princes, who respected only authority and forces greater than their own. The ruling influence of the higher clergy at the court of Henry III and the renewed flow of grants to the church had estranged them from the empire. It is likely also that these eternally belligerent men were lagging behind the prelates in the development of their agrarian resources. The prelates had a vested interest in peace, and under royal protection they improved and enlarged their estates by turning forests into arable land and also by offering better terms to freemen in search of a lord. The bishops' market and toll privileges brought them revenues in money, which many of the lay princes lacked. So far, however, the princes' military power, their chief asset, had remained unchallenged. Now, for the first time, they also had to face rivals within their own sphere of action. Henry III and the young Henry IV began to rely on advisers and fighting men drawn from a lower tier of the social order—the poorer, freeborn nobility of Swabia and, above all, the class of unfree knights, known as *ministeriales*. Those knights had first become important as administrators and soldiers on the estates of the church early in the 11th century. Their status and that of their fiefs was fixed by seignorial ordinances, and they could be relied on and ordered about, unlike the free vassals of bishops and abbots. Beginning with Conrad II, the Salian kings used *ministeriales* to administer their demesne, as household officers at court and as garrisons for their castles. They formed a small army, which the crown could mobilize without having to appeal to the lay princes, whose ill will and antipathy toward the government of the *Reich* grew apace with their exclusion from it.

Having come of age, Henry IV used petty south German nobles and his *ministeriales* to recover some of the crown lands and rights, which the lay princes and certain prelates had acquired during his minority, particularly in Saxony. His recovery operations went further, however, and a great belt of lands from the northern slopes of the Harz Mountains to the Thuringian Forest was secured and fortified under the supervision of his knights to form a compact royal territory, where the King and his court could reside almost continuously. The south German magnates were thus kept at a distance when Henry and his advisers struck at such neighbouring Saxon princes as Otto of Northeim and the Billung family.

Formation of a royal territory

The storm broke in 1073. A group of Saxon nobles and prelates and the free peasantry of Eastphalia, who had to bear the brunt of statute labour in the building of the royal strongholds, revolted against the regime of Henry's Frankish and Swabian officials. To overcome this startling combination and to save his fortresses, the King needed the military strength of the south German princes Rudolf of Rheinfelden, duke of Swabia; Welf IV, duke (as Welf I) of Bavaria; and Berthold of Zähringen, duke of Carinthia. Suspicious and hostile at heart, they took the field for him only when the Eastphalian peasantry committed outrages that shocked aristocratic caste feeling everywhere. Their forces enabled Henry to defeat the Saxon tribal rebellion near Langensalza in June 1075. But, when the life-and-death struggle with Rome opened only half a year later, the south German malcontents deserted Henry and, together with the Saxons and a handful of bishops, entered into an alliance with Gregory VII. Few of them at this time were converted to papal reform doctrines, but Gregory's daring measures against the King gave them a chance to come to terms with one another and to justify a general revolt.

The civil war against Henry IV. On February 22, 1076, the Pope absolved all men from their oaths to Henry and solemnly excommunicated him. In October Gregory's legates met the German lords at Tribur (modern Trebur) to decide on the future of the King, whom his last adherents now abandoned. Although Henry was absolved by Gregory at Canossa in January 1077, the princes two months later nonetheless elected Rudolf of Rheinfelden to rule in his place.

The war that now broke out lasted for almost 20 years. A majority of the bishops, most of Rhenish Franconia (the Salian homeland), and some important Bavarian and Swabian vassals sided with Henry. He thus held a central position, dividing his south German from his Saxon enemies, who could not unite long enough to destroy him. With the death in battle of Rudolf of Rheinfelden (1080) and the demise of another anti-king, Hermann of Salm (1088), the war in Germany degenerated into a number of local conflicts for the possession of bishoprics and abbeys. It almost died down in 1098, when the south German adherents of the papacy came to terms with Henry

Division of Henry's enemies

for the time being, though without recognizing his antipope Clement III. Throughout these years the crown, the churches, and the lay lords had to enfeoff more and more *ministeriales* in order to raise mounted warriors for their forces. Though this recruitment and frequent devastations strained the fortunes of many nobles, they knew how to recoup themselves by extorting more fiefs out of neighbouring bishoprics and abbeys. The divided German Church thus bore the brunt of the costs of civil war, and it needed peace almost at any price.

Henry V and results of the conflict. The Salian dynasty and the rights for which it fought were saved because Henry IV's son and heir himself seized the leadership of a last rising against his father (1105). This manoeuvre enabled Henry V (reigned 1106–25) to continue the struggle for the crown's prerogative over the empire's churches against the inexorable demands of the papacy. The conflict now shrank into a legalistic dispute over the right to invest bishops and abbots with their dignities and the secular possessions attached to them. As the struggle continued, the princes became the arbiters and held the balance between their overlord and the Pope. In 1122, acting as intermediaries and on behalf of the *Reich*, they forced the temporary concessions known as the Concordat of Worms out of the Holy See and its German spokesman, Archbishop Adalbert of Mainz, the bitter personal enemy of Henry V and the territorial rival of the Hohenstaufen sons of Henry's sister Agnes. But by then the princes had for the most part defeated efforts to restore royal rights in Saxony and to stem the swollen jurisdictions and territorial powers of the aristocracy elsewhere.

German loss of power after the death of Henry V

When Henry V, the last Salian, died childless in 1125, Germany was no longer the most effective political force in Europe. The brillant conquest states of the Normans in England and in Sicily and the patient, step-by-step labours of the French kings were achieving forms of government and concentrations of military and economic strength that the older and larger empire lacked. The papacy had dimmed the empire's prestige, and Rome became the true home of universalistic interests. When Pope Urban II preached the first crusade in 1095, Henry IV, cut off and surrounded by enemies, was living obscurely in a corner of northern Italy. The Holy See, by its great appeal to the militant lay nobility of western Europe, thus won the initiative over the empire. At this critical moment the *Reich* also lost control in the Italian bishoprics and towns, just when their population, trade, and industrial production were expanding fast. Germany did not even benefit indirectly from the crusaders' triumphs, although some of their leaders (*e.g.*, Godfrey of Bouillon and Robert II of Flanders) were vassals of the Emperor. The civil wars renewed for a time the relative isolation of the southern and central German regions.

Internally, the crown had saved something of the indispensable means of government in the control over the church; but it was a bare minimum, and its future was problematic. The ecclesiastical princes henceforth held only their temporal lands as imperial fiefs, for which they owed personal and material services. As feudatories of the empire, they came to represent the same interests toward it as did the lay princes; at least, their sense of a special obligation tended to weaken. The king's jurisdiction continued to exist side-by-side and in competition with that of the local powers. The great tribal duchies survived as areas of separate customary law. Each developed differently, and the crown could not impose its rights on all alike or change the existing social order. The most tenacious defenders of this legal autonomy had been the Saxons; but it also prevailed in Swabia, where distinct territorial lordships grew fast.

The Gregorian reform movement therefore aggravated the age-old contradictions in Germany's early medieval constitution. But its monastic culture and its intellectual interests were anything but barren. Both sides fought with new literary weapons to work on public opinion in cathedrals and cloisters and perhaps also in the castles of the lay aristocracy. In their hard-hitting polemical writings they attempted to expound the fundamental theological, historical, and legal truths of their cause. The agitation did something to disturb the cultural self-sufficiency of the German laity. It drove many of the south German nobles to maintain direct connections with the Holy See, and, whether they wanted it or not, they had to fall in with the aspirations of the religious leaders. The reform movement of the 11th and 12th centuries, it might almost be said, very nearly completed the conversion of Germany that had begun five centuries before.

GERMANY AND THE HOHENSTAUFEN, 1125–1250

Dynastic competition, 1125–52. The nearest kinsmen of Henry V were his Hohenstaufen nephews—Frederick, duke of Swabia (1105–47), and his younger brother Conrad, the sons of Henry's sister Agnes and Frederick, the first Hohenstaufen duke of Swabia. Some form of election had always been necessary to succeed to the crown, but, before the great civil war, nearness to the royal blood had been honoured whenever a dynasty failed in the direct line. By 1125, however, the princes, guided by Archbishop Adalbert of Mainz, no longer respected blood right. Affinity with Henry V was no recommendation to them, and hereditary succession seemed to lower their authority in the government of the *Reich*. Instead of Frederick they chose the duke of Saxony, Lothair of Supplinburg (reigned as king 1125–37, reigned as emperor 1133–37). Like the Hohenstaufen, he had risen by a lucky marriage and a successful career of continuous fighting into the first rank of dynasts; but, unlike them, he had served the cause of the Saxon opposition to the Salians.

Disregard for blood right

With the enormous Northeim and Brunonian inheritances behind him, Lothair III (sometimes called Lothair II) could humble the Hohenstaufen brothers (1134) after marrying his only daughter and heiress to a Welf, Henry the Proud. Even without his dazzling alliance, the Welfs, already dukes of Bavaria and possessors of vast demesnes, countships, and ecclesiastical advocacies there and in Swabia, were somewhat better off than their Hohenstaufen rivals. On the death of Lothair in 1137, however, the fears of the church and a few princes turned against the Welfs. Instead of Henry the Proud, who now held the duchies of Saxony and Bavaria and the Matildine lands in Italy, they chose Conrad (reigned 1138–52), who had been Lothair's unsuccessful Hohenstaufen opponent.

The battle against the Welfs, which Conrad III put foremost on his political program, was abandoned with his death in 1152, when an election once again decided the succession and the political situation in Germany for the next 30 years. The princes then chose Frederick I Barbarossa (reigned as king 1152–90), the son of Conrad's elder brother Frederick and the Welf princess Judith. Frederick I agreed to share power in Germany with his Welf cousin Henry the Lion. The price of his election was dualism. In 1156 the duchy of Bavaria, which Conrad had tried to wrest from the Welfs, was restored to Henry the Lion, already undisputed duke of Saxony. The Babenberg margrave of Austria, Henry's rival, had to be compensated with a charter that raised his margravate into a duchy and gave him judicial suzerainty over an even wider area. Taken out of the Lion's duchy, it was to be held as an imperial fief that might descend both to sons and daughters. A perpetual principality, it served as a model for the aspirations of many other lay princes.

Colonization of the east. The history of Germany in the 12th and 13th centuries is one of ceaseless expansion. A conquering and colonizing movement burst across the river frontiers into the swamps and forests from Holstein to Silesia and overwhelmed the Slav tribes between the Elbe and the Oder. Every force in German society took part: the princes, the prelates, new religious orders, knights, townsmen, and peasant settlers. Agrarian conditions in the older lands of Germanic occupation seem to have favoured large-scale emigration. With a rising population, there was much experience in drainage and wood clearing but a diminishing fund of spare land to be attacked in the west. Excessive subdivision of holding impoverished tenants and did not suit the interests of their

The conquering and colonizing movement of the 12th and 13th centuries

lords. Sometimes also, seignorial oppression is said to have driven peasants to desert their masters' estates. They certainly found a better return for their labour in the colonial area: personal freedom, secure and hereditary leasehold tenures at moderate rents, and, in many places, quittance from services and the jurisdiction of the seignorial advocate.

The colonists brought with them a disciplined routine of husbandry, an efficient plow, and orderly methods in siting and laying out their villages. Very soon, even the Slav rulers of Bohemia and Silesia, were competing for immigrants. First and foremost, however, the princes of the Saxon and Thuringian marches sought to attract settlers for the lands that they had conquered and the towns that they had founded to open up communications and trade routes. The older regions of the *Reich*, moreover, had not only peasants but also men of the knightly class to spare —soldiers who needed fiefs and lordships to uphold their rank. Both could be gained beyond the Elbe under the leadership of successful princes. The Germanized east thus became the home of fair-sized principalities in the 13th century, while all along the Rhine River Valley the rights of government were tending to be scattered over smaller and less compact territories. The Ascanian dynasty, for instance, which under Albert the Bear began to advance into Brandenburg, by 1250 not only ruled over a broad belt of land up to the Oder River but had already established itself on the eastern banks ready for further advances. Farther south the Wettin margraves of Meissen busied themselves with settlements and town foundations in Lusatia.

For a time Henry the Lion, as duke of Saxony (1142–80), overshadowed all these rising powers, and the Welf profited as much by the ruthless use of his resources against weaker competitors as by his own efforts in Mecklenburg. As his protection was alone worth having in northeastern Germany, the newly established Baltic bishoprics were at his mercy, and he alone could attract the traders of Gotland to frequent the young port town of Lübeck, which he extorted from one of his vassals in 1158.

The *Reich*, too, possessed demesnes in the east, notably the Egerland, Vogtland, and Pleissnerland in the Thuringian March. The Hohenstaufen kings therefore took some part in opening up these regions. They, too, founded towns and monasteries on their thickly wooded lands and established their *ministeriales* as burgraves and advocates over them. But, in this, as in many other things, they only competed with the princes. They did not and could not control the eastward movement as a whole.

Adapted from J.R. Strayer and D.C. Munro, *The Middle Ages, 395–1500*, 5th ed., p. 521, copyright © 1970; by permission of Appleton-Century-Crofts, Educational Division, Meredith Corporation



German expansion to the east, 800–1400.

Hohenstaufen policy in Italy. In the other great field of German expansion in the 12th century—Lombardy and central Italy—the emperors and their military following alone counted. The rural population of Germany had no direct interest in the wars waged to recover and exploit ancient regalian rights over the growing Lombard city communes. The connection between the German crown, the empire, and dominion over Italy has indeed been regarded as a disaster for Germany, and the ever-increasing concern of the Hohenstaufen dynasty with the south as its most tragic phase. But, although Frederick Barbarossa's policy was opportunistic, he had really very little choice. Having bought off the Welfs and reconciled other great families with yet more concessions and lastly endowed his own cousin, Conrad III's son Frederick, with Hohenstaufen demesnes in Swabia, he had to try to mobilize their goodwill for the empire while it lasted. He now aimed to set up a regime of imperial officials and captains who were to exact dues and to control jurisdiction that the communes had usurped from the failing grasp of the bishops. The Germans in Italy did not bring valuable accomplishments to poor and savage tribesmen, but they attacked economically advanced and better developed communities, to which they had nothing to offer in return for the rights and taxes they demanded. Military power was their chief asset in Lombardy, and they used it ruthlessly.

The disastrous Italian policy

For the Hohenstaufen *ministeriales* the rule of their masters in northern and central Italy was a career. Because they could be deployed continuously, they became the backbone of the imperial occupation. A handful of minor dynasts also served Barbarossa for many years in the powerful and profitable commands that he established. The German bishops and certain abbots still had to supply men and money, and some of them threw themselves wholeheartedly into the war: for instance, Rainald of Dassel and Philip of Heinsberg, archbishops of Cologne from 1159 to 1167 and from 1167 to 1191, respectively, who, as archchancellors for Italy, had a vested interest in it. But the support of the lay princes was fitful and sporadic. Even at critical moments they could not be counted on unless they individually agreed to serve or to send their much-needed contingents for a season. The refusal of the greatest of them, Henry the Lion, in 1176 brought about the Emperor's defeat at the Battle of Legnano and spoiled many years' efforts in Lombardy.

The fall of Henry the Lion and the estate of princes. Forced to retreat before the Lombard League in 1177, Barbarossa cooled toward his Welf cousin, whom he could justly blame for some of his setbacks. Dualism in Germany had outlived its purpose. Hitherto, the enemies of Henry—the princes, bishops, and magnates of Saxony—had been unable to gain a hearing against him at the Emperor's court days. By 1178, however, the Emperor was ready to help them. Outlawed (1180), beaten in the field and deserted by his vassals, Henry had to surrender and go into exile in 1182. His duchies and fiefs were forfeited to the *Reich*.

Surrender and exile of Henry the Lion

His fall left a throng of middling princes face-to-face with an emperor whose prestige, despite reverses, stood high and whose resources had greatly increased since he began to reign. The princes were nonetheless the chief and ultimate gainers from the events of 1180. The final judgment by which Henry the Lion lost his honours was not founded on folk law but on feudal custom. The princes who condemned him regarded themselves as the first feudatories of the empire, and they decided on the redistribution of his possessions among themselves. During the 12th century the stem duchies of the Ottonian period finally disintegrated. Within their ancient boundaries not only bishops but also lay lords succeeded in eluding the authority of the dukes. In their large immunities they themselves wielded stem-ducal powers. To enforce the imperial peace-laws became both their ambition and their justification. Everywhere, the greater lay dynasties and even some bishops tried to acquire a ducal or an equivalent title, that would enable them to consolidate their scattered jurisdictions and, if possible, to force lesser freelords to attend their pleas.

These highest dynasts had interests in common, and they closed their ranks not only against threats from above but also against fellow nobles who had been less successful in amassing wealth, counties, and advocacies and who did not possess the superior jurisdiction of a duke, a margrave, a count palatine, or a landgrave. They and they alone were now called princes of the empire. To lend a certain cohesion to their varied rights, they were willing to surrender their houselands to the *Reich* and receive them back again as a princely fief. For the emperor it was theoretically an advantage that men so powerful in their own right should owe their chief dignity and most valued privileges to his grant. It opened the possibility of escheats, or land seizures, for in feudal custom the rules of inheritance were strict. But in Germany the political misfortunes of rulers succeeded, by and large, in ensuring that ancient caste feeling and notions of inalienable right conquered the principles of feudal law. By 1216 it was established that the emperor could neither abolish principalities nor create princes at random.

The "heirs" of Henry the Lion had to fight a ceaseless battle to establish and maintain themselves. In Bavaria the Wittelsbachs had received the vacant duchy, but they were not recognized as superiors by the dukes of Styria or by the dukes of Andechs-Meran. In Saxony the Archbishop of Cologne was enfeoffed with Henry the Lion's ducal office and with all his rights in Westphalia, while an Ascanian prince, Bernard of Anhalt, received the eastern half of Henry's duchy. Neither Bernard nor the Archbishop, however, could make much out of their dukedoms, except in the regions where they already had lands and local jurisdictions. All over the *Reich* these and regalian rights, such as mints, fairs, tolls, and the right of granting safe-conducts, were the substance of princely power. To possess them as widely as possible became the first goal of the abler bishops and lay lords.

The Hohenstaufen conflict with the papacy, 1159–1215. The attempt to establish a direct imperial regime in Italy antagonized the papacy once again and led to a new struggle with Rome, the ally of the Lombard communes. Political and territorial rather than ecclesiastical interests were at stake; but the popes could only fight as heads of the universal church, defending its liberty against a race of persecutors, and they had to employ their characteristic weapons—excommunication, propaganda, and intrigue. Nonetheless, the German bishops stood by Barbarossa and, for the most part, followed him in maintaining a prolonged schism against Pope Alexander III. Unsuccessful in Lombardy, the centre of Hohenstaufen ambitions after 1177 shifted to Tuscany, Spoleto, and the Romagna. This redoubled the fears and the resentment of the popes, particularly when Frederick's son and chosen successor, Henry VI (reigned 1190–97), became after 1189 the legitimate claimant to the Sicilian kingdom through his wife Constance, the sole surviving legitimate heiress.

With their backs to the wall, the popes had to make what use they could out of any opposition to the Hohenstaufen. Their chance came in 1197 when Henry VI died prematurely, leaving a three-year-old son, Frederick, to succeed him. To escape the chaos of a minority regime, the bulk of the German princes and bishops in 1198 elected the boy's uncle Philip of Swabia; but an opposition faction in the Lower Rhenish region, led by the Archbishop of Cologne and financed by Richard I of England, raised an anti-king in Otto IV, a younger son of Henry the Lion. Pope Innocent III had to enlarge on his rights over imperial coronations and become a partisan in the German electoral feud if he wished to defend his recovered holdings in Italy against Hohenstaufen claims. Territorial interests in the Romagna tempted the papacy to exploit the weaknesses of the empire's constitution, the uncertainties of electoral custom, and the lack of strict legal norms in Germany. During the war for the crown, much hard-won demesne and useful rights over the church had to be sacrificed by the rivals to bribe their supporters.

Frederick II and the princes. Henry's son Frederick II entered Germany to regain his own against Otto IV in

1212 and secured the crown in 1215. Despite promises to divide his inheritance, he kept the Kingdom of Sicily and the empire together, and thus he also became locked in the inevitable life-and-death struggle with the papacy. The Hohenstaufen demesne in Swabia, Franconia, and Alsace and on the Middle Rhine was still very considerable, and Frederick even recovered certain fiefs and advocacies that had been lost during the earlier civil wars. Their administration was improved, and they provided valuable forces for his Italian wars. An edict of imperial peace in 1235, moreover, showed that the Emperor had not become a mere competitor in the race for territorial gain. But, except for brief intervals, the princes and bishops were left free to fight for the future of their lands against one another and against the intractable lesser dynasts who refused to accept their domination. The charters that Frederick had to grant to the ecclesiastical princes (the so-called *Confoederatio cum principibus ecclesiasticis*, 1220) and later to all territorial lords (*Constitutio*, or *Statutum in favorem principum*, 1232) gave them written guarantees against the activities of royal demesne officials and limited the development of imperial towns at the expense of episcopal territories. But the charters were not always observed, and until 1250 the crown remained formidable in southern Germany, despite the anti-kings Henry Raspe and William of Holland, whom the papacy caused to be elected by the Rhenish archbishops in Germany in 1246 and 1247.

The Reich after the Hohenstaufen catastrophe. Frederick II died in 1250, in the midst of his struggle against Pope Innocent IV. His son Conrad IV left the north in 1251 to fight for his father's Italian possessions. William of Holland, anti-king from 1247 to 1256, was thus without a rival in an indifferent Germany, which had lost interest in its rulers. The bishops' cities and the towns, many of them founded on royal demesne, could not be absorbed. Their economic power challenged the age-old aristocratic order in German society, and, deprived of royal protection, they banded together to defend their autonomy. Within the nobility each rank tended to acquire some of the personal rights of its betters. The Hohenstaufen breakdown after 1250 left a gap in Swabia that no rising territorial power was able to fill. Countless petty lords and imperial *ministeriales* of the southwest succeeded in holding their seigniories as immediate vassals of the *Reich*. Their independent territories often survived for centuries.

The *ministeriales* elsewhere, too, ceased to be the dependable servants that they once had been. Many free nobles voluntarily joined their ranks, and the knights thus assimilated the rights of the free aristocracy. They became the governing class of the territorial principalities, the standing councillors of their masters, whose household offices and local justice they monopolized and held in fee for many generations. Without the consent of this territorial nobility, the princes could neither tax nor legislate. Even the less important *ministeriales*, who only administered manors for their lords, entrenched themselves as hereditary bailiffs, who kept surplus produce for themselves and usurped seignorial dues, so that it paid the owners to commute the labour services of their villeins into money rents and to lease out those portions of the demesne that the unfree peasants had cultivated for them. Even then, however, the hereditary officials could not be easily dislodged. Finally, the ambitions of the princes themselves did not aim above the patrimonial policies of the past. They were acquisitive and attemped to build up their territories by usurpation, inheritance, marriage treaties, and escheats. They also tried, where possible, to administer their lands with officials whom they could depose at will. Yet, they did so not to found sovereign states but chiefly to provide for their families. Again and again, they divided their dominions among sons, who, in turn, founded cadet lines and set them up on a fraction of the principality.

By 1250 there was thus no really effective central authority left in Germany. The prince-bishoprics had become fiercely contested prizes between neighbouring dynasties, often vassals of the see (*i.e.*, the bishopric). But constant feuds, disorder, and insecurity did not, by any means, frustrate the immense energies of the Germans in the 13th century. Eastward expansion continued under the leadership of the princes and, above all, of the Knights of the Teutonic Order. Their advance into Prussia went hand in hand with the opening up of the Baltic by the merchants of Lübeck. It is possible that three centuries of complete security from foreign invasion made it unnecessary for the German aristocracy to learn the virtues of political self-discipline and subordination; but it would be a great mistake to judge Hohenstaufen Germany solely by its failure to achieve political and administrative unity. (K.J.L.)

The loss of effective central authority

II. Germany from 1250 to 1493

1250 TO 1378

The extinction of the Hohenstaufen dynasty and the rise of the Habsburgs and Luxembourgs. The death of Frederick II in 1250 and of his son Conrad IV in 1254 heralded the irreversible decline of Hohenstaufen power in Germany and in the conjoint kingdoms of Naples and Sicily. Conrad's infant son Conradin, heir to Naples and Sicily, remained in Germany under the guardianship of his Bavarian mother. His uncle Manfred seized the reins of government in both Italian kingdoms and in 1258 formally supplanted Conradin by engineering his own coronation in Palermo. Manfred's defiance of papal claims to suzerainty over the kingdoms impelled the French-born Pope Urban IV to grant them to Charles of Anjou, brother of Louis IX of France. Papal taxation of the French clergy and loans from Florentine bankers enabled Charles to raise a large mercenary army for an expedition to Italy. Manfred, deserted by his barons, was defeated and slain near Benevento in 1266. Conradin then rallied his German supporters and led them across the Alps. But Conradin's financial resources were inadequate; unpaid troops deserted, and his depleted following was routed by Charles near Tagliacozzo (1268). Conradin was captured as he fled toward Rome, convicted of lese majesty (a form of treason), and beheaded in the public square at Naples.

Decline of Hohenstaufen power

The Great Interregnum. In Germany, the death of Frederick II ushered in the Great Interregnum (1250–73), a period of internal confusion and political disorder. The anti-kings Henry Raspe (landgrave of Thuringia, 1246–47) and Count William of Holland (ruled 1247–56) were elected by the leading ecclesiastical princes at the behest of the papacy. William's title was recognized initially only in the lower Rhineland, but his marriage to Elizabeth of Brunswick in 1252 ensured his acceptance by the interrelated princely dynasties of north Germany. The death of the Hohenstaufen Conrad IV left William without a rival in Germany. His growing strength and independence enabled him to escape from the tutelage of his ecclesiastical electors and to devote himself to purely dynastic policies. He pursued his feud with Margaret, countess of Flanders, over their conflicting territorial claims in Zeeland at the mouth of the Rhine. He renewed the attempts of his dynasty to obtain complete mastery of the Zuider Zee by thrusting eastward at the expense of Friesland; he died at the hands of the Frisians in 1256.

Pope Alexander IV forbade the election of a Hohenstaufen but interfered no further with the succession. Hence the initiative was taken by a small group of influential German princes, lay and ecclesiastical, acting out of self-interest and personal advantage. None desired the election of a ruler powerful enough to threaten their growing independence as territorial princes; nor did they single out a German candidate, who might prove to be as uncontrollable as William. Archbishop Conrad of Cologne approached Richard, earl of Cornwall, brother of Henry III of England. Richard's bribes and assurances of future favour bought him the votes of the archbishops of Cologne and Mainz, the count Palatine of the Rhine, and Otakar II of Bohemia. He was formally elected in 1257 and crowned king at Aachen (Aix-la-Chapelle). Three months after Richard's election, Alfonso X of Castile, who aspired to the empire in order to strengthen his foothold in Italy, was chosen in equally corrupt fashion by

Initiative of the German princes

the Archbishop of Trier, the Duke of Saxony, the Margrave of Brandenburg, and the devious Otakar.

The candidates were distracted by the turbulence of the aristocracy in their respective countries—Richard paid four fleeting visits to Germany; Alfonso failed to appear at all. Both appealed to the papacy for confirmation of their election. Their claims were summarized in Urban IV's bull *Qui coelum* (1263), which assumed that the exclusive right of election lay with the seven leading princes involved in the double election of 1257.

The rise of the Habsburgs and Luxembourgs. *Rudolf of Habsburg.* When Richard died in 1272 the electoral princes were spurred into action by Pope Gregory X, who desired the election of a German monarch sympathetic toward a crusade for the recovery of the Holy Land. The princes, dreading an overly powerful king, rejected the advances of Philip III of France and Otakar. They chose instead Rudolf of Habsburg (1273), a minor count of Swabia who lacked the strength to regain the crown domains the electors had usurped during the Interregnum. Papal diplomacy persuaded Alfonso X to abandon his pretensions to the throne; but Otakar denounced the election on the ground that the Duke of Bavaria had voted as lay elector in his stead. Rudolf allied himself with the Wittelsbach family of Bavaria and with other envious neighbours of Otakar, who was defeated and slain (1278). The duchies of Austria and Styria, overrun by Otakar during the Interregnum, were declared vacant and conferred jointly on Rudolf's sons Albert and Rudolf (1282). These acquisitions placed the Habsburgs in the first rank of the German territorial princes and lent impetus to a gradual shift in the political centre of gravity from the Rhineland to east Germany. But the growing Habsburg power disquieted the electoral princes, who frustrated the King's attempts to secure the election of his elder son Albert in 1287 and of his younger son Rudolf in 1290.

Growing Habsburg power

Adolf of Nassau. On the death of Rudolf in 1291, the electors averted the danger of a hereditary Habsburg monarchy by choosing Count Adolf of Nassau as his successor. Adolf, possessing only a small patrimony to the south of the river Lahn, strengthened himself financially by promising military aid to and receiving subsidies from both sides in the current Anglo-French war. He took possession of Meissen when the cadet branch of the Wettin dynasty died out, and he used his foreign subsidies to purchase Thuringia in 1295. He was thus able to adopt a more independent attitude toward his electors. On June 23, 1298, five of the electors pronounced Adolf unfit to rule and deposed him; on the following day they elected Albert of Austria in his stead. Albert marched westward from Austria at the head of a large army, and in a battle at Göllheim, Adolf was slain and his supporters fled.

Albert I of Habsburg. By restoring the Habsburg Albert I (ruled 1298–1308) to the kingship, the electors placed themselves in jeopardy. The new ruler, backed by the ample resources of his Austrian dominions, was more powerful and unscrupulous than his predecessor. The electors regarded his treaty of friendship with Philip IV of France (1299) as a move to enlist French support for the election of his son Rudolf as his successor in Germany. His attempt to seize Holland and Zeeland as a vacant fief of the empire was rightly interpreted by the electors as an effort to establish Habsburg influence on the lower Rhine (1300). The four prince-electors of the Rhineland (the archbishops of Mainz, Trier, and Cologne and the count Palatine) conspired to depose Albert. But Albert wrecked the design by decisive military action (1301–02), and he sealed his victory over the electors by obtaining confirmation in 1303 of his election from Pope Boniface VIII in return for an unprecedented oath of fealty and obedience to the papacy. Albert subsequently renewed Adolf's claims to Meissen and Thuringia, but his authority there was still disputed when he died by assassination in 1308. Albert temporarily tamed the electoral princes, placated the papacy, and renounced intervention in Italy; but this policy foundered at his death, and the electors were given a fresh opportunity to reassert their influence over the German monarchy.

Oath of fealty and obedience to the papacy

Henry VII of Luxembourg. The princes, released from Albert's heavy hand, sought a servant, not a master. Archbishop Baldwin of Trier sponsored the candidature of his brother, Count Henry of Luxembourg, who was elected at Frankfurt (modern Frankfurt am Main) in 1308 as Henry VII. The house of Luxembourg (Luxemburg) was not a major territorial power, and Henry lost no time in exploiting his new status to extend its possessions. Under his direction the Diet of Frankfurt (1310) closed the long-disputed question of the Bohemian succession by awarding the kingdom, with the consent of the Bohemian estates, to Henry's son John. Thus, in common with the Habsburgs, the main weight of Luxembourg interests gravitated eastward. But Henry, unlike his Habsburg predecessors, dreamed of a restoration of the ancient authority of the empire in Italy. His Italian expedition (1310–13) opened brilliantly, and in 1312 he was crowned Holy Roman emperor at Rome. But the old fear of German domination stiffened the resistance of the Italian states. Pope Clement V was alarmed by Henry's preparations to invade the kingdom of Naples, a papal fief, and threatened excommunication. A renewed collision of empire and papacy seemed imminent when Henry died in 1313.

The growth of territorialism under the princes. The decline of Hohenstaufen influence in Germany, the Interregnum, and the rapid alternation of dynasties on the German throne created favourable conditions for the territorial princes, lay and spiritual, to gain power. Frederick II had purchased the support of the princes by lavish grants of crown lands, chiefly in the Rhineland and Thuringia; in 1220 he procured the cooperation of the ecclesiastical princes in the election of his son Henry as king and eventual heir to the empire by renouncing his regalian rights of building castles, issuing coinage, and imposing tolls on merchandise in their territories. Henry himself had extended similar concessions to the lay princes in 1231.

Thenceforth the direct action of royal authority was virtually precluded in the princely domains. The princes were at liberty to multiply castles and toll stations, establish mints, exploit mineral deposits, and settle all judicial cases except those transferred on appeal to the court of the emperor. But the machinery of administration under the prince and his council (*Hofrat*) was still rudimentary. Public taxation was intermittent and restricted to emergency occasions, and it was subject to the consent of the three estates of the principality (clergy, nobles, townspeople), which were consulted separately by the prince. The estates grasped the opportunity to ventilate their grievances and to press their advice upon the prince. The emerging territorial state was thus under the dual government of the prince and the estates, and its development was to be heavily influenced by a shifting balance of power between them.

Downfall of royal authority in princely domains

Constitutional conflicts in the 14th century. The death of Henry VII led to a disputed election and a civil war in Germany. The electors' impulse to choose another lesser count as king was checked by the houses of Habsburg and Luxembourg, which pressured the prince-electors to choose between their candidates. The pro-Habsburg majority elected Frederick the Handsome, duke of Austria. The minority withdrew their support from Henry VII's son John and transferred it to a more formidable candidate, the Bavarian duke Louis of Wittelsbach, who had recently broken an Austrian invasion of his duchy.

Electoral custom did not yet acknowledge the majority principle. The papacy, which had claimed the right to adjudicate disputed elections since 1201, was vacant. Hence the two claimants settled their differences by the sword. In 1322 Louis defeated and captured his rival at Mühldorf, but his triumph in Germany merely raised the curtain on a long and bitter dispute with the papacy.

Pope John XXII, guided by canon law and precedent, affirmed that Louis might not legally rule until confirmed by the papacy; thus the disputed election of 1314 and the absence of papal approbation invalidated Louis's royal title and his right to govern. Louis contended, however, that election by a majority conferred a legitimate title

and administrative power and did not require papal confirmation. His defiance of the Pope exposed him to excommunication (1324) and to the procedures of canon law, whereby he was required to submit entirely to the papal terms before absolution could be granted. Louis warned the electors that their rights were endangered by the subjection of the elections to papal confirmation. Six electors responded in the Declaration of Rhens (1338), proclaiming as an ancient custom of the empire that election by a majority was valid and that the king-elect assumed his administrative power immediately, without the intervention of papal approbation. Under Louis's direction the declaration was repeated at the subsequent Diet of Frankfurt as an imperial law, and offenders against it were declared guilty of lese majesty.

The Declaration of Rhens

John XXII and his successors were unyielding. In 1343 Pope Clement VI made diplomatic overtures to Charles of Luxembourg, heir to the Bohemian throne, with the object of procuring his election to the German kingship in Louis's stead. The electors, led by Baldwin of Luxembourg, archbishop of Trier, began to desert Louis one by one. The Pope thereupon urged a new election. Charles assured the Pope secretly that he would await papal confirmation of his forthcoming election before exercising governmental power in the Italian possessions of the empire, but, despite intense pressure by Clement, he would accept no such restriction with regard to Germany. In 1346 only two electors remained faithful to Louis: his son Louis of Brandenburg and his kinsman Rudolf, count Palatine of the Rhine. The other five assembled at Rhens on July 11 and elected Charles under the title of Charles IV. The new king was spared a lengthy conflict with his rival, who died of apoplexy in 1347.

Charles IV and the Golden Bull. Charles IV (ruled 1346–78) readily perceived that disputed elections exploding into civil war had been a standing malady of the German body politic since 1198 and that the stability of the German monarchy depended largely upon the degree of cooperation achieved with the territorial princes, more especially with the prince-electors. On his return from his imperial coronation as Holy Roman emperor (1355) he promulgated, with the consent of the German assembly of estates or diet (1356), a basic constitutional document—known as the Golden Bull from its pendant gold seal (*bulla*). Charles's double objective was to minimize areas of dispute in future elections and to strengthen his ties with the electors. Unanimity among the electoral princes had always been difficult to attain; hence the validity of election by majority vote, a principle already set forth in the Declaration of Rhens, was reaffirmed. The territories of the lay electors were declared indivisible and heritable only by the eldest son. Thus, partitions of land by family agreement and consequent uncertainty concerning the holder of the electoral vote were eliminated. In conformity with ancient custom the archbishop of Mainz was to convene the electors and to request them to name their favoured candidate. He was to announce his own choice after the other electors had given their vote verbally, and so he could cast the deciding vote in the event of a tie. The election was to be held in Frankfurt, the royal coronation in Aachen (Aix-la-Chapelle).

Provisions of the Golden Bull

The membership of the electoral body was fixed at the traditional number of seven: the archbishops of Mainz, Cologne, and Trier, the count Palatine of the Rhine, the king of Bohemia, the margrave of Brandenburg, and the duke of Saxony. When the throne was vacant the count Palatine would be regent in south Germany and the duke of Saxony in the north; thus the long-standing papal claim to govern the empire during a vacancy was tacitly rejected. The question of papal confirmation of elections was ignored; neither Charles nor his electors were prepared to yield, but an open affirmation of their position would have been ill-received by the papacy, which had played a leading role in Charles's election.

The Golden Bull consolidated and extended the territorial power of the electors. Their right to construct castles, issue coinage, and impose tolls was confirmed. They could judge without appeal. Conspiracy or rebellion against them was deemed high treason. They were to meet the ruler once yearly as supreme advisory council on affairs of state. The formation of city leagues against them was specifically prohibited. On the basis of these enactments the Golden Bull has been called the Magna Carta of German particularism. The electors in their capacity of territorial lords were its chief beneficiaries; the rest of the princes were envious and strove thenceforth to acquire an equally large measure of territorial sovereignty.

Rudolf IV of Austria ordered his chancery to fabricate a series of imperial charters, including two from Julius Caesar and Nero, as evidence of his virtual independence of the empire. Charles IV submitted them for examination to the Italian humanist Petrarch, who declared the charters to be spurious. Rudolf took up arms and was bought off by the recognition of his claim to Tirol (1364).

The election of Charles's son Wenceslas (Wenzel) as king in 1376 (two years before Charles's death) was a striking example of the Emperor's skill in securing the cooperation of the electors for his dynastic purposes. The election of an emperor's son as king of the Romans during the father's lifetime had not occurred since 1237; the prince-electors, in their anxiety to prevent any single dynasty from strengthening its grip on the succession, had checked all subsequent attempts. But unprecedented bribes, concessions, and a renewed prohibition of city leagues by Charles overcame the opposition of the electors. Pope Gregory XI had previously announced that the election would be invalid without papal confirmation. Charles, in concert with the electors, speeded the election and subsequent coronation of his son and then submitted an antedated request for confirmation to the Pope, who countered these devious tactics by delaying confirmation; it was still under consideration at Gregory's death in 1378. The decline of the papacy during the Great Schism (1378–1417) precluded the vigorous assertion of its right of confirmation, which became a mere formality and was subsequently tacitly abandoned.

Election of Wenceslas

Decline of the German monarchy. Charles IV's power was based primarily upon the territorial possessions of the house of Luxembourg, which he greatly extended by the purchase of the electorate of Brandenburg (1373). The German monarchy was a source of dignity and influence, but in terms of land and revenue it was outranked by Charles's hereditary domains in the east and northeast. The Golden Bull, replete with privileges to the electors, attacked none of the fundamental problems of the monarchy: dwindling crown lands, slender revenues, lack of an army and of an expert bureaucracy.

The financial problem was acute and of long standing. The succession of disputed elections between 1198 and 1257 had compelled the various claimants to purchase support by grants of royal land and revenues; the attempt by Rudolf of Habsburg to recover possession of crown lands alienated since 1245 had been opposed by his electors, who were unwilling to set an example by surrendering their own considerable acquisitions. At every election the votes of the princes had been bought by the grant or pledge of royal rights and property; thus, every king began his reign with a financial millstone round his neck and could attain freedom of action only by the possession or acquisition of extensive dynastic territories. The system of pledging crown lands involved the transference of the land and its revenues to the creditor. These revenues did not reduce the original debt, and the alienation tended to become permanent. The imperial cities (*Reichsstädte*) had been heavily taxed by Rudolf, and before his acquisition of Austria they had furnished the bulk of his revenue. His less provident successors had pledged them in a few cases to the local territorial princes and had thus lost the right of taxation. Charles IV carefully cultivated his dynastic revenues from Bohemia, but he lavishly expended crown assets in Germany to expand his family possessions. His financial exploitation of the cities for purely dynastic purposes naturally stiffened their resistance to taxation. By 1400 the annual revenues from all the German crown possessions averaged only 30,000 florins.

Financial problems

The enforcement of the public peace, a taproot of royal

power in other countries, also slipped gradually from the hands of the German monarchs. The German monarchy possessed no executive officials comparable with the English sheriff or justice of the peace, and it was diverted from its guardianship of law and order by recurrent conflicts with the papacy and by its absorption in purely dynastic matters. Consequently, the proclamation and enforcement of the peace fell into the hands of regional associations of cities and of the individual territorial princes. Thus the monarchy was prevented from using its function as defender of the public peace as an entering wedge to invade the jurisdiction of the municipalities and the territorial lords.

In sum, the German rulers were being gradually deprived of their triple role of feudal suzerains, defenders of the church, and keepers of the peace. The sweeping privileges granted to the princes in 1220 and 1231 had undermined their position as feudal suzerain. Their bitter struggles with the papacy cast doubt on their credibility as protectors of the church. They allowed their powers as guardians of the public peace to slip into the hands of others.

The continued ascendancy of the princes. By Charles IV's death in 1378, the division of Germany into loosely defined territorial principalities had reached an advanced stage.

Division into territorial principalities

Southern Germany. In south Germany, the dissolution of the Hohenstaufen duchy of Swabia gave territorial predominance to the Habsburgs, whose original possessions lay in Alsace, Breisgau, the Vorarlberg, and Tirol. Rudolf's acquisition of Austria and Styria (1282) had more than doubled the Habsburg patrimony and established its centre of gravity in eastern Germany. The Habsburg's rivals and neighbours to the north, the counts of Württemberg, had combined with the Swabian nobles to foil the attempt of Rudolf to revive the defunct duchy of Swabia for one of his sons (the counts, insatiably acquisitive and the inveterate enemies of the cities of the region, were finally raised to ducal status in 1495). The margraves of Baden were chiefly preoccupied with the southward expansion of their territory on the upper Rhine at the expense of the independent small nobles and cities of Swabia.

These three large entities contained lesser lordships, which were in constant danger of absorption by marriage, purchase, or feud. Bavaria, granted to the house of Wittelsbach as a duchy in 1180, was strengthened by the acquisition of the Palatinate in 1214; but subsequent testamentary partition restricted this important gain to the Upper Palatinate.

Central Germany. In central Germany, the margraves of Meissen of the Wettin dynasty thrust steadily eastward and received the electorate of Saxony in 1423, when the Ascanian line of electors died out; in the west they obtained Thuringia (1263) and clung to it tenaciously despite repeated royal attempts to oust them by claiming it as a vancant fief. The landgraves of Hesse, though surrounded by powerful neighbours, contrived to make modest territorial gains at the expense of the Wettin dynasty and the archbishops of Mainz. East and south of Hesse, the Rhine–Main region was a land of great ecclesiastical princes: the archbishops of Mainz, Trier, and Cologne; the bishops of Speyer, Worms, Würzburg, and Bamberg; and the wealthy abbots of Fulda and Lorsch. It abounded in counts of the second rank, dominated by a great secular prince, the count Palatine of the Rhine. The area contained four electorates and was therefore of crucial political importance.

Northern Germany. In north Germany, the dukes of Brunswick dissipated their strength by frequent divisions of their territory among heirs. Farther east the powerful duchy of Saxony was also split by partition between the Wittenberg and Lauenburg branches; the Wittenberg line was formally granted an electoral vote by the Golden Bull of 1356. The strength of the duchy lay in the military and commercial qualities of its predominantly free population. But the vigour of its eastward expansion into the Slav lands beyond the Elbe tended to diminish its involvement in the internal politics of the *Reich*.

Eastern Germany. In east Germany, the duchy of Mecklenburg, Germanized by a steady stream of immigrants, was drawn deeply into Scandinavian affairs and in 1363 provided Sweden with a new royal dynasty in the person of Albert of Mecklenburg. The electorate of Brandenburg, purchased by Charles IV and bequeathed to his second son, Sigismund, was dominated by a disorderly and rapacious nobility. Sigismund granted this dubious asset in 1415 to his faithful ally Frederick, burgrave of Nürnberg. The kingdom of Bohemia remained the durable territorial core of the Luxembourg dominions, and its silver mines at Kuttenberg, under German supervision, vastly increased crown revenues. The Slav population resented increasingly the economic and cultural influence of the German minority, and this created antagonisms profoundly disturbing to the monarchy.

Continued dispersement of territory. Inside the various territories the consolidation of the princely authority was far from complete. The principalities were often ragged in outline and territorially dispersed because of the accidents of inheritance, grant, partition, and conquest. Everywhere lesser nobles disputed the power of the prince and formed associations in defense of their rights and fiefs. In the ecclesiastical princedoms the ascendancy of an archbishop or a bishop was contested by the cathedral chapter, which had become a preserve of the nobility. The self-governing cities fought to protect their chartered liberties and drew together in formidable leagues to resist princely encroachments. Thus the princes, trying to enforce their authority, tended to consolidate the opposition and to excite potential or open hostility.

Opposition to princely power

In this crucial struggle the great secular potentates impaired their own strength by persisting in the Germanic custom of dividing their territory among their sons instead of transmitting it intact to the eldest. By 1378 the Bavarian lands of the house of Wittelsbach were shared between three grandsons of Louis IV. In 1379 the wide possessions of the Habsburgs were partitioned by family agreement between Albert III and his younger brother Leopold.

The ecclesiastical princes, dedicated to celibacy and elected by their cathedral chapters, could not hand on their lands to their descendants. But their policies and aspirations were no less worldly than those of the secular princes, and most of them contrived to install their relatives in rich canonries and prebends.

1378 TO 1493

Internal strife among cities and princes. The electors had voted for Wenceslas reluctantly during Charles IV's reign, fearful that the monarchy might become a perquisite of the house of Luxembourg. Most of the other princes shared their concern over the continued ascendancy of the dynasty.

Wenceslas. Wenceslas (ruled 1378–1400) inherited a variety of problems, which grew after his father's statesmanlike hand had been removed. Wenceslas' habitual indolence and drunkenness, vices that increased as he grew older, excited the pious indignation of his critics. His prolonged periods of residence in Bohemia betrayed his lack of interest in German affairs and allowed the continuous friction between princes, cities, and nobility to develop into open warfare.

Conflict between princes and cities

The collision of princes and cities was prompted by vital issues of long standing. The flight of the rural population from servile tenures on the land to the free air of the cities reduced the labour force and impaired the revenues of territorial lords. Others who stayed on the land accepted the protection and jurisdiction of the neighbouring city as "external" citizens (*Ausbürger, Pfahlbürger*) and thus withdrew themselves and their land from seignorial control. Only the most powerful cities (*e.g.*, Nürnberg, Rothenburg) were able to extend their extramural territory to a substantial degree by force. But all strove to expand the area of their jurisdiction at the expense of local lords, partly to prevent village industries from competing with the city guilds.

A second major issue was the insistence of territorial lords on imposing tolls on city merchandise in transit

through their possessions. In theory, tolls on road and river traffic were exacted in return for the protection of merchants and their goods. But the multiplication of toll stations without legal warrant hampered trade and provoked innumerable disputes, which often culminated in the seizure of merchants and merchandise by exigent lords.

The third and immediate cause of the crisis lay in the financial policy of Wenceslas himself. His Bohemian revenues, though large, were strained by the great sums payable to the electors in return for his elevation to the kingship. Hence he attempted to tap the resources of the imperial cities by demanding heavy taxes, and he threatened to mortgage recalcitrant cities to the neighbouring princes, their capital enemies.

On July 4, 1376, an alliance of 14 imperial cities of Swabia was formed under the leadership of Ulm and Constance for mutual protection against unjust taxes and alienation from the empire. The Swabian League counted 40 members by 1385 and was linked with similar coalitions in Alsace, the Rhineland, and Saxony. Wenceslas' initial hostility to the league faded as its membership increased, and in 1387 he gave it his verbal and unofficial recognition. He feared to offend the territorial princes by extending full recognition; further, a clause of the Golden Bull had declared all city leagues to be illegal. Thus he temporized and awaited the outcome of the approaching trial of strength between cities and princes. On August 28, 1388, the princes of Swabia and Franconia routed the largely mercenary forces of the Swabian League at Döffingen, near Stuttgart. The stipendiaries of the Rhenish League were put to flight by the count Palatine Rupert II near Worms on November 6.

The cities triumphantly withstood the ensuing siege operations, but their economy was injured by the forays, ambuscades, and blockades instituted by the princes. The protracted campaigns also exhausted the financial resources of the princes. When Wenceslas intervened in 1389, both parties were ready for peace. At the Diet of Eger (May 2) he ordered them to desist and declared the city leagues to be dissolved. The contestants complied. The princes were satisfied with the prospective disbandment of the cities, and the cities feared the consequences of further resistance. But Wenceslas' opportunism was relished by neither side; the princes disliked his political flirtation with the cities, and the cities resented his final championship of the cause of the princes.

Wenceslas' alienation of cities and princes

Wenceslas' early gestures of support for the cities rankled with the electors, who in 1384 and 1387 discussed the advisability of replacing him by an imperial vicar or regent. But Wenceslas learned of the plan and conveyed his opposition; nor could the electors unite on their choice of a regent. Some electors turned to a more drastic solution: his deposition. In 1394 Rupert II and the archbishop Frederick of Cologne considered the election of Richard II of England but failed to win the support of their electoral colleagues. In the following year, however, Wenceslas' elevation of Gian Galeazzo Visconti, imperial vicar of Milan, to the status of duke was assailed as a dismemberment of the empire and enabled the electors to pose as the indignant defenders of the integrity of the *Reich* against a wasteful and profligate king. Wenceslas attempted to conciliate the princes by appointing his younger brother Sigismund as German regent (1396). But the Milanese issue enabled Rupert and Frederick to enlist the support of the archbishops of Mainz and Trier for their proposed deposition of Wenceslas. The death of Rupert in 1398 occasioned some delay. But at length the electors compiled a lengthy series of charges against the King, and in September 1399 they openly proclaimed their intention of deposing him.

At this critical stage further proceedings were temporarily checked by serious differences concerning the choice of Wenceslas' successor. The favoured candidate of the Rhenish electors was the count Palatine Rupert III, himself an elector. But another elector, Duke Rudolf of Saxony, and a powerful group of north German princes contended that the electors could not raise one of their own members to the kingship. The Golden Bull had declared otherwise, but Rudolf held his ground and declined to participate in the subsequent proceedings. On June 4, 1400, the four Rhenish electors invited Wenceslas to Oberlahnstein to consider measures for the reform of the empire and threatened to release themselves from their oath of allegiance if he failed to appear. The King's efforts to rally support for his cause were utterly fruitless, and he decided to stay in Bohemia. On August 20 Archbishop John of Mainz, on behalf of the four electors, publicly proclaimed the deposition of Wenceslas as an unfit and useless king and freed his German subjects from their allegiance to him. On the following day the three archbishops elected Rupert in Wenceslas' stead. Rupert's consent to his election was presumed to furnish the necessary majority required by the Golden Bull.

Deposition of Wenceslas

There was no solid constitutional basis for the electors' claim that they possessed the right to depose their king: the dethronement of Frederick II in 1245 had been effected by the papacy. Their political purpose was to disentangle the monarchy from the grip of the house of Luxembourg and to repatriate it to the Rhineland. They succeeded because Wenceslas could exercise no leverage in German affairs after alienating both the princes and the cities.

Rupert. Rupert (ruled 1400–10) lacked the skill and resources necessary to revive the drooping power of the German monarchy. His title was not beyond dispute while Wenceslas lived, and the territorial princes and cities were therefore slow to acknowledge him. Pope Boniface IX, maintaining that only a pope might legally depose a German monarch, withheld his approbation of Rupert. An expedition against Wenceslas (1401) failed before the walls of Prague. Rupert then embarked upon an Italian expedition (1401–02), hoping to obtain the imperial crown from the pope and thus dispel the cloud of uncertainty that hung over his title. The enterprise was crippled by lack of financial means, Boniface's conditions were exorbitant, and Rupert returned to Germany without the coveted imperial coronation. Fortunately, he had little to fear from Wenceslas, who was fully occupied in protecting his Bohemian throne from the machinations of his ambitious younger brother Sigismund. Far more dangerous was the degeneration of Rupert's relations with the Rhenish electors. In 1405 he offended Archbishop John of Mainz by refusing him military aid in his war against Hesse and Brunswick. Consequently the Archbishop united all the enemies of Hesse and Brunswick in the League of Marbach, which included 18 imperial cities. Rupert contended that coalitions of cities were prohibited by the Golden Bull, and he denounced the league as illegal. The dispute was arrested by the mediation of the Archbishop of Cologne, but the memory rankled. Rupert's prospects darkened still further in 1408, when he lent his support to Pope Gregory XII against the cardinals who wished to summon a general council to end the Great Schism in the church. The archbishops of Mainz and Cologne and the vast majority of the German prelates favoured the conciliar solution and strongly approved the policy of the cardinals. Wenceslas shrewdly followed suit and in return received assurances from the cardinals that the future general council would recognize him as German king. The powerful proconciliar party in the German Church proceeded to agitate openly for the restoration of Wenceslas to the throne. But the threat of civil war was averted by Rupert's death on May 18, 1410.

Sigismund. On the death of Rupert the movement for the reinstatement of Wenceslas immediately lost headway. The Rhenish electors, having deposed Wenceslas ten years previously on ground of his unfitness, could not re-elect him without admitting their inconsistency. But the House of Luxembourg was powerful and would assuredly throw its full weight against any non-Luxemburg candidate to the German throne. The four electors agreed on the expediency of selecting Rupert's successor from the Luxembourg dynasty but disagreed on the choice of candidate. Rupert's successor, the count Palatine, and the archbishop of Trier elected Wenceslas' brilliant but unreliable brother, Sigismund, at Frankfurt on September 20, 1410. Eleven days later, the archbishops

of Cologne and Mainz elected Wenceslas' turbulent and treacherous cousin, Jost of Moravia. Jost died the next year, and Wenceslas agreed to accept Sigismund on condition that he himself retained the title of German king. But Sigismund ignored the reservation and assumed the disputed title. Wenceslas' protests were greeted with indifference in Germany and quickly died away. A second election of Sigismund at Frankfurt (July 21, 1411) gave him an ample majority and removed all doubt concerning the validity of the previous election.

Sigismund's character

Sigismund was energetic, versatile, and intelligent; but long experience never blunted his rashness in rushing into new projects, and his financial incapacity never ceased to astonish his contemporaries. His pursuit of personal power and dynastic possessions was unceasing and was conducted with complete unscrupulousness. His kingdom of Hungary and his later acquisition, Bohemia, were his primary concerns, and the interests of Germany were constantly set aside in their favour. The disastrous reigns of his predecessors, Wenceslas and Rupert, had emphasized Germany's basic problems: the weakness of the monarchy, the friction between princes and cities, and the unchecked growth of lawlessness and disorder. During his long reign (1410–37) Sigismund appeared less and less frequently in Germany and, as a largely absentee ruler, did little to correct these evils.

The Hussite controversy. Sigismund's prolonged absences were caused in great part by the explosive Hussite controversy in Bohemia. The Czech Church in Bohemia had long retained a marked individuality and much autonomy in its liturgy. This independent temper in ecclesiastical affairs was being slowly fused in the late 14th century with a rising sentiment of nationality among the Czechs. The upsurge of feeling took the negative form of a growing hatred of the German minority, which dominated the towns by virtue of its economic power and cultural influence. The luxury and immorality of the Bohemian clergy were castigated by a series of religious reformers such as Conrad of Waldhauser, Thomas of Štítný, John Milíč of Kroměříž (Kremsier), and Matthew of Janov. The teachings of Conrad and Milíč assumed a strongly puritanical tinge: in opposition to the wealthy sacramental church with its external means of grace they held up the ideal of the primitive church in a condition of apostolic poverty and the exclusive authority of the Bible as the foundation stone of faith and belief. These three movements met and intermingled in the person of Jan Hus.

Jan Hus. A graduate in divinity of Charles IV's foundation of the University of Prague, Hus was appointed incumbent of the Bethlehem Chapel in Prague (1402) and immediately attracted wide attention by his sermons, which were delivered in Czech in accordance with the foundation charter of the chapel. In 1403 he strongly defended a number of extracts drawn from the religious writings of the Englishman John Wycliffe. Czech opinion in the university solidly supported Hus. But the more numerous German masters carried the day, and the teaching of the controversial extracts was forbidden. Similarly, when Pope Gregory XII's cardinals rebelled against the Pope and demanded a general council to terminate the schism in the papacy (1408), the Czech members of the university aligned themselves with the cardinals, while the Germans stood with the Pope. On matters of general policy the masters of the university voted by "nations," of which there were four: Bohemian, Bavarian, Saxon, and Polish, the last consisting in fact largely of Germans. Thus the Germans controlled three votes, the Czechs only one. King Wenceslas reversed the proportion by decree, the German masters and students seceded to found their own university at Leipzig, and the mutual enmity deepened.

Hus's excommunication

In 1410 Hus was excommunicated by Archbishop Zbyněk of Prague but refused to appear at the papal court for judgment and continued to preach. Two years later he protested against the sale of indulgences, was placed under papal sentence of excommunication, and the city of Prague was subjected to a papal interdict. Hus left the capital at Wenceslas' request but preached throughout the land and vastly enlarged his following. From the lower Czech clergy who popularized Hus's doctrines, the masses learned that the German minority were intruders, foes of Bohemia and of the true religion. Lesser nobles who had lost their lands by mortgage or purchase to the prosperous German burghers of the cities were readily converted. The more self-interested members of the upper nobility were attracted by Hus's proposed reduction of the Czech Church to apostolic poverty, which would bring the rich territorial possessions of the higher clergy within their grasp.

The Council of Constance

The rising ferment in Bohemia disquieted the heir apparent, Sigismund, and he intervened with the suggestion that Hus should expound and justify his opinions to the Council of Constance (1414–18), recently convened to heal the schism in the papacy. Hus accepted, and Sigismund furnished him with a comprehensive safe conduct. The conciliar commission that examined Hus focussed the debate on two issues: the unauthorized Bohemian practice of extending communion in both kinds (bread and wine) to the laity, and the points of agreement between Wycliffe and Hus. Hus declined to retract his Wycliffite opinions until they were refuted by Holy Writ, and thus he defied the authority of the council in matters of doctrine. He was arrested on November 28, 1414, and died courageously at the stake on July 6, 1415. Sigismund's protests against the breach of his safe conduct were silenced by the argument that pacts with convicted heretics were not binding.

The Hussite wars. The death of Hus enshrined him at once as a martyr and a national hero in the memory of his followers. They raised a storm of denunciation against Sigismund and expressed their resentment by widespread attacks on orthodox priests and churches. The Catholics retaliated in kind, and Bohemia was in a state of civil war when the death of Wenceslas (August 16, 1419) brought Sigismund to the tottering throne. The new king's talent for conciliation and compromise was useless in the heated religious atmosphere. Pope Martin V urged him on against the Hussites and promised him imperial coronation as his reward. Under his prompting, Sigismund raised a motley host in Germany and launched it into Bohemia under the banner of a papal crusade (March 1, 1420). But the invaders were thrown back from the walls of Prague, and on July 7, 1421, Sigismund was declared deposed by the Bohemian assembly of estates. The shock of defeat forced Sigismund to attempt a fuller mobilization of German resources. Under the traditional system, princes and cities had been allowed to fix at their own discretion the quota of men provided by each when a royal campaign was in prospect. Naturally, both estates used their discretionary power to reduce contributions to a minimum. In 1422, however, Sigismund himself fixed the strength of the contingents demanded from the individual princes and cities throughout Germany. The response was disappointing. In 1426 the King raised his requirements, but to no effect. Hence the yearly campaigns against the Hussites were waged largely by mercenary armies. To meet the rising costs, the Diet of Frankfurt was persuaded in 1427 to vote a general tax, the so-called Common Penny. But there was little enthusiasm in Germany for the crusade, massive evasions of payment occurred, and the strength of local feeling hampered the coercion of defaulters.

Hussite raids into Germany

In 1429–30 the irrepressible Hussites swept through Saxony, Thuringia, and Franconia in a destructive foray. Sigismund, exploiting the general alarm, reverted to the older system and demanded contingents from each prince and city. The response improved, and a large army invaded Bohemia, only to meet complete disaster at Taus (Domažlice in modern Czechoslovakia) in 1431. It was evident that the veteran Hussites could not be crushed by force. Sigismund therefore welcomed the opportunity to transfer the problem of reconciling the Hussites with the church to the Council of Basel (1431–49). The Hussite extremists, the Taborites, were inflexible. They condemned the hierarchical system of church government and affirmed the priesthood of all true believers. Hence the council conducted its long and arduous negotiations

with the majority party among the Hussites, the Calixtines or Utraquists, who were prepared to accept the grant of communion in both kinds as a basis of settlement. The Utraquist nobles annihilated the protesting Taborites at the Battle of Lipany (May 30, 1434), made peace with the council by the Compact of Iglau (July 5, 1436), which conceded them communion in both kinds, and reunited with the Roman Catholic Church. The Utraquist nobles extracted far better terms from Sigismund as the price of their recognition. He agreed to accept the guidance of Czech councillors in governmental affairs, to admit only Czechs to public office, to grant an amnesty for all offenses committed since the death of Wenceslas, and to allow the Czechs a large measure of autonomy in their civil and religious life. It is unlikely that the slippery Sigismund intended to honour these pledges, but they cleared the way for his triumphant return to Prague in August 1436.

Effects of the wars on Germany

In Germany, the Hussite threat had clearly revealed the inadequacies of the existing financial and military systems. But the incentive to press Sigismund's reforms to a successful conclusion faded when the Hussite peril was scotched by negotiation. The general apathy was demonstrated in 1434, when Sigismund proposed to the princes a land peace embracing the whole of Germany. The abolition of private wars and feuds by such a peace was undeniably a paramount necessity. But the princes themselves were among the chief offenders against law and order, and their nominal approval of the plan deceived no one. Sigismund himself, increasingly absorbed in crucial negotiations with the Hussites, did not persevere, and the project gathered dust in the imperial archives. The impulse he gave to the cause of reform did not, however, fade entirely, though Sigismund did not live to see the sequel. His death on December 9, 1437, terminated the tenure of the German throne by the House of Luxembourg and opened the door of opportunity to the Habsburg dynasty.

The Habsburgs and the imperial office. *Albert II.* In the absence of a male heir, Sigismund had named his son-in-law Albert of Habsburg, duke of Austria, as his successor. Albert was able and vigorous, and the union of the territories of the two dynasties enabled him to exert considerable leverage in German politics. Albert declared his neutrality in the current dispute between Pope Eugenius IV and the Council of Basel on the subject of conciliar sovereignty and thereby evaded an issue on which the electors were strongly divided; thus, on March 18, 1438, he was unanimously elected at Frankfurt. The electors attempted to elicit from the new king an undertaking that he would grant privileges to his subjects only with their advice and consent. They also submitted a project for the division of Germany into four new administrative units (*Kreise*) in which the enforcement of the land peace would be entrusted to captains of princely rank. Albert judged that the princes were seeking to enlarge their power and influence under the guise of introducing reforms for the common good. The German cities also doubted the impartiality of the princes as custodians of law and order. Both proposals were therefore stillborn. The King hastened from Frankfurt to defend his kingdom of Hungary, endangered by Turkish raids on Siebenbürgen (Transylvania in modern Romania). The campaign was brought to a premature close by the death of the King on October 27, 1439.

Frederick III. Albert II had left only an infant son, and the leadership of the House of Habsburg passed to his cousin Frederick, duke of Styria. Inside the electoral college the Duke was vigorously supported by his brother-in-law Frederick of Saxony and was elected unanimously on February 2, 1440. The choice of Frederick tightened the hold of the Habsburgs on the German kingship. It also brought to the throne a ruler who, absorbed in dynastic concerns and in astrology, had no more than a passing interest in Germany. Under the absentee government of Frederick III, the feuds among the princes and the collisions between the princes and the cities developed into savage wars accompanied by widespread ravaging and pillage. All paid lip service to the need for peace; but who was to enforce it? Was it to be enforced by the monarchy, which lacked power and executive machinery? Was it to be enforced in the courts of the princes, whose judicial impartiality was suspect? Were complaints against the princes to be heard and decided in the king's court (*Hofgericht*)? Or must they be adjudicated by the council (*Hofrat*) of the prince concerned? The right to enforce peace effectively was a major source of power to the holder: hence the struggle between Frederick and the princes was long, bitter, and inconclusive.

The Turkish threat

These issues were brought to a head by the rapid westward progress of the Ottoman Turks after their victories at Varna on the Black Sea (1444) and at Kossovo in Serbia (1448). The Habsburg kingdom of Hungary and Frederick III's own duchy of Styria lay full in the path of the invaders. In 1453 the fall of Constantinople extinguished the Eastern Empire and aroused fears in Germany that the Western Empire would meet with the same fate. The King used the opportunity to demand financial aid against the Turks from the diet, the German assembly of estates. Under the leadership of the princes, the diet reminded him that Germany's capacity for defense was weakened by the current internal anarchy. In 1455, six electors proposed to the King the establishment of an imperial court of justice in which all three estates (electors, princes, and cities) should be represented. Frederick dismissed the scheme as an attempted invasion of his authority and stubbornly maintained his disapproval in a series of stormy interviews.

In time an increasing number of princes became convinced that reform would make no significant progress until Frederick was removed. As early as 1460 the Wittelsbach princes urged his deposition in favour of George of Poděbrady, the able and resourceful king of Bohemia. To check the danger, Frederick began to dole out reforms with a sparing hand. In 1464 he consented to make the court of the treasury (*Kammergericht*) independent of his person, to staff it with representatives of the three estates, and to extend its jurisdiction into fields other than financial. It was the acquisition of Austria in 1463 on the death of his brother Albert that finally proved his undoing. The unruly Austrian nobility early took the measure of Frederick and thereafter disregarded his authority. On the east and south the duchy was imminently threatened by the expanding kingdom of Hungary under its land-hungry ruler Matthias Corvinus. The southern borders of the Habsburg lands were also ravaged by the Turks. Frederick's continuing irresolution and passivity encouraged Matthias Corvinus, who had already seized a portion of Bohemia, to launch a campaign against Austria. The Austrian nobility made no move against him, and Vienna fell to him in 1485. Frederick fled to Germany and made pitiful appeals for help to the princes. His misfortune provided the party of reform with a long-awaited opportunity. Led by Berthold of Henneberg, the able and resolute archbishop of Mainz, they pressed the aging and afflicted Frederick to relinquish the kingship in favour of his son Maximilian. Solaced somewhat by the assurance of a Habsburg succession, he gave a reluctant acquiescence, and Maximilian was elected on February 16, 1486. Frederick retained the title of emperor, held since his imperial coronation at Rome in 1452. But he played no part in the government of Germany, and his death on August 19, 1493, passed almost unnoticed.

Loss of Austria

Developments in the individual states to c. 1500. *The princes and the Landstände.* In the various principalities the outcome of the struggle between the territorial princes and the assembly of estates (*Landstände*) was not fully decided by 1500. The vigour of the conflict arose partly out of the contrasting conceptions of government held by the protagonists. The secular princes looked upon their lands as private possessions that could be divided by agreement among their sons and drew little distinction between their private and their public revenues. The three estates regarded themselves as the corporate representatives of the whole territorial community and maintained that actions by a prince affecting their interests and privileges should be subject to their con-

sent. They therefore opposed the partition of the territory by family pact among the princes' sons. The inadvisability of breaking up the principalities into petty territorial lordships was at length conceded by the more prudent princes. By 1500 the rulers of Bavaria, Brandenburg, Saxony, and Württemberg had accepted the principles of territorial indivisibility and primogeniture.

Dispute over taxation

In financial matters the imposition of extraordinary taxes (*Notbeden*) remained the crucial issue between the princes and the estates. The mounting cost of war and administration outstripped the ordinary revenues of the ruler, plunged him deeply into debt, and compelled him to seek financial aid from the estates with increasing frequency. In the absence of a clear distinction between public and private revenue, the estates often contended that the deficit was a private debt of the prince and disclaimed responsibility. Needy princes were thus forced to buy temporary solvency by concessions that later shackled them in their dealings with the estates. The estates regarded the *Notbede* strictly as an occasional emergency tax and insisted that it should be reasonable in amount. Indeed, the estates of Bavaria and Brunswick extracted from their respective princes in the course of the 14th century a formal recognition of their right of armed resistance to extortionate taxation. Similarly, any prince who broke his agreements with the estates was subject to the right of resistance. Thus the aims of these territorial assemblies were in part negative. They sought to preserve the privileges of the three orders, to restrain the power of the prince, and to limit taxation. But they were also actively interested in good government, and the more enlightened rulers usually issued their ordinances only after consultation with the estates.

The princes proceeded against these powerful and often turbulent bodies with great caution. They persistently demanded that territorial assemblies convene only at the summons of the prince. They discountenanced the widespread conviction that absentees from the assembly were not bound to pay the taxes that it voted. In consequence, the peasants, who were not represented except in the Swiss cantons Baden, Friesland, and Tirol, remained in the grasp of the princes' tax collectors. In these directions the princes had generally made notable advances by 1500. But in the vital matter of the *Notbede* they were still obliged to bargain with the estates as equals. They had nowhere attained their ultimate objective: to transform the tax into a regular imposition voted automatically by the estates on demand.

Beyond the confines of the assembly of estates the attempts of the princes to curb their overmighty subjects aroused vigorous resistance. The noble vassals, proud and unruly, readily combined against any prince who sought to tamper with their liberties. Wise rulers deflected the nobles' energies into useful channels by employing them as stipendiaries. Hence even the most powerful princes—the Habsburgs in Austria and the Hohenzollerns in Brandenburg—proceeded circumspectly, and the difficult task of bringing the nobility to heel was far from completed in 1500. The cities of the princely territories defended their independence no less stubbornly. The princes revoked their charters, influenced municipal elections, and forbade the cities to associate in self-defense. The struggle was most intense in the north and east, where the Hohenzollern dynasty of Brandenburg emerged as the chief foe of municipal freedom. In 1442 the elector Frederick II ("Iron Tooth") crushed a federation of Brandenburg cities and deprived its leader, Berlin, of its most valued privileges. In the Franconian possessions of the dynasty, Albert Achilles of Hohenzollern waged a destructive war (1449–50) against a city league headed by Nürnberg. He suffered a resounding defeat in a pitched battle near Pillenreuth (1450). The elector John Cicero took up the battle 38 years later, when the cities of the Altmark in west Brandenburg refused to pay an excise tax on beer voted by the assembly of estates.

The "Beer War"

He discomfited the cities in the ensuing "Beer War" and radically revised their constitutions to his own advantage. On the other hand, the great cities of south Germany, enriched by the Italian trade, were more than a match for the local princes: the Wittelsbach dukes of Bavaria were decisively worsted by Regensburg in 1488.

The growth of central governments. Between 1300 and 1500 the organs of central government in the territorial states became more specialized and diversified. The parent body was the advisory council (*Hofrat*) of high nobles and ecclesiastics, whom the prince consulted at his discretion. Its business was not differentiated, and there was no division of labour among the councillors. It met at the summons of the prince and did not convene at regular intervals. Its membership was not fixed, and some advisers did not attend except at special invitation. Others were regional councillors who attended the prince only when he appeared in their locality. A body so unspecialized and fluctuating was ill-adapted to cope with the increasingly complex problems of central government. Hence in the 14th and 15th centuries a professional element of "daily" or permanent councillors was introduced. They were usually legists, trained in Italy or in the newly founded universities of Prague (1364), Vienna (1365), Heidelberg (1386), Rostock (1419), and Tübingen (1477). They were well versed in Roman law, which, with its centralizing and authoritative precepts, provided a congenial climate for the growth of the powers of the territorial princes everywhere save in Saxony and Schleswig-Holstein, where the ancient customary codes were deeply rooted. Financial administration, which required specialized skills, was placed under the direction of a separate department of government, the chamber (*Hofkammer*). An inner ring of favoured advisers, the privy council (*Geheimrat*), was also instituted to counsel the prince on affairs of state. The besetting weakness of the new administrative structure was financial. Few princes followed the example of the Hohenzollern dynasty in drawing up an annual budget and requiring financial officials to submit regular accounts to the government.

German society, economy, and culture in the 14th and 15th centuries. *Transformation of rural life.* Despite the impressive advance of trade and industry in the later Middle Ages, German society was still sustained chiefly by agriculture. Of an estimated population of 12,000,000 in 1500, only 1,500,000 resided in cities and towns. Agriculture exhibited strong regional differences in organization. The more recently settled areas of the north and northeast were characterized by great farms and extensive estates that produced a surplus of grain for export through the Baltic ports. The south and southwest was a region of denser population, thickly sown with small villages and the "dwarf" estates of the lesser nobility. In the northeast the great landlords, headed by the Knights of the Teutonic Order, tightened their control of the originally free tenants, denied them freedom of movement, and ultimately bound them to the soil as serfs. In the south the heavy urban demand for grain chiefly benefitted the larger peasant proprietors, who sold their surplus production in the nearest town and used their gains to acquire more land. The lesser peasantry, with their smaller holdings, practiced chiefly subsistence farming, produced no surplus, and therefore failed to benefit from the buoyant urban demand. The frequent division of the patrimony among heirs often reduced it to uneconomically small fragments and encouraged an exodus to the cities. On the other hand, landless day labourers who survived the Black Death in the mid-14th century were able to command higher wages for their services.

Serfdom in the north

In south Germany, the strain of transition in rural society was heightened by the policies of the landlords, lay and ecclesiastical. Confronted by labour shortages and rising costs, many landlords attempted to recoup themselves at the expense of their tenants. By means of ordinances passed in the manorial courts they denied to the peasantry their traditional right of access to commons, woods, and streams. Further, they revived their demands for the performance of obsolete labour services and enforced the collection of the extraordinary taxes on behalf of the prince. The peasants protested and appealed to custom, but their sole legal recourse was to the manorial court, where their objections were silenced or ignored. Ecclesiastical landlords were especially grasping,

Peasant discontent in south Germany

and peasant discontent assumed a strong anticlerical tinge and gave rise to the localized disturbances in Gotha (1391), Bregenz (1407), Rottweil (1420), and Worms (1421). Disturbances recurred with increasing frequency in the course of the 15th century on the upper Rhine, in Alsace, and in the Black Forest. In 1458 a cattle tax imposed by the archbishop of Salzburg kindled a peasant insurrection, which spread to Styria, Carinthia, and Carniola. In Alsace, the malcontents adopted as the symbol of revolt the *Bundschuh*, the wooden shoe usually worn by the peasants. They also formulated a series of specific demands, which included the abolition of the hated manorial courts and the reduction of feudal dues and public taxes to a trifling annual amount. On these fundamental points there was little room for compromise, and the outbreaks were stifled by the heavy hand of established authority. But the rigours of repression added fuel to peasant discontent, which finally burst forth in the great uprising of 1524–25 (see below).

The nobility. The lesser nobility included two distinct elements. The imperial knights (*Reichsritter*) held their estates as tenants in chief of the crown. The provincial nobility (*Landesadel*) had lost direct contact with the crown and were being compelled by degrees to acknowledge the suzerainty of the local prince. The imperial knights had been extensively employed by the Hohenstaufen emperors in military and administrative capacities and were chiefly concentrated in the Hohenstaufen possessions in Swabia, Franconia, Alsace, and the Rhineland. With the extinction of the Hohenstaufen dynasty they lost their function and rewards as a nobility of service. Their revenues from their small estates sank in purchasing power as prices rose. Caste prejudice prevented them from seeking an alternative role in trade or industry. Resentful of the decline in their fortunes and fiercely independent, they clung grimly to their remaining privileges: exemption from imperial taxes and the right to indulge in private war. They stubbornly resisted the persistent attempts of the princes to reduce them to subject status, and in Trier and Württemberg especially they were given valuable aid by the provincial nobles. For purposes of defense or aggression the imperial and provincial knights combined freely in powerful regional leagues, usually directed against the local princes or cities. In the course of their chronic feuds with the cities, many knights became mere highwaymen. Many others, who had been forced to sell their estates or who were encumbered with debts, took service in Germany or Italy as mercenaries (*Soldritter*). In east Germany the knights, though equally unruly, were far more affluent. The knightly estate (*Rittergut*) was larger and produced a profitable surplus for export. The knights sat in the assembly of estates, and taxation by the prince required their consent. They were therefore well entrenched against the encroachments of princely power.

Urban life. Urban society in 15th-century Germany was concentrated in some 3,000 cities and towns. About 2,800 of the total were extremely small, with populations varying from 100 to 1,000. Of the remainder, no more than 15 cities contained more than 10,000 inhabitants. In this restricted group three were pre-eminent. Cologne reached its peak in the 13th century with a population of 60,000, but sank to 40,000 by 1500 following internal disputes, expulsions, and steady emigration. In 1500 Augsburg was the most populous German city, with a resident population of 50,000. Third place was held by fast-growing Nürnberg, which counted 30,000 souls. The social unity of the citizen body had been most marked in the 13th century, when the guilds joined the dominant patrician families (*Geschlechter*) in wresting the right to form an independent city council (*Stadtrat*) from the lord of the city. In the 14th century the guild masters, methodically excluded from the council by the patrician oligarchy, broke into open revolt in Speyer (1327), Strassburg (1332), Nürnberg (1348), and elsewhere. In its economic aspect the ensuing conflict embodied an attempt by the guildsmen as industrial producers to free urban industry from the tight control exercised by the merchant patriciate. By 1500 the guilds almost everywhere had gained varying degrees of representation in the city council.

The guild system

In the meantime the guilds themselves had become increasingly oligarchical and exclusive as the established masters restricted the entry of new members in order to reduce competition. The ascent of journeymen and apprentices to the rank of master was obstructed by the imposition of excessive fees, and in many guilds membership became virtually hereditary. In consequence, the journeymen began to associate in fraternities of their own to press their demands for higher wages and a shorter working day. The masters denounced the fraternities as illegal, compiled blacklists of leading agitators, and formed intercity associations to enforce low wage rates. The day labourers and casual workers outside the guild structure had no protective organization and suffered heavily in periods of economic depression. The surviving tax records of the German cities, though not wholly reliable guides, nevertheless suggest wide extremes of wealth and poverty. In late 15th-century Augsburg, 2,985 of a total of 4,485 households (66 percent) were recorded on the tax rolls as exempt from taxation on the ground of insufficient means. At the other extreme stood the enterprising and prosperous business dynasties of the Fugger and the Welser. But not all wealthy citizens lacked public spirit: hospitals, almshouses, and charitable foundations multiplied within the city walls. The spreading evil of mendicancy was combatted by stringent legislation against able-bodied beggars in Esslingen (1384), Brunswick (1400), Vienna (1442), Cologne (1446), and Nürnberg (1447).

The decline of the church. The vigour and assertiveness of secular society in Germany was exercised increasingly at the expense of the clergy and the church. Among the upper clergy more than 100 archbishops, bishops, and abbots were temporal rulers. The prelates were usually sons of the nobility and did not allow election to church office to interfere with their aristocratic ardour for war and territorial acquisition. They were expert in the accumulation of benefices and were notoriously lax in the performance of their spiritual duties. Their influence was freely used to advance their kinsmen and partisans among the greater and lesser nobles, who poured into the cathedral chapters and ruled the abbeys. Thus the monasteries were filled with unspiritual persons, who were distinguishable from the lay aristocracy only by a nominal celibacy. Among the secular princes, the rulers of Austria, Brandenburg, and Saxony wrested from the papacy, gravely weakened by the conciliar movement of the 15th century, a right of appointment to a fixed number of bishoprics and lesser church offices. All the lay princes, in defiance of canon law, imposed their extraordinary taxes on the clergy. The steady invasion of the church by secular interests was also exemplified by the moral and material condition of the lower clergy. The Black Death of 1348–49 had decimated the ranks of the more dedicated priests, who ministered to their plague-stricken flocks instead of seeking safety by flight. The new recruits who rushed into holy orders were often self-seeking and spiritually unqualified. As the inflow continued, the problem of clerical unemployment and inadequate stipends attained greater proportions. Many were compelled by need to accept ill-paid livings. Others obtained no benefice at all and lived precariously as chantry priests or as itinerant chaplains. Their moral and intellectual defects were bitterly assailed by church reformers and by an increasingly well-informed laity. Pious Christians, especially in the cities, began to turn away from the priesthood in their search for spiritual comfort and to seek relief in mysticism or in lay associations practicing a simple, undogmatic form of Christianity.

Trade and industry. The most impressive achievements of the German economy between 1200 and 1500 lay in trade and industry. German trade benefitted from the Hundred Years' War between France and England, which diverted northbound Mediterranean merchandise from the customary Rhône valley route to the eastern Alpine passes; by the fierce internal warfare between the Italian city-states, which weakened their supremacy in

long-distance trade; and by the rapid economic development of "colonial" eastern Europe between the Baltic and the Danube. The north German trade was chiefly based on staple commodities such as grain, fish, salt, and metals. But the south German merchants, in their capacity as middlemen between Italy and the rest of Europe, had taken the lead by 1500. They combined trade and industry in the great Ravensburg Trading Company (1380–1530), which produced and exported Swabian linen and laid the foundation of the fortunes of the Höchstetter, Herwart, Adler, Tucher, and Imhof families. The most important independent concern was that of the Fugger, whose founder, Hans Fugger, began his career as a linen weaver in Augsburg. The Fuggers' accumulated profits provided capital for moneylending and banking, which they conducted with the aid of business techniques borrowed from the more advanced Italians.

The Ravensburg Trading Company

Cultural life. In the absence of a strong centralized monarchy to act as a focus, German culture continued to be regional in character and widely diffused. The mysticism of Meister Eckehart, Johann Tauler, and Heinrich Suso, which commanded all men to look for the kingdom of God within themselves, flourished chiefly in the cities of the Rhineland, where lack of diligent pastoral care forced Christians to call upon their own inner resources. In the same region social and moral satire attained an urgent and vivid realism. Sebastian Brant (1458–1521), born at Strassburg, spared no class in his epic on human stupidity, the *Narrenschiff*, or *Ship of Fools*. But it was in the thriving cities of south Germany, as yet little affected by Italian Humanism, that late Gothic culture reached magnificent heights in art, architecture, and sculpture. Albrecht Dürer, born in Nürnberg in 1471, challenged his generation with his evocative engraving of "Melencolia I" in which a brooding figure with closed wings sat idly amid a chaos of scientific instruments and meditated on the futility of human endeavour. In architecture, the hierarchical elaboration of the late Gothic style maintained its ascendancy and even made a notable conquest in Italy with the construction of the great cathedral of Milan, begun in 1387. The sculptured carvings of Tilman Riemenschneider (1468–1531) in the castle of Würzburg revealed the anxiety, the deep piety, and the religious sensibility of Christian men engaged upon a spiritual pilgrimage that was to continue to the Reformation and beyond. (C.C.B.)

III. Germany from 1493 to c. 1760

THE REFORMATION, TO 1555

Imperial politics

Maximilian I. Maximilian I, emperor from 1493 to 1519, the "last knight," has exercised German imagination as few other emperors have done. Dignified, affable, and showy, famous as a leader of the mercenary *Landsknechte* (infantry), as a hunter, and as a patron of the arts, he hatched fantastic plans to escape his unceasing financial embarrassment. In spite of all his failures, chance continued to deal him winning cards. His favourite project was to lead a European army against the Turks, but this could not be realized. Charles VIII of France invaded Italy in 1494; and this endangered not only the imperial fief of Milan but also the communications with Rome, which the Emperor valued, as he considered himself the protector of Christendom. Maximilian was not able to maintain his rights, despite the long series of wars and the continually changing alliances with the Italian states, the papacy, Switzerland, Spain, England, or France against whomever held Milan at the time. The German princes refused to follow Maximilian, as they feared that he would use them only to strengthen Habsburg interests. It is true that the estates in the Diet of Worms, guided by the archbishop of Mainz, Berthold von Henneberg, in 1495, granted a uniform and general tax, the "common penny," for the establishment of an imperial army and agreed to the erection of a supreme court to supervise the execution of the "permanent public peace" forbidding private feuds; but, in return, Maximilian had to agree to the setting up of a council of regency that was to supervise the Emperor. The estates, however, were neither organized nor resolute enough to impose their will for long on the Emperor, while the empire, without an imperial civil service, was too weak to secure its own revenue by taxation. The war with the Swiss Confederacy, which refused to tolerate Habsburg territorial possessions on its soil, virtually severed the connection between Switzerland and the empire in 1499, though Swiss independence was not formally recognized until the Peace of Westphalia in 1648.

Dynastic politics

In order to strengthen the material basis of his power, Maximilian was forced to expand the hereditary Habsburg territories. He never forgot this necessity, not even when he had to mortgage important sources of income, such as the Tirolese silver mines, to his creditors, the main creditor being the banking house of Fugger. By his new marriage to Mary of Burgundy (1477), daughter of Charles the Bold, he had brought to his house the possessions in the Netherlands. The marriages (1496 and 1497) of his son Philip I the Handsome to Joan the Mad and of his daughter Margaret to Joan's brother John, heir to the Spanish crowns (who died, however, within a few months of his marriage), brought about that linking of Spain, the Netherlands, and Austria that formed the basis for the worldwide empire of Charles V. Similarly, the betrothal of his grandson Ferdinand in 1516 to the Jagiello princess Anna created a claim on the succession not only to Hungary (where the rule of the Jagiello dynasty was weakened militarily and financially by the wars with the Turks and by the opposition of the Magyar nobles) but also to Bohemia, since her father was king of both. In return, however, Maximilian had to agree that the Knights of the Teutonic Order should do homage for Prussia to Anna's uncle Sigismund I the Old of Poland. Maximilian's arbitration between Bavaria-Munich and the Palatinate for the disputed succession to Bavaria-Landshut was paid for by his acquisition of the towns of Kitzbühel, Rattenberg, and Kufstein and the bailiwicks of Hagenau and Ortenau (1504). The territory around Lienz, which escheated to him after the death of the last count of Görz, linked Tirol and Austrian Swabia in the west with Austria proper in the east.

Beginning of the Reformation. Maximilian's reign saw the beginning of the Reformation. Criticism of the church had continued unabated since the great reforming councils of the 15th century. The states of western Europe had long ago made concordats with the Holy See permitting them to draw on the rich property of the church for government expenditures and forming, in fact, state churches largely independent of Rome. Similar moves on behalf of the *Reich* were unsuccessful as long as its rulers did not give up their pretension to the secular universal empire and, therefore, were not able to afford to renounce the power of the universal church. The only gainers were the territorial princes and the towns; they used the emergency powers of all secular authorities to reform the church in their territories but still allowed the papacy's complicated financial demands. Hence, from 1456, the imperial and territorial diets repeatedly formulated the "grievances (gravamina) of the German nation." They complained bitterly that the church was but an enormous financial institution that administered the means of salvation only with an eye to material profit. Complaints were also made against the privileged social and economic position of the clergy. In its higher ranks the church had become a welfare organization for the younger sons and daughters of the nobility. In some districts one-third of the soil was ecclesiastical property. Although the radical sermons of the mendicant friars glorified the ideal of a poor church, many prebendaries, monks, and nuns lived the pleasant life of drones, enjoyed special legal status, and were free of all civic burdens. This criticism found new food as the lives of high and low ecclesiastics became more secular and less edifying.

Luther's Theses

Martin Luther gave to this public criticism of the church a voice that could not be quieted and thus compelled a complete reconstruction of the church. The young monk and theologian, after many solitary struggles, had learned from his study of the Bible that men are justified not by the accumulation of pious works but by trusting in the mercy of God. His opposition to the traditional customs

of the church became public because of the abuses connected with the sale of indulgences. Albert of Brandenburg (1490–1545), son of the elector Johann Cicero, was a pluralist, holding the archbishopric of Magdeburg and Mainz and the administration of the bishopric of Halberstadt; this breach of canon law had to be paid for by increased dues to Rome. In order to reimburse the Fuggers and to obtain some additional income, he permitted a clever salesman, the Dominican Johann Tetzel, to sell indulgences in his dioceses by every dirty trick; the indulgence in question had been promulgated by Pope Julius II in 1506 on the occasion of the papal jubilee and had been renewed by Pope Leo X to obtain funds for the rebuilding of St. Peter's Cathedral in Rome. It is not quite certain that Luther, a professor of theology in Wittenberg, affixed his Ninety-five Theses to the door of the palace church on October 31, 1517. This was the usual way of inviting an academic discussion concerning a doctrine that was not yet a dogma of the church and that, according to Luther, endangered the sacrament of penance. Nevertheless within a few weeks the theses were printed and distributed all over Germany; against Luther's wish they were interpreted as an attack on the Roman Catholic Church. The papacy opened proceedings against him, but he was protected by his territorial prince, the elector Frederick III the Wise of Saxony, a key figure in the contemporary struggle for the imperial crown.

The election of Charles V. Maximilian's succession as an emperor was contested by his grandson Charles I of Spain, by Francis I of France, and for a short period by Henry VIII of England. Charles had been in Spain since 1517; after the death of his father, Philip the Handsome (1506), he united Burgundy with the Kingdom of Castile and its newly discovered lands in America; after the death in 1516 of his maternal grandfather, Ferdinand II of Aragon, Charles joined Aragon, Naples, and Sicily to his realms; and, after the death of Maximilian (1519), he and his brother Ferdinand I shared the Austrian hereditary lands together with claims on Bohemia and Hungary. Francis I of France, threatened by this extension of Habsburg power into the area of the Meuse and Scheldt and across the Pyrenees, did not intend to give up France's claim to lead Europe. Pope Leo X, as the temporal ruler of the papal states, wished neither for a renewal by the Habsburgs of the Hohenstaufen combination of the empire and Naples nor for the foundation by Francis I, who held Milan, of a French imperial line. He therefore advocated the candidacy of Luther's protector, Frederick the Wise, who was, however, too clearly aware of his own limitations and those of his power to agree to this. Both the French and the Habsburgs offered enormous bribes to the electors, but the contest was decided unanimously in favour of Charles, who became emperor as Charles V. German public opinion was also in his favour. The princes insisted on a series of solemn promises on the Emperor's part, contained in a "capitulation," both in order not to be drawn into the imminent war between him and France and in order to limit his rights by means of an oligarchic constitution for the *Reich*. If he broke the capitulation, the estates were to have the right to oppose him by force. For the remainder of his life Charles tried to fuse the dynastic and the imperial ideals, but he was pushing both ideas beyond their limit in the very period that saw the rise of the national states in western Europe.

The Diet and Edict of Worms (1521). In the meantime, the Lutheran affair assumed the proportions of a national movement. Luther became increasingly confirmed in his opinion that the Christian Church, according to Holy Writ, was not a visible external organization but a small group of people separated from the mass of nominal Christians by their belief in the divine revelation witnessed by the Bible. Luther felt called to a reform of the existing church, not to its complete reconstruction.

Support for Luther

The Humanists, fighting against the Schoolmen and the friars, believed him to be their partisan. The great Dutch Humanist Erasmus counselled moderation, but Ulrich von Hutten hoped to win both Luther and the young emperor over to his fight against Rome after the Swabian League had driven the quarrelsome Duke Ulrich from Württemberg and transferred the administration of Württemberg to Austria. Public opinion felt certain that the long-awaited reform of church and state would at last be realized. Luther, the most popular professor at the University of Wittenberg, where he allied himself with Philipp Melanchthon in an attempt to reform the curriculum, symbolized his formal separation from the church by burning a copy of the papal bull excommunicating him.

Charles came from Spain to be crowned at Aachen in October 1520 and to open his first Diet at Worms in 1521. He left Spain in a state of revolt. The regency, led by Cardinal Adrian of Utrecht (the future Pope Adrian VI), managed only slowly to gain the upper hand. Charles's aim was to renew the universal empire on the basis of the universal Habsburg power. At the Diet the estates obtained the agreement of the self-confident emperor to the establishment of a Reichsregiment, or Governing Council. The council, however, was to function only during his absence from the *Reich*, was to be led by an imperial viceroy, and was to be barred from dealing with foreign affairs. Since the capitulation that had secured Charles's election had stipulated that no subject of the empire could be declared an outlaw—a procedure that normally followed papal excommunication—without public trial, Luther's case had to be heard. When Luther steadfastly refused to recant his doctrines, the Edict of Worms was promulgated outlawing him and forbidding the reading and sale of his books. On his way home to Wittenberg, Luther, for his own protection, was secretly taken to the Wartburg Castle on the elector Frederick's order.

Immediately after the Diet of Worms, Charles made alliances against France, first with Pope Leo X, then with Henry VIII of England. He reconquered Milan and Genoa and defeated and captured Francis at Pavia in 1525. The Treaty of Madrid, which the imprisoned king signed to secure his release, did not last long, since the new pope, the vacillating Clement VII, formed the League of Cognac with Milan, Venice, and France in order to obtain small territorial gains. The sack of Rome by German and Spanish mercenaries in 1527 forced the Pope and France to make peace (treaties of Barcelona and Cambrai, 1529). Charles received the imperial crown at the Pope's hand in Bologna before he left again for Germany, the last elected king to be so crowned.

Lutheran church organization and the Peasants' Revolt (1524–25). Germany had seen far-reaching changes while Charles was absent. Luther's doctrines of the priesthood of all believers and of the Bible as the sole norm of life had shaken the bases of society as it had formerly been constituted. Wherever the authorities did not interfere, Evangelical congregations, independent in doctrine and discipline, sprang up throughout Germany. For, while Luther stayed in the Wartburg, his Wittenberg congregation listened to the "Zwickau prophets" led by Nickolaus Storch, who taught, as did Thomas Müntzer in Thuringia, that the true authority was not the Bible but the "inner light" given by God to the faithful. Other men, appealing to Luther's writings as their authority, used the general insecurity to further their own ends. The wealthy Franz von Sickingen, a friend of Hutten and, like all knights, hemmed in between the rising cities and the territorial princes, hoped that the impending changes would bring him a principality between the Nahe River and Alsace. He made war against the Archbishop of Trier but was himself besieged and killed in his castle of Landstuhl with the assistance of Philip the Magnanimous of Hesse. The Swabian League razed the castles of the knights allied with him throughout Franconia and the Odenwald (1523). The greatest upheaval was caused by the Peasants' Revolt that began in 1524 in the southern area of the Black Forest and spread in 1525 through southern Germany (except Bavaria), Hesse, Thuringia, Saxony, and Tirol. Citing Luther's plea for the "liberty of Christian men," the peasants demanded the restoration of their customary rights and destroyed abbeys and manor houses. Luther, however, attacked them in passionately worded pamphlets. The individual peasant bands, badly

The Knights' War

led, were easily defeated, with enormous loss of life by the armies of the territorial princes.

The experience of the Peasants' Revolt proved that the Reformation could not advance by itself outside the cities. Ecclesiastical visitations of rural parishes showed that the common man was rude and hostile to all religiosity, while the clergy were ignorant, negligent wasters of church property. With Luther's encouragement each territory now established its own state church; they all differed according to the character of the prince concerned and according to the changing political circumstances. Unsuitable priests were dismissed; inventories of ecclesiastical property were drawn up; consistories supervised the clergy, with the prince as *summus episcopus* ("supreme bishop") and judge of appeal; the monasteries were made to bear the cost of schools and churches; and provision was made for education, for church discipline, and for a uniform development of dogma. As none of the rulers wanted to give up uniformity of public service or of dogma, their subjects who differed from them on grounds of conscience had to emigrate.

The Diets of Speyer (1526 and 1529) and Augsburg (1530). During the long absence of the Emperor, the estates tried to find a political solution to the religious differences, but Charles refused to call a German national council. The governors of both the Reformed and Roman Catholic territories were convinced that in one territory there should be only a unanimous faith; otherwise new revolts would arise. Some of the Roman Catholic territories, united in the Regensburg Convention (1524) and in the Dessau League, promised one another mutual succour in the execution of the Edict of Worms, but this provoked the formation of the Torgau League of Evangelicals. At the Diet of Speyer of 1526, Luther's followers were willing to uphold the union with Rome if they were permitted to treat institutions and ceremonies based on the Bible as essential but to regard those that were manmade as not essential and only to be tolerated until the calling of a general council. Charles objected again, for he wished all ecclesiastical changes only to be settled by a council called by both the Emperor and the Pope. He needed them both for the realization of his idea—a universal Christian empire. As these two powers were actually in a state of war, the estates bound themselves to follow only their conscience until the next council.

The Zwinglian movements

Besides the state churches based on Lutheran doctrines, there arose another type of reformed organization, the Zwinglian, in which the civil community was itself identified as the legal embodiment of the church and in which the citizen was equated with the Christian. Huldrych Zwingli had abolished the mass and removed the sacred images in the Swiss city of Zürich. Zürich thus had to face the hostility of neighbours who had remained Roman Catholic and, for economic reasons, were not prepared to follow Zürich's example in cancelling their treaty for the provision of Swiss mercenaries to France. In Switzerland also there arose the radical Anabaptists, whose congregations also rapidly spread throughout southern Germany, Thuringia, Hesse, Silesia, and Moravia, into the Alpine valleys and down the Rhine into the Netherlands; everywhere they were cruelly persecuted by both Roman Catholic and Reformed authorities; a general war of religion was avoided only with difficulty. The Diet of Speyer of 1529 failed to produce any recess to which all the estates could agree: the Roman Catholic majority resolved to leave the solution of the religious quarrel to a future council, to oppose Zwingli's doctrines, and to keep the Edict of Worms; but the "protesting" minority appealed to the Emperor and to the council and, in the meantime, did not wish to do anything against their conscience. Philip the Magnanimous, landgrave of Hesse, invited Luther and Zwingli to Marburg, on the river Lahn, in the hope of reconciling their differences about the Eucharist so as to make possible a political alliance with the Swiss, but the attempt failed. Lutherans and Zwinglians presented to the Emperor different formularies of faith at the Diet of Augsburg (1530). Though drawn up in a conciliatory spirit, they did not lead to any compromise, since the Evangelicals remained firm on all questions of conscience. Yet the calling of a council could not be expected from the contemporary papacy, which could only foresee a diminution of its power as the likely result. Thus failed the attempts to reform the church in the empire. Those who wanted to recover the lost areas for Roman Catholicism had to use force. To be prepared against force, Philip of Hesse and the elector of Saxony, John the Steadfast, formed, in 1531, the Schmalkaldic League, consisting of the north German Protestant princes, Strassburg, and a number of south German cities. It proved impossible to form a counterleague, as the political interests of the Roman Catholic princes concerned were too divergent.

Protestant and imperial politics after Augsburg. Ferdinand, Charles V's brother, had been made his successor in the Austrian hereditary lands in 1522 and had become king of part of Hungary in 1526, after the death of his brother-in-law King Louis II in the Battle of Mohács against the Turks in 1529; and in 1526 a quarrel about the succession to Louis II's other kingdom, Bohemia, had ended with Ferdinand's election as king against his rivals, the Bavarian dukes. In 1531 Charles saw to it that Ferdinand was elected king of the Romans, or successor designate to the empire, despite strong Roman Catholic and Lutheran opposition. In the same year, the south German cities lost their Swiss backing when Zwingli was killed in the Battle of Kappel against the Roman Catholic cantons. When the Turks invaded Hungary again in 1532, Ferdinand was forced to buy the indispensable Protestant support by the Religious Peace of Nürnberg; a truce was called in doctrinal matters until the meeting of a council (now very far distant) or the next Reichstag. Philip of Hesse succeeded, with the help of French subsidies, in restoring Duke Ulrich to Württemberg, where he immediately introduced the Reformation (1534). Such successes gained new members for the Schmalkaldic League. The Reformation, in the form of strictly regulated state churches, spread along both shores of the Baltic, into Silesia, and to the lower Rhine.

Charles against the Turks

Charles, full of the dream of a universal empire, held it his highest duty to fight infidels and heretics. He was victorious against the Barbary corsairs of North Africa who had been harassing the coastal cities of Spain and ruining Spanish commerce in the Mediterranean, but this led to a renewal of hostilities with France. After futile campaigns in Provence and in the Netherlands, the ten-year Truce of Nice was mediated by Pope Paul III (1538). Paul, however, refused Charles's demand for a council. Consequently, Charles extended the religious truce to all the new adherents of the Augsburg Confession (Frankfurt Agreement, 1539), for he needed the aid of the Protestants against the Turks and now, moreover, had to compete not only with France but also with William, duke of Cleves-Jülich-Berg, heir to the rich duchy of Gelre. Religious colloquies at Hagenau, Worms, and Regensburg came to nothing. King Ferdinand had to watch impotently while the Ottoman sultan Süleyman I occupied Hungary.

In 1542 Francis I of France felt strong enough to make his fourth war against Charles V. The Schmalkaldic League remained inactive while Charles occupied Cleves and reintroduced Roman Catholicism; but Philip of Hesse and the elector, John Frederick I of Saxony, together expelled the Emperor's firmest partisan, Henry II of Brunswick-Wolfenbüttel, from his duchy and proceeded to evangelize it. At the same time Herman of Wied, archbishop-elector of Cologne, attempted to introduce Protestantism in his diocese with the advice of Melanchthon and the Strassburg reformer Martin Bucer. The Protestant faith was openly accepted in the bishoprics of Münster, Osnabrück, Paderborn, and Minden and was professed by Maurice, of the then ducal line of Saxony, and by Otto Heinrich of the Palatinate Neuburg.

The Schmalkaldic War. The Peace of Crépy (1544) between Charles V and Francis of France created the preconditions for a thorough overhaul of German affairs. While the Protestants were fooled by renewed religious talks at Regensburg, Charles got ready for war. The papacy had called a general Council to Trent for 1545

but hoped to escape this irksome promise by actively supporting the Emperor with money and troops. Charles obtained the neutrality of Bavaria and of Duke Maurice by vague promises. There was no doubt about the Emperor's purpose, yet the Schmalkaldic League proved unable to reform its cumbersome organization. The Emperor took the field against the league in the summer of 1546. Hesse and electoral Saxony could not bring themselves to risk the effective army they and their allies had quickly assembled; consequently, the Emperor was able to reinforce himself from Italy and the Netherlands. Impatient, Duke Maurice brought the decision when he, together with King Ferdinand, invaded unprotected Saxony in order not to lose the electoral dignity that had been promised to him. John Frederick fled but was caught near Mühlberg on April 24, 1547, and Philip of Hesse was made prisoner in July. Charles V had reached the zenith of his power, ready, with the help of the council, to lead back the strayed sheep into the fold of the church.

Charles V at the height of his success

The Augsburg Interim. The papacy, however, was not willing to see the Emperor all-powerful. Charles V had asked the Pope to consider and redress first the "grievances," but, when the long-awaited council had at last met, it had begun by redefining the creed and the apostolic traditions of the church. Indeed, Paul III had withdrawn his troops even before the Battle of Mühlberg and transferred the Council to Bologna in order to withdraw it from imperial influence. Thus, it was without papal backing that Charles V forced the Interim on the estates at the Diet of Augsburg (1547–48). This conciliatory formula restored Roman Catholic ritual in general but conceded the eucharistic cup to the laity and allowed priests who were already married to keep their wives pending the final decision of the general council. The Roman Catholic princes refused to accept this before the council had actually decided, and Roman Catholic priests were reintroduced by force in the Protestant areas and towns of southern Germany; but the common people either remained faithful to the expelled ministers (unless they went into exile) or renounced sermons and sacraments altogether where the authorities did not find a means of circumventing the Interim.

Pope Julius III recalled the Council to Trent in 1551, but the Protestant envoys, now admitted to its sessions, had to limit themselves to the mere presentation of their confessions of faith and to protests against decrees that could no longer be revised. Deeply disappointed, the Emperor had to recognize that the general council was unable to provide a viable solution for the religious conflicts in Germany.

Maurice of Saxony's war (1552). All the estates agreed in opposing the imperial absolutism, the "beastly Spanish servitude," as long as the Emperor, contrary to his election capitulation, kept Spanish troops in the *Reich*. The soul of the resistance movement was the ambitious Maurice of Saxony, whose Protestant faith was chiefly a means of affirming his princely independence and rise to power. His Treaty of Chambord with Henry II of France (1552) provided him with large subsidies for his war against the Emperor in exchange for the cession to France of the imperial cities of Metz, Toul, Verdun, and Cambrai. Taken by surprise, the Emperor fled from Innsbruck to Villach, pursued by the troops of Saxony, Hesse, and Brandenburg-Kulmbach; the council of Trent disintegrated out of fear of the approaching army. Yet, in the negotiations at Passau, the opposition achieved only a renewed truce until the calling of the next Diet, which was to decide whether the religious conflict was to be ended by decision of the Council, by a recess of the Diet, or by a religious colloquy. Maurice was killed in the Battle of Sievershausen in July 1553, against Albert II Alcibiades of Brandenburg–Kulmbach, who had continued the war on France's order, burning and killing as he pleased.

Charles V's abdication and the Peace of Augsburg (1555). Charles besieged Metz unsuccessfully, but his hopes for a universal empire were revived in 1554 by the marriage of his son Philip with the English queen Mary I; by combining the resources of the Netherlands, England, Spain, and the *Reich*, Charles hoped to encircle and defeat France. The marriage, however, remained childless, so that the Emperor's expectations came to nothing. When Henry II of France found a supporter in the new pope, Paul IV (elected 1555), Charles abdicated, worn out by all his failures. He handed over the Netherlands to his son Philip in 1555, Spain to Philip in 1556, and his imperial authority to his brother Ferdinand in 1556. He then withdrew to Estremadura, near the monastery of San Yuste. Thus, he admited that the medieval ideal of the unity of Christendom was no longer valid.

Division of the Habsburg dominions

In 1555, despite Charles's protests, the Diet of Augsburg sanctioned the existing state of affairs. The Peace of Augsburg renounced the idea of uniformity of doctrine within the empire, acknowledged the coexistence of Roman Catholicism and Lutheranism, and promised no toleration for Zwinglians, Calvinists, or Anabaptists. Lawyers later elaborated the formula *cujus regio ejus religio* ("the prince's religion is the religion of his dominions"), to be applied to the individual territories. The so-called ecclesiastical reservation stipulated that, if an ecclesiastical prince became Lutheran, he had to renounce his office; thus his change of religion would not affect his subjects as a secular prince's change would. Church property secularized before 1552 was to remain so. The cities that had accepted the Interim had to tolerate both Roman Catholicism and Lutheranism within their boundaries. The Peace of Augsburg was a deep disappointment to the high hopes entertained at the beginning of the Reformation. Its authors held it to be only a temporary solution, but nobody could really believe that a final religious reconciliation would be brought about in the future. The idea of one Christian Church had to be given up now that two creeds were legally established side-by-side.

The recess of the Diet of 1555 also decided the conflict between emperor and estates in favour of the estates. By its "executive ordinance" the "imperial circles" (administrative districts established by Maximilian) were given powers to administer law and to maintain order within their areas and to execute the decisions of the Imperial Chamber, or Reichskammergericht. In the southern and western areas, where there were the greatest number of quasi-independent authorities, this "circle" organization provided at least some safeguard against the religious wars threatening to spill over into Germany from the west. The vital forces making for a renewal of political life were now to be found in the individual territories of the empire rather than in the empire as a whole. They were not active in foreign affairs. Within the territories, however, newly created administrative organizations, centrally directed and staffed with trained lawyers, began to issue numerous laws covering every aspect of life. The subjects were treated as uniformly as possible within the boundaries of each state, and the habit of obedience to orders was instilled; the territorial estates, composed of nobility and towns, gradually saw their rights severely curtailed, especially that of deciding taxes. As regards ecclesiastical policy, the territorial princes, "viceregents of God," felt responsible for maintaining uniformity of belief: whatever their momentary belief might be, it was to be the only truth recognized, upheld, and propagated in their states. Regarded by contemporary opinion as an essential safeguard against heresy and revolution, the doctrinal unity thus enforced was used as an instrument of state policy to strengthen the unity of the territory and so to augment its power. The religious struggles in Germany came to be settled on the territorial plane, and the resultant territorial distribution of the various creeds survived, by and large, into the 20th century.

Powers of the "imperial circles"

The independence of the numerous though varied states was hardly called into question any more, so that emperor and empire faded into the background, having shed their universal claims. The emperors had to cooperate with independent institutions of the *Reich* and to reach an agreement with the diets and the representatives of imperial circles. The emperors' power was not enough to stop the centrifugal forces nor to inaugurate a development toward the formation of a national state.

The emperor and the *Reich*

THE COUNTER-REFORMATION AND THE THIRTY YEARS' WAR

Ferdinand I and Maximilian II. The partition of Charles V's dominions meant that henceforth there were two Habsburg lines—the Austrian and the Spanish—but both were still Roman Catholic, and they were closely linked with one another despite occasional tension and opposition. Ferdinand I was no less devoutly Roman Catholic than Charles V had been, but reasons of state found him readier to make necessary concessions. His son and successor, Maximilian II, emperor from 1564 to 1576, was still readier for compromise as he leaned toward Protestantism, though he never confessed it publicly. During their reigns, however, the leadership in the world of ideas, which Luther and the proponents of universal empire had combined to give to Germany, passed to the spiritual forces, which took advantage of the peace that now prevailed in the *Reich*.

The aims of the Counter-Reformation

The Roman Catholic Counter-Reformation, inaugurated by the Council of Trent and promoted by Spain, Rome, and Italy, put forward the idea of the hierarchical structure of the church, led by the papacy, against the Protestant doctrine of the priesthood of all believers. The reformed papacy, moreover, possessed in the Society of Jesus its most reliable army, which turned its attention immediately to the most threatened spot, Germany; Jesuit settlements had been founded as early as 1544 in Cologne, Vienna, Ingolstadt, and Prague—that is, in areas where the religious decision was still in the balance. Moreover, the relationship between the Habsburgs and the Bavarian dukes created a Roman Catholic bloc in southern Germany with its political and intellectual centre in Munich. Though support from the rest of Germany was always most readily forthcoming for the defense of the empire against the Turks, who bore most heavily on Austria, Ferdinand and Maximilian, nevertheless, avoided offensive war in order to prevent further religious concessions to the Protestant Austrian estates.

The Lutheran princes east of the Weser River, meanwhile, received effective help from Denmark. Electoral Saxony, Brandenburg, and Pomerania not only secularized ecclesiastical property and established Lutheran state churches within their territories but also began, from 1555, to absorb neighbouring bishoprics and abbeys. Despite endless disputes, which paralyzed the diets, this proved a successful method of circumventing the ecclesiastical reservation that had been meant to protect such ecclesiastical principalities. Disunity, however, grew among the German Protestants when French and later Dutch refugees brought Calvinism first to the Palatinate (1559) and then to Nassau, Hesse, and the Lower Rhine Valley. Soon after Luther's death (1546), doctrinal strife had broken out between the followers of the old radical Lutheranism and those who preferred Melanchthon's compromise formulas. As the princes could decide what form of religion they wished to establish, this theological quarrel had political consequences. Electoral Saxony, Württemberg, and Brunswick-Wolfenbüttel were eventually to end the constant doctrinal disputes among the Lutherans by accepting, after a long struggle, the *Formula concordiae* of 1577, after Rudolf II's accession as emperor. This did not prevent them from regarding the Calvinists as more serious opponents than the Roman Catholics. When the Calvinist John Casimir of the Palatinate tried to form a Protestant league, Saxony wrecked the plan. John Casimir was the foremost of the German princes to intervene on the Hugenot side in the wars of religion in France.

Calvinism in Germany

Wherever Austrian power proved ineffective, the empire suffered great losses. As the Teutonic Order had been secularized as early as 1525, its calls for help were ignored, so that Courland became a Polish fief in 1562, while Livonia was divided between Poland and Russia, and Estonia passed under Swedish rule. On the western frontier the French kept the cities of Metz, Toul, and Verdun and extended their influence over Lorraine. When the revolt of the Netherlands broke out against Spanish rule, the empire did not intervene, though the Netherlands belonged to the Burgundian circle of the empire and though the leader of the revolt, William I the Silent, was a prince of the empire (Nassau and Orange). Only the Calvinist Palatinate gave William support; otherwise, the Protestant estates merely appealed for Maximilian II's mediation, which Spain rejected.

Rudolf II, the Cologne War, and Matthias. Maximilian was succeeded by his eldest son, Rudolf II, emperor from 1576 to 1612. By this time the conflict in the Netherlands was having its effect on the strife of the creeds in the Rhineland. The fluctuating religious position there was decided by events in the archbishopric of Cologne (Köln), where Archbishop Gebhard wanted to marry and so, in order to circumvent the ecclesiastical reservation, tried to secularize the diocese (1582–83). Cologne's conversion to Protestantism would have given the Protestants a two-thirds majority in the electoral college, with far-reaching repercussions on the election of an emperor. Ernst of Bavaria, bishop of Freising, Hildesheim, and Liège, was consequently put forward by Bavaria, Spain, and the papacy to take Gebhard's place. Gebhard put his troops under the leadership of John Casimir of the Palatinate but received no help from the Lutheran princes. Defeated in 1584, he withdrew to the Netherlands and then to Strassburg. The Spaniards and the Dutch, however, continued the war for years, devastating the country; Münster and the other Westphalian bishoprics fell to Ernst of Bavaria, along with Cologne.

Rudolf II had been educated in Spain; he was devoutly Roman Catholic, interested in intellectual matters, of good political judgment, but of an odd nature. He fled from human contact, was afraid to act, and lived alone in his castle in Prague until his reason gave way and he became incapable of making political decisions. The violent means by which Catholicism was being restored in all the Austrian territories and in Hungary provoked the opposition of the Protestant territorial estates everywhere and fanned the discord between Rudolf and his brother Matthias, who in 1606 was declared the head of the House of Habsburg in Rudolf's place. It seemed as if the Habsburg dominion would disintegrate into separate small territories, each controlled by its estates, thus paralyzing Austria's capacity for action. As Rudolf, the wearer of the imperial crown, was thus incapacitated, imperial institutions slowly ceased to function.

The absolute paralysis of Emperor and empire revived the old plan to form religious associations. Two such groupings soon opposed one another: the Protestant Union, led by the Palatinate (1608); and the Catholic League, led by Maximilian I of Bavaria (1609). The former looked for help to France, the latter to Spain; and both armed themselves, as France and Spain seemed about to begin a European war over the succession to Cleves-Jülich-Berg and its undecided religious allegiance. The assassination of Henry IV of France (1610) prevented the outbreak of war, and Matthias (emperor from 1612 to 1619) tried to ease the tension by his policy of "compositions"—that is, of small concessions. Finally, in 1614, the duchy of Cleves-Jülich-Berg was peacefully partitioned, without imperial mediation, by the late duke's heirs: John Sigismund of Brandenburg, who had just turned Calvinist, and Wolfgang William of Palatinate Neuburg, who had just turned Catholic.

The Catholic League and Protestant Union

Ferdinand II and Bohemia. Since neither Matthias nor his surviving brothers had legitimate heirs, it was planned that the Habsburg and the imperial succession should pass to their cousin Ferdinand of Styria, a pupil of the Jesuits and a convinced champion of the Counter-Reformation who had just re-Catholicized inner Austria by force. Philip III of Spain renounced his own claims and offered help to Ferdinand in exchange for the cession of the Austrian territories in Alsace and Ortenau, which would link the Spanish Netherlands with the Spanish Franche-Comté and the routes to Italy and thus extend the arc of encirclement around France. The Protestant Bohemian estates, however, objected strongly to recognizing Ferdinand as their future king, and in 1618, during a dispute over local grievances, some radical nobles in Prague threw the imperial governors out of the windows of the Hradčany Castle. The Bohemians then pre-

pared for war, and in 1619, after the death of Matthias, proceeded to elect the Calvinist Frederick V, elector palatine, as king of Bohemia on August 26, two days before the Habsburg candidate was elected emperor as Ferdinand II. Ferdinand had the support of Spain, Poland, the papacy, and, especially, the Catholic League.

Bohemian defeat at White Mountain

Their combined armies defeated the isolated Frederick at the Battle of the White Mountain near Prague in November 1620. Draconic measures destroyed Protestantism in Bohemia and the remaining Austrian territories, and the influence of the estates was abolished. The army of the Catholic League pursued the last partisans of Frederick, the "winter king," on their flight to the Netherlands. Maximilian of Bavaria was granted the electoral dignity, which had been the condition of the help that he gave to the Emperor.

The Thirty Years' War. The Bohemian revolt and its suppression are conventionally regarded as the beginning of the complex European struggle designated as the Thirty Years' War (*q.v.*). The surprising resuscitation of the power of the German Habsburgs, the occupation by the Spaniards of the Rhenish Palatinate, and the resumption of war in 1621 between Spain and the United Provinces after a 12-year truce called the European powers into the arena, for they did not wish a revival of the empire of Charles V. Thus, the warfare in Germany became a general European war. Christian IV of Denmark wished to acquire the bishoprics of Verden, Bremen, Osnabrück, and Halberstadt, which were under Protestant administrators and surrounded by secular principalities. Against him Ferdinand put a new army into the field under the leadership of Albrecht Wenzel von Wallenstein (*q.v.*). This soldier of fortune had acquired an enormously compact mass of lands in Bohemia by a rich marriage and augmented it from the confiscated estates of Protestant nobles. After his elevation to the dukedom of Friedland (1625), he turned this property into a huge armaments factory. Because of his enormous wealth, Wallenstein became independent of the imperial treasury and was thus, at first, indispensable to Ferdinand. Johann Tserclaes, Graf von Tilly, the general of the Catholic League, defeated Christian near Lutter (1626); Wallenstein occupied Jutland, Mecklenburg, and Pomerania; Christian was forced to make peace at Lübeck (1629), regaining his lands but renouncing his alliances with the north German princes and his claims on the bishoprics in Lower Saxony.

The rise of Wallenstein

Now at the height of his power, Ferdinand issued the Edict of Restitution (1629), enforcing again the ecclesiastical reservation and ordering all bishoprics and abbeys secularized since 1552 to be restored to the Roman Catholic faith. This edict showed the Emperor determined to revolutionize all existing political conditions; the independent and aggressive spirit that he revealed provoked even the opposition of the Catholic League and especially that of Maximilian of Bavaria. At a meeting at Regensburg (1630) the electors, backed diplomatically by France, forced Ferdinand to dismiss Wallenstein, to reduce the imperial army, and to accept the electors' control of foreign and military policy. Wallenstein's dismissal made it easier for the Swedish king Gustavus II Adolphus to decide to land in Pomerania. The King of Sweden was brought to this decision by Sweden's struggle for the *dominium maris Baltici* (supremacy in the Baltic sea), and Gustavus' intervention was also by his desire, based on his deep religious convictions, to liberate the north German Protestants and to obtain the large subsidies offered by France. Of the Protestant states only Hesse-Kassel voluntarily joined the Swedish king; Saxony was forced into an alliance because of counterpressure from Vienna. The princes feared for their freedom and dreaded the Emperor's revenge. In 1631, in the first of the battles of Breitenfeld, Gustavus Adolphus defeated Tilly, who had taken Magdeburg; the victory was so complete that the Catholic party collapsed completely in northern Germany and the Catholic League dissolved itself. The Swedes now occupied Munich and threatened Vienna.

Entry of Sweden into the war

In this danger, Ferdinand asked Wallenstein to recruit another army. Wallenstein, however, proceeded very cautiously, since he wanted both to demonstrate his indispensability and to avoid jeopardizing his new army. Gustavus Adolphus restored Protestantism in large areas of Germany but fell mortally wounded in the Battle of Lützen (1632). The Swedes, however, under the chancellor Axel Oxenstierna, continued the war, united the south German estates in the League of Heilbronn, and again threatened the Austrian hereditary lands. The Vienna court began to distrust Wallenstein because of his hesitant strategy. He negotiated with Saxony and Sweden without, however, following a clear line of policy. Suspected of high treason, he was murdered in 1634 at Ferdinand's instigation by some of his officers. The imperial army, still intact, and the newly arrived Spanish auxiliary troops defeated the Swedes at Nördlingen and again occupied southern Germany to the Rhine. Peace was made at Prague between the Emperor and Saxony (1635), and most German states agreed to its terms in time. The Peace of Prague fixed the religious divisions as they had existed in 1627 (an arrangement originally limited to 40 years and not applicable to Calvinists), renounced the Edict of Restitution, and gave the emperor command over a new imperial army to be provided by the estates and to be used against the foreign powers.

Despite this renewed effort, German strength did not suffice to enforce peace. War ravaged the country for another 13 years; it was no longer fought for German questions but to divide the spoils between Sweden and France. A renewed France under the Cardinal de Richelieu declared war against the Spanish world power on all fronts in 1635, regardless of doctrinal considerations but simply following the dictates of French reason of state. France gained footholds in Savoy, Mantua, and Parma and thus controlled Spanish Milan and the passes over the Alps into southern Germany. The United Provinces and those states of the empire that stood aside from the war were strengthened by subsidies from France. Along the western frontier of Germany, Richelieu occupied Lorraine and pushed his occupation troops through Alsace to the Rhine. In order to keep the Swedish army in the German theatre of war, Richelieu mediated a prolongation of the Swedish–Polish truce for another 26 years and granted annual subsidies for the Swedish armies fighting in the empire. There were meetings of the electors and of the Diet, but neither the Emperor nor the estates were able to bring about a general pacification. One prince after another made separate treaties with the foreign powers.

Continuation of the war

The Peace of Westphalia (1648). Ferdinand II was succeeded by his son Ferdinand III, emperor from 1637 to 1657. Nine years after his accession, with Spain's power sapped by internal revolts, the enemies of the House of Habsburg won the upper hand in Germany. In 1648 the French had crossed Bavaria and reached the Inn River, and the Swedes were again attacking Prague when the news of the Peace of Westphalia at last arrived.

Earlier peace feelers had been in vain, but, from 1644 onward, serious negotiations had been taking place among the interested powers. First, Spain concluded a separate peace with the United Provinces of the Netherlands, the sovereignty of which was recognized (though the Franco-Spanish war was to go on for 11 more years until the Treaty of the Pyrenees in 1659). The conditions offered by the Emperor and the states of the empire to the Swedes at Osnabrück and to the French at Münster were accepted, and peace was signed in October 1648. In addition to 5,000,000 thalers to pay off their troops, the Swedes obtained the archbishopric of Bremen, the bishopric of Verden, the town of Wismar, and Hither Pomerania; the French obtained the Habsburg rights in Alsace and the Lorraine bishoprics of Metz, Toul, and Verdun, together with the fortress of Breisach and the right to occupy the fortress of Philippsburg. The Swiss and the Dutch, now sovereign powers, severed all links with the empire.

Within the empire, Bavaria retained the electoral dignity and the Upper Palatinate; the Elector of Brandenburg received the bishoprics of Kammin, Halberstadt, and Min-



Germany in 1648.

den, with the reversion to that of Magdeburg, in compensation for what he had had to cede to Sweden; and the dukes of Mecklenburg received the bishoprics of Ratzeburg and Schwerin, also in compensation. Otherwise, the territorial status quo of 1618 was restored. The religious status quo of 1624 was acknowledged, except for the Austrian hereditary lands and Bavaria, and Calvinism was recognized as one of the creeds of the empire. Constitutionally, the Emperor's power was considerably reduced in favour of the estates or members of the empire, even in matters of foreign policy. To prevent majority decisions in denominational disputes, the estates were divided between a Corpus Evangelicorum, or Protestant group, and a Corpus Catholicorum, both of which had to be in agreement if a decision was to be valid for the empire as a whole. Finally, the fact that France and Sweden were guarantors of the peace laid Germany open to foreign intervention.

THE EMPIRE IN DECAY, 1648–1721

Emperor and empire after 1648. After the Peace of Westphalia, the expression emperor and empire (*Kaiser und Reich*) ceased to mean the whole body politic with its monarchical head but emphasized the contrast between the emperor, who had no power in the empire, and the empire, which had little use for the emperor. Free to form associations among themselves or with foreign princes (provided that they were not directed against empire or emperor), the rulers of the member states had become nearly fully sovereign and independent in international law. The emperor was reduced to the rank of an honorary president of an aristocratic republic; he was unable either to make laws or to levy taxes for the *Reich* without the consent of the Imperial Diet, or Reichstag. On the other hand, the territorial diets declined nearly everywhere—with the special exceptions of Mecklenburg and Württemberg, where the dukes were almost powerless. Therefore, the territorial princes could now build up their military power and finances without interference from above or below and so consolidate the basis not only for increasing independence but also for absolute government within their own territories.

Division into small principalities

As a result of the Peace of Westphalia, Germany was divided into about 300 entities with practically sovereign

rights and the quality of *Reichsstandschaft—i.e.*, that of being represented in the Reichstag as an estate of the empire. In addition to these quasi-sovereignties, there were nearly 1,500 other minor lordships that, without having *Reichsstandschaft*, enjoyed *Reichsunmittelbarkeit*, or "immediate" dependence on the *Reich*, with no suzerain other than the emperor. With no direct power over the *Reich*, the emperor was little more than the nominal guardian of law and order (which he could scarcely enforce) and the fountain of honour (for elevations in rank and the granting of privileges).

The territorial powers in the 17th century. The greatest of the secular powers in the *Reich* was the House of Habsburg, which ruled over a vast area in the southeast comprising Bohemia, Austria, Styria, Carinthia, Carniola, Istria, Trieste, and Tirol and also held numerous hereditary possessions in southern Swabia extending to the banks of the Rhine (in the area of Breisgau). These lands, however, had from time to time been distributed among various branches of the house and were not finally reunited until 1665 (after the extinction of the Tirolese branch). Moreover, the central authority for the Habsburg possessions, which Maximilian I had tried to set up, exercised little control, as many of the component territories kept their own administrations and diets. The Kingdom of Hungary was a Habsburg territory outside the *Reich*.

The rise of the Hohenzollerns

The decades after the Thirty Years' War saw the beginning of the rise of Brandenburg-Prussia to pre-eminence among the territorial powers of the *Reich*. From the middle of the 17th century, the House of Hohenzollern was outpacing the houses of Wittelsbach and Wettin in the struggle for the second place in the *Reich*. The Hohenzollern margraves of Brandenburg had inherited Cleves in western Germany, together with Mark and Ravensberg in Westphalia, in 1614, and the duchy of Prussia, which was a Polish fief outside the *Reich*, in 1618. The union of Ducal Prussia with Brandenburg was fundamental to the rise of the Hohenzollern monarchy to the rank of a great power in Europe. John Sigismund's grandson Frederick William, the Great Elector of Brandenburg, by military intervention in the Swedish–Polish War of 1655–60 and by diplomacy obtained the ending of Poland's suzerainty over Ducal Prussia at the Treaty of Oliva (1660). This made the Hohenzollerns sovereign over Ducal Prussia, whereas Brandenburg and their other German territories were still nominally parts of the *Reich* under the theoretical suzerainty of the Holy Roman emperor. Frederick William was also able to set up a centralized administration in Prussia and to secure control of the duchy's financial resources, thus reducing the power of the Prussian *Landstände*, or estates. These estates, comprising the landowning nobility and the oligarchies of the towns, among which Königsberg was paramount, had previously run the duchy's affairs.

When Frederick III of Brandenburg succeeded his father, the Great Elector, in 1688, Brandenburg-Prussia consisted of four separate groups of lands, scattered across Europe from the lower Rhine in the west to the Memel River in the east and unconnected by any territorial link. The central and largest group comprised Electoral Brandenburg (Altmark, Mittelmark, and Neumark), with the secularized bishoprics of Magdeburg and Halberstadt adjoining it in the southwest and Hinterpommern (Farther Pomerania) in the northeast; the second group comprised the duchy of Cleves, on the lower Rhine, and the countship of Mark, on the Ruhr River (these two territories were not contiguous); the third group comprised the countship of Ravensberg and the secularized bishopric of Minden, between the upper Ems and the middle Weser; and finally there was Ducal Prussia in the east, independent of the *Reich*.

The most significant achievement of the Great Elector's son was to secure the royal dignity for himself, as Frederick I, king in Prussia; he crowned himself at Königsberg on January 18, 1701, after the emperor Leopold I had consented to his assuming this new status in order to ensure his goodwill in the forthcoming War of the Spanish Succession. Ducal Prussia thus became the basis of his rank as sovereign king; and the other Hohenzollern possessions, though theoretically they remained within the German *Reich* and under the ultimate overlordship of the emperor, came to be treated in practice as part of the Prussian kingdom rather than as distinct from it.

Among the other territorial powers, the House of Wettin was divided into so many branches that its head, the elector of Saxony, ruled only a portion of its lands. The same was true of the House of Wittelsbach: its largest and most compact territory was Bavaria, whereas its scattered lands in the Palatinate were distributed among junior branches of the family. The House of Welf, which had the leading role in northwestern Germany, also divided its lands in Brunswick and Hanover among various branches (seven or sometimes even more); and the House of Hesse was similarly divided. In the extreme southwest, the House of Zähringen, with the margraviate of Baden, and the House of Württemberg were the most important dynasties.

Most of these great princes in the course of the 17th and 18th centuries succeeded in reducing the old-established powers of the diets in their territories, though they did not dare to suppress them outright. Government and justice, however, in the hands of paid officials, were far more efficiently administered in the greater principalities than they were in the lesser ones. The lesser secular princes ruled their subjects patriarchally but employed quantities of officials out of all proportion to the size of their miniature territories and maintained sumptuous courts, for the upkeep of which they were ever preoccupied with devising new taxation. The ecclesiastical princes, in general, governed no better than the secular ones, since their lack of legitimate issue made them sometimes careless of the future well-being of their countries. Elected by the chapter with which he was to share the administration and which was composed largely of younger sons of the local nobility, a bishop or abbot would often use his position to further the interests of his family by settling his brothers, cousins, and nephews on church property or bringing them into the chapters.

The major imperial cities

Some of the imperial cities were towns of major importance with a long-standing commerce, among them Lübeck, Hamburg, Bremen, Nürnberg, Augsburg, Ulm, Frankfurt am Main, and Strassburg. Most, however, were little country towns in southwestern Germany, with 3,000 inhabitants or less, which had long ago lost any significance. Overshadowed by the capitals rising around the princely residences, the imperial cities were for the most part governed by a few patrician families who monopolized all positions of influence and profit and used them to their personal advantage.

Finally, there were the dominions of the "immediate" knights of the empire (not to be confused with the knights of orders, whose master was an ecclesiastical prince), nearly all in southwestern Germany. These knights were large landowners who exercised sovereign rights over their manor and perhaps some adjacent village.

In the first decade after the Peace of Westphalia, the primary concern of all the estates was the maintenance of peace. As the war between France and Spain lasted until 1659 and as grave issues were arising in the north and in the east, Germany could easily have been involved again in a general war. As no one believed in the ability of the emperor to safeguard the empire from this danger, there grew up a network of alliances between the different territories for mutual defense. The elector of Mainz, Johann Philipp von Schönborn, succeeded in uniting the most important Catholic and Protestant princes in a great defensive alliance. This first Rheinbund, or League of the Rhine, was signed in August 1658 for three years and had for its object the full execution of the Peace of Westphalia, the prevention of foreign wars, and the defense of its members' own territories; but, in fact, the alliance, which the emperor regarded as directed against his authority, soon became largely dependent on France. It was frequently renewed and lasted until 1667.

During the First Northern War (1655–60), Emperor Leopold, in the Catholic interest, supported the Catholic king of Poland, John Casimir; the elector of Branden-

burg, Frederick William the Great, who at the outset had supported Sweden, later entered into an understanding with Poland and the Emperor. This warfare, however, was waged, for the most part, outside the territory of the empire.

Rise of Louis XIV

Germany was far more deeply distrurbed by the course of events on its western frontier, where Louis XIV of France was preparing to invade the United Provinces. When Louis induced the archbishop of Cologne, Maximilian Heinrich of Bavaria, and the bishop of Münster, Christoph Bernhard of Galen, to cooperate with his invasion of the United Provinces, the neighbouring German states, fearing that they would be involved in the war, invoked the assistance of the Emperor and the empire. Moreover, Louis XIV had, without any legal grounds, driven the duke of Lorraine, Charles IV, who was a prince of the empire, out of his duchy. The empire, thus, had good cause to intervene.

Yet Leopold hesitated to take any action against Louis XIV. Foreseeing the extinction of the Spanish branch of the Habsburgs, his main policy was to assure its succession for his own branch. This would require the French king's acquiescence, for which Louis let him hope, so that Leopold did not feel himself at liberty to oppose him. Hence, the Dutch found their sole support in the elector of Brandenburg, Frederick William, who had grown up in Holland and whose first wife had been a princess of Orange. By cutting the dikes and flooding their country, the Dutch were able to avert the French attack in 1672. At the urgent request of the estates, Leopold determined to send an army, under Raimondo Montecuccoli, for the defense of the imperial frontier on the Rhine but with instructions to maintain the defensive. Finally, having allied himself with Spain, the Emperor secured the declaration of war by the *Reich* against France in 1674.

The war was waged chiefly in the Austrian Netherlands and on the Rhine. The Swedes, in alliance with Louis, invaded Brandenburg from Pomerania to restrain Frederick William, who was participating personally in the war against France. At the same time, Louis entered into relations with Poland, with Turkey, and also with Hungary, which was discontented with Habsburg rule. The Emperor then found himself threatened in the rear. By his victory at Fehrbellin on June 28, 1675, Frederick William drove the Swedes out of his territory and occupied nearly the whole of Swedish Pomerania. The French, on the other hand, were for the most part victorious on the western front, and, eventually, Louis was able to induce first the Dutch and then Spain to conclude a separate peace (1678). When the Emperor could no longer hope to win anything by continuing the war, he acceded to the Treaty of Nijmegen in February 1679.

Convinced by these experiences that the *Reich* was too weak to withstand encroachments on its territory, Louis next proceeded to set up special courts, the so-called chambers of reunion, for the purpose of determining what lands had at any former time belonged to districts already ceded to him. On the strength of these investigations, he declared that the countship of Mömpelgard (Montbéliard), the whole of Alsace, and certain districts in the Palatinate and in the electorate of Trier belonged, by right, to France. French armies occupied the territories, and the imperial city of Strassburg was seized on September 28, 1681, and at once erected into a powerful French fortress. The Emperor and the Reichstag contented themselves with ineffectual protests.

The Emperor's failure to safeguard the integrity of the *Reich* at this crisis is partly to be explained by his preoccupation with the dangers threatening his own territories. The Turks in Hungary were planning an invasion of Austria, and in the summer of 1683 they appeared with a powerful army before the walls of Vienna. The imperial generals, however, managed to hold off the attack; and, finally, John III Sobieski of Poland and the electors Maximilian II Emanuel of Bavaria and John George III of Saxony arrived with an army to defeat the Turks and free the city.

When Louis XIV sought to bring the Palatinate within his grasp, claiming it as the inheritance of his sister-in-law Elizabeth Charlotte of Orléans, war once more broke out in the west. French armies invaded the Palatinate, and the Emperor allied himself with England and the Dutch. The War of the Grand Alliance inflicted severe losses on the French but did not result in any decisive victory. Once more Louis was able to sow dissension among his enemies and isolate Germany, and the Emperor had to conclude the Treaty of Rijswijk in 1697, with but small gain to himself. Louis restored Lorraine to the rightful duke, Leopold Joseph, and abandoned his claims to the Palatinate and to the districts outside Alsace that had been declared to have once formed part of it. But he retained the whole of Alsace and also Strasbourg.

The War of the Spanish Succession (1701–14). The last Habsburg king of Spain, Charles II, died childless on November 1, 1700. The question of succession, which had already occupied the European diplomats for decades, had now to be settled. Both Leopold I and Louis XIV each supported a different candidate for the Spanish throne. Louis supported his grandson Philip, while Leopold maintained the right to the Spanish inheritance of his son, the archduke Charles; this conflict erupted in 1701 in the War of the Spanish Succession. The English and the Dutch, led by William III of Orange, stadtholder of the Netherlands and from 1689 also king of England, supported the Archduke's claim, while the French had the help of the Hungarians, of the English Jacobites, and of a number of German princes, chief of whom was Maximilian Emanuel of Bavaria. In the first years of the war, the advantage lay with the French, who might have struck a deadly blow at Austria by a concerted attack from the Rhine and Italy. This danger was averted by the great Anglo-Austrian victory in the Battle of Blenheim on August 13, 1704. Bavaria was occupied by the imperial troops, and Maximilian Emanuel fled to Brussels.

Battle of Blenheim

Leopold I died in May 1705. His eldest son, Joseph I, succeeded him as emperor (1705–11). The war against France was pursued with great successes throughout Joseph's reign. The archduke Charles landed in Catalonia and by the summer of 1706 was in Madrid. After French defeats in the Netherlands and in Italy, Louis XIV, in 1708 expressed his readiness to renounce on behalf of his grandson all claims to the Spanish throne and to agree to a restoration of the Franco-German frontier on the line laid down in the Peace of Westphalia in 1648, which would have meant restoring Strasbourg. The allies, however, demanded that French troops should help them to expel Philip V from Spain. When Louis refused this, they broke off negotiations. Similarly, after the Battle of Malplaquet (1709), when Louis even offered to give up his acquisitions in Alsace and the bishoprics of Metz, Toul, and Verdun, together with money to be used for driving Philip V out of Spain, the allies stood by their former demands, so that negotiations once more broke down.

The death of the emperor Joseph I (April 1711) changed the situation. As he left no son, his younger brother the Archduke succeeded him as the emperor Charles VI (1711–40). Now, if Austria, the empire, and the Spanish inheritance were all to be united under one ruler, there was danger that such a disproportionate concentration of power would threaten the European balance of power more seriously than would the establishment of a French dynasty in Spain; and neither England nor the Dutch nor Savoy felt disposed to prosecute the war for such an objective. A peace congress met at Utrecht in 1712, and the chief treaties were signed in 1713. These awarded only a portion of Lombardy, the Neapolitan mainland, and the former Spanish Netherlands to Charles VI; and no mention was made of a restoration of the old Franco-German frontier. Meanwhile, the Emperor and the *Reich* were still at war with France, but, when Landau and Freiburg im Breisgau had been lost, they were forced in the treaties of Rastatt and Baden (March and September 1714) to assent to the terms of Utrecht. The frontiers of the German *Reich* remained as laid down in the Treaty of Rijswijk; the Elector of Bavaria recovered his

territories; and the prospect of winning back the old German territories in the southwest had completely vanished. Thus, the War of the Spanish Succession had overthrown the supremacy enjoyed by France in Europe without yielding any profit to Germany.

The Great Northern War (1700–21). At the same time, another long-standing quarrel was being fought out in the north and the east. In 1697 the elector Frederick Augustus I of Saxony, after conversion to Catholicism, had been elected king of Poland as Augustus II the Strong. In alliance with Denmark and with the Russian tsar Peter I the Great, he began a war against Charles XII of Sweden in the summer of 1700, with the object of breaking the power that the Swedes had been exercising in the Baltic area since the Thirty Years' War. By the spring of 1706, however, Charles XII was in possession of the greater part of Poland and, in agreement with a section of the Polish nobility, had set up Stanisław (Leszczyński) I as king in opposition to Frederick Augustus. He then marched through Silesia into his enemy's German territories, occupied a great part of Saxony, and established his headquarters, during the winter of 1706–07, in Altranstädt near Leipzig. But the peace that Frederick Augustus had to sign, renouncing the Polish throne, did not long remain in force. When Peter the Great of Russia threatened to occupy Poland, Charles marched against him and was defeated at Poltava (1709). The Swedes were driven from Poland, and their allied enemies invaded Swedish territory on all sides. The elector Frederick III of Brandenburg, who with the emperor Leopold I's consent had assumed the title of king in Prussia as Frederick I in 1701, also took part in the attack. Pomerania, Bremen, and Verden, Sweden's possessions in Germany, were seized.

The treaties of Stockholm and Nystad

After Charles XII's death (1718), the conclusion of peace was possible. Sweden had to surrender Bremen and Verden to Hanover (1719) and western Pomerania south of the Peene to Prussia (Peace of Stockholm, 1720). Of greater importance for future Russo-German relations, however, was the Treaty of Nystad (1721), whereby Sweden ceded Estonia, Livonia, and Ingria to the Tsar, thus giving Russia a firm foothold on the shores of the Baltic and thereby a position that became more and more threatening to Germany. Sweden, meanwhile, retained the northern part of western Pomerania and Rügen.

AUSTRO-PRUSSIAN RIVALRY IN THE 18TH CENTURY

Charles VI and the Pragmatic Sanction. From his youth the emperor Charles VI had been regarded by the emperor Leopold as a future king of Spain and had been brought up in the spirit of Spanish Jesuitism. In his reign (1711–40) Austrian policy became more and more obviously inspired by the desire to strengthen Habsburg influence in Italy and also to win more territory along the lower reaches of the Danube. In this latter aim Charles was at first successful through the war that he entered in 1716 in support of Venice against the Turks. By the Treaty of Passarowitz (1718), thanks to Prince Eugene's victories, Austria obtained the Banat, Little Walachia, northern Serbia (with Belgrade), and part of Bosnia.

Imperial authority in Germany, meanwhile, was based both on the imperial constitution and on the territorial power of the Emperor and his house. As the various Austrian lands grew into a political organism, the interests of the *Reich* and the interests of the House of Habsburg diverged more and more. This is clearly visible in the disputes between the Aulic Council, led by Ludwig Phillipp, Graf von Sinzendorf, and the Imperial Chancery under Friedrich Karl, Graf von Schönborn, nephew of the archbishop of Mainz, Lothar Franz von Schönborn. The smaller secular and ecclesiastical states of the *Reich* were concerned to retain and to strengthen imperial power, but the larger states, especially Hanover and Brandenburg, which wanted to expand, were opposed to this. Four of the most powerful electors were now sovereign rulers: the Elector of Brandenburg was king in Prussia; the Elector of Hanover was king of Great Britain; the Elector of Saxony was king of Poland; and the Emperor himself was king of Hungary. Moreover, the kings of Denmark and of Sweden were also princes of the empire. Whereas the smaller states regarded the imperial power as the only guarantee of their "liberty" against aggression, the greater states regarded that power as a danger to their own liberty, fearing that the Emperor might erect a "despotic" or truly monarchical regime. Whenever the interests of the more powerful states were involved, the Emperor found few means at his disposal to render effective legal and constitutional decisions. Indeed, the evolution of the territorial states undermined the imperial constitution and the feeling for the unity of the *Reich*. There was no common political history for Germany in this period, though Germany's destiny as a whole was to be deeply influenced by international affairs.

Attempts to keep the dynasty intact

When Charles had gone to Spain to enter into his inheritance there, a secret family compact had been made to regulate the succession, the *Pactum mutuae successionis* (1703). After Spain had been lost and Charles had succeeded to Austria, it became necessary to adapt the precedence of the heirs and the order of their succession to the changed circumstances if Austria was to remain a unitary state and to be saved from partition among the various claimants. This was done not only in the interest of Austria and of the dynasty but also in that of Germany and, indeed, of Europe, since the European balance of power depended on the survival of the undivided Austrian state. Charles spared no effort to ensure that after his death his wishes respecting the succession would be carried out. These wishes he embodied in a special law, the Pragmatic Sanction, for which he secured the approval of the diets in all his territories.

Charles did not consider this security enough; he sought to have the Pragmatic Sanction recognized by the great powers and approved by the Reichstag. The approval of the Reichstag was especially difficult to obtain, because two of the most important princes in Germany, the future electors Frederick Augustus II of Saxony and Charles Albert of Bavaria, had, in 1719 and 1722, respectively, married Emperor Joseph I's daughters and so had an immediate personal interest in frustrating the execution of Charles VI's wishes. In these circumstances, the Emperor particularly desired Prussia's consent to his plan. Frederick William I of Prussia was quite willing to fall in with the Emperor's wishes but demanded in return that the Emperor should help him to pursue his claim to part of the inheritance of the duchies of Jülich and Berg on the Lower Rhine. The Emperor appeared to assent to this and so secured Frederick William's agreement to the Pragmatic Sanction by the Treaty of Berlin (December 23, 1728). Soon, however, it became clear that the Emperor had made contrary promises to the rival claimants to Jülich and Berg. This estranged Prussia from Austria, and Frederick William allied himself with the Emperor's enemies.

The War of the Polish Succession (1733–35). In 1733 Augustus II of Poland (that is, Frederick Augustus I of Saxony) died. While Austria and Russia declared themselves in favour of the succession of his son Augustus III (Frederick Augustus II), a number of Polish nobles chose Augustus II's former opponent, Stanisław Leszczyński, whose daughter was married to the young French king, Louis XV.

When the War of the Polish Succession broke out, the French used the opportunity to acquire at last the long-disputed Duchy of Lorraine. The Spanish Bourbons also entered the war in order to realize their claims to the Italian duchies of Parma and Piacenza. Though the *Reich* as a whole took the Emperor's side, the Wittelsbach rulers of Bavaria, the Palatinate, and Cologne refused to do so. Old and ill, Prince Eugene of Savoy could only hold the defensive with a small army, and he advised the Emperor to make peace. As the French minister Cardinal André-Hercule de Fleury also wanted peace, the preliminary agreements of 1735 became the basis of the final Treaty of Vienna (1738). The Emperor had to abandon Sicily and Naples, which were placed under the rule of one of the Spanish princes, Don Carlos, though the duchies of Parma and Piacenza were ceded to Austria. Of special significance to Germany was the Em-

peror's consent to the cession of Lorraine, which was made over to Stanisław Leszczyński as compensation for his renunciation of the Polish crown but which was to revert to France after his death. The reigning duke of Lorraine, Francis Stephen, who was married to Charles VI's heiress, Maria Theresa, was compensated with the Grand Duchy of Tuscany. France, in return, recognized the Pragmatic Sanction—with the important reservation that it only did so insofar as it did not conflict with established third-party rights.

The War of the Austrian Succession (1740–48). When Charles VI died on October 20, 1740, his daughter Maria Theresa at once assumed the government of the countries belonging to the House of Habsburg. The interests of power politics, however, proved stronger than the treaties by which her father had sought to secure the succession for her. Though Germany and even Europe required that the integrity of her inheritance should be maintained, the electors Frederick Augustus II of Saxony (Augustus III of Poland) and Charles Albert of Bavaria protested against her accession and were supported by the French, who desired to see a partition of the Austrian territories; a still more pressing danger threatened Maria Theresa when the new Prussian king joined her opponents. Frederick II the Great of Prussia at first offered Maria Theresa his help against her enemies if she would cede to him a part of Silesia to which Prussia had some claim by virtue of old dynastic agreements. When she rejected his proposal, he determined to occupy the disputed Silesia territory by force.

The aggression of Frederick II of Prussia

Frederick crossed the Silesian frontier in December 1740, advanced as far as Breslau, and defeated an advancing Austrian Army near Mollwitz in April 1741. At the same time, the Bavarians, supported by a French Army, advanced as far as Linz and even seized Prague with the help of the Saxons. Next, the German electors, under French and Prussian influence, elected as emperor not Maria Theresa's husband, Francis Stephen, but the elector Charles Albert of Bavaria (January 24, 1742), who assumed the name of Charles VII. When Frederick, having occupied all Silesia and invaded Moravia, won the Battle of Chotusitz against an Austrian army, Maria Theresa thought it prudent to open negotiations—though her troops had regained Linz and even invaded Bavaria. England, also at war with France over colonial questions, acted as intermediary, and, thus, the Treaty of Breslau (June 11, 1742) was concluded between Austria and Prussia. Austria ceded the greater part of Silesia along with the Countship of Glatz (now Kłodzko, Poland) to Prussia and retained only the principalities of Troppau (now Opava, Czechoslovakia) and Teschen. In return, Frederick promised his neutrality. The First Silesian War, which ended with this peace, established the military reputation of Frederick the Great. It was the first armed contest between the two greatest German states that had developed out of the old *Reich*—states that had long regarded each other with distrust and jealousy.

The struggle for the inheritance of Charles VI continued. The Austrian Army captured Prague, freed Bohemia from the invaders, and even captured Munich, the Bavarian capital. A British Army defeated the French at Dettingen and advanced from Hanover as far as the Rhine. Then, when the Austrians, led by Prince Charles of Lorraine, were advancing from southern Germany with the intention of crossing the Rhine, Frederick the Great decided to intervene again, lest Maria Theresa, after a complete victory over her other enemies, should try to wrest back Silesia. Having signed a new alliance with France, he invaded Bohemia in August 1744 but had to withdraw, because the expected simultaneous advance of the French Army did not take place. Then, Charles VII died (January 20, 1745), and his son, the elector Maximilian III Joseph of Bavaria, at once made peace with Maria Theresa, recovering his ancestral domains in return for renunciation of all his claims to the Austrian inheritance. Frederick Augustus II of Saxony had already abandoned his claims and made peace with Austria, and, as the French were fully occupied in Flanders, Frederick found himself alone opposed to the main force of Austria. Strengthened by Saxon troops, the Austrians attacked Silesia but were defeated near Hohenfriedeberg on June 4, 1745. Once more Frederick invaded Bohemia, and, by the end of the year, a great part of Saxony was in his possession. On December 15, his chief general, Leopold of Anhault-Dessau, won a fresh victory over the Austrians and Saxons at Kesselsdorf, near Dresden.

The majority of the electors had, in the meantime, elected Francis Stephen of Lorraine as emperor. As Francis I, he was nominal head of the *Reich* from 1745 to 1765. Maria Theresa, who now saw that it would not be so easy to retake Silesia and who laid great stress on the recognition of her husband as emperor by Frederick, was ready to reopen negotiations. On December 25, 1745, the Second Silesian War was brought to a close by the Treaty of Dresden, by which Frederick retained Silesia and recognized Francis I as emperor.

The election of Francis

The war against France lasted nearly three years more, but French victories on land were more than offset by the British victories at sea. Finally, by the Treaty of Aix-la-Chapelle (1748), Maria Theresa was recognized as her father's sole heiress but gave certain frontier districts in Lombardy to Savoy and the Duchy of Parma to the Spanish prince Philip. Maria Theresa had to come to terms mainly because Great Britain wanted to end the war.

The Seven Years' War (1756–63). Though peace was, thus, outwardly restored, tension between Austria and Prussia remained, since Maria Theresa had never abandoned her hopes of regaining Silesia. These hopes were shared by her chief minister, Wenzel Kaunitz, Fürst von Kaunitz-Rietberg, who looked on Prussia as the natural enemy of Austria. Neither party, however, wished to resume the contest without the help of powerful allies. Kaunitz had already established relations with Russia, and his special concern was to induce France, Austria's old enemy, to support his schemes. The French, meanwhile, despite their Prussian alliance, viewed the growing power of Prussia with dislike. Frederick, for his part, was convinced that he would have to defend Silesia against an Austrian attack and so was anxious to detach the British from their traditional support of Austria if he could do so without impairing his formal friendship with France, Great Britain's principal rival.

In 1755, when French and British settlers in North America had already come to blows and the Austrian government showed itself unwilling to undertake the protection of Hannover in the event of Anglo-French hostilities in Europe, the British government sought a promise of help from Prussia. Since it was to be a purely defensive agreement, Frederick considered that he could enter into it without breaking faith with France and, on January 16, 1756, concluded the Convention of Westminster, which stipulated that Prussia should help Great Britain if the French attacked Hanover and that Great Britain should support Frederick if the Austrians attacked Silesia. The French, however, saw this convention as a defection by Frederick from his French commitments, and, on May 1, 1756, at Versailles, the defensive alliance for which Kaunitz had so long laboured was concluded between France and Austria.

The Convention of Westminster

Kaunitz, with the zealous support of Russia, set to work in Paris to turn the Austro-French defensive alliance into an offensive one for the complete destruction and partition of Prussia; and Frederick the Great learned that an attack on Prussia by Austria, Russia, and France was being planned for the following spring. Frederick, resolved to anticipate his enemies, gave the order to his troops to cross the Saxon frontier. Thus began the Seven Years' War (*q.v.*), in which Prussia fought against Austria, France, Russia, Saxony, and Sweden. Frederick's sole ally was Great Britain–Hanover, the support of which consisted mainly in a subsidy of 4,000,000 thalers a year. It was only his generalship and his determination not to consent to any dimunition of his territory that made it possible for him to survive.

In the autumn of 1756 Frederick occupied Saxony and compelled the Saxon army to surrender at Pirna. In 1757 he invaded Bohemia and besieged Prague but was him-

self defeated at Kolín on June 18 and forced to withdraw. At the same time Russian troops threatened East Prussia, one French Army overran Hanover, and a second French Army, in conjunction with an imperial army, advanced on Berlin. Frederick defeated the French and imperial forces at Rossbach on November 5 and then hastened to Silesia, where, by the Battle of Leuthen (December 5), he recovered Breslau from the Austrians. The moral effect of Rossbach was even greater than its military effect, for it brought many members of the *Reich* to "a Fritzian way of thinking" (Goethe): it was not only a victory over the French but also a victory over the *Reich*, which, notwithstanding the smaller states' leadership in cultural matters, had declined in political importance while Austria and Prussia rose to power.

The following years, however, brought Prussia to the verge of disaster. While the French advanced as far as the Weser, Austrian and Russian armies together defeated Frederick at Kunersdorf on August 12, 1759. With his army well-nigh exterminated, Frederick thought that Prussia was bound to fall; however, disunity among leading Russian and Austrian commanders gave him time to recruit a new army. The imperial army entered Dresden and occupied a part of Saxony, but the illness of the empress Elizabeth hindered the operations of the Russians, since it was common knowledge that the sympathies of the heir apparent, the future Peter III, were with Frederick. Nevertheless, it became daily more difficult for Frederick to obtain reinforcements and money. After the fall of the elder William Pitt's ministry, the British government began to negotiate for peace with France and ceased its financial support of Prussia from April 1762.

By this time, however, Frederick's position had been decisively improved by the death of Elizabeth of Russia on January 5, 1762: her successor, Peter III, made peace with Frederick and entered into an alliance with him. Though Peter was murdered a few months later, his successor, Catherine II the Great, also thought that neither a strengthening of Austria nor the destruction of Prussia would serve Russia's true interests. She withdrew her military support from Frederick but did not renew the alliance with Austria.

Furthermore, after signing the preliminaries of peace with Great Britain (November 1762), the French lost all interest in the war with Prussia and withdrew their troops from Germany. Hence, Maria Theresa gave up all hope of a decisive victory and entered into negotiations. Peace was signed at Hubertusburg on February 15, 1763. Frederick evacuated Saxony but retained Silesia.

The importance of the Seven Years' War in German history lies in the failure of Austria's attempt to destroy Prussia before Prussian power was consolidated. The hostility between the two greatest German states continued to exist and to influence powerfully the whole future political development of Germany. (Ed.)

IV. Circa 1760–1871

GERMANY FROM C. 1760 TO 1815

Germany in the middle of the 18th century was a country that had been drifting in the backwaters of European politics for more than a hundred years. The decisive roles in the affairs of the continent were played by those great powers, such as France, England, and Spain, whose economic resources and commercial connections provided a solid foundation for their military might. The states of central Europe, on the other hand, floundered in a morass of provincialism and particularism. All the forces that had contributed to the rise of powerful national monarchies west of the Rhine were lacking in the east. In the Holy Roman Empire the central government was losing rather than gaining strength, the princes were enlarging their authority at the expense of the crown, and business initiative was being discouraged by the lack of political unity and by the remoteness of the major trade routes.

The Holy Roman Empire in the late 18th century

Civic power became increasingly concentrated in the hands of local governments controlled by aristocratic overlords, ecclesiastical dignitaries, or municipal oligarchs. The history of Germany between the Thirty Years' War and the French Revolution is largely the sum total of the histories of dozens upon dozens of small political units, each enjoying virtually full rights of sovereignty. The rulers of these gingerbread principalities, copying the example of the royal court of France or Austria, built costly imitations of the palaces of Versailles and Schönbrunn, which today are the delight of tourists but which were once the curse of an impoverished peasantry. The tradition of princely authority, an instrument of national greatness in western Europe, encouraged national divisiveness in central Europe. The petty rulers of Germany legislated at will, levied taxes, concluded alliances, and waged wars against each other and against the emperor. Policies pursued in Munich, Stuttgart, Dresden, or Darmstadt reflected policies originating in Paris, Vienna, London, or Madrid, but without seeking a goal greater than the promotion of particularistic interests.

Political institutions designed theoretically to express the will of the nation continued to function, yet they had become an empty shell. The Holy Roman emperor was still elected in accordance with a time-honoured ritual that proclaimed him to be the successor of Caesar and Augustus. The splendid coronation ceremony in Frankfurt am Main, however, could not disguise the fact that the office conferred on its holder little more than prestige. Since all the emperors except Charles VII were Habsburgs by birth or marriage, they enjoyed an authority that had to be respected. But that authority rested not on the prerogative of the imperial crown but on the possession of hereditary lands stretching from Antwerp in the west to Debrecen in the east. The sovereigns of the Holy Roman Empire, in other words, were able to play an important role in German affairs by virtue of their non-German resources. And since Germany was not the main source of their strength, Germany was not the main object of their concern. The emperors tended to regard the dignity bestowed upon them as a means of furthering the interests of their dynastic holdings. The imperial Diet meeting in Regensburg had also become an instrument for the promotion of particularistic advantage rather than national welfare. It continued in theory to express the will of the estates of the realm meeting in solemn deliberation. In fact it had degenerated into a debating society without authority or influence. The princes had ceased to attend the sessions, so that only diplomatic representatives were left to discuss questions for which they were powerless to provide answers. The other central institutions of the empire, such as the imperial cameral tribunal in Wetzlar, languished in indolence. Constitutionally and politically Germany around 1760 resembled Poland in that a once vigorous and proud state had become weakened by internal conflict to the point that it invited the intervention of its more powerful neighbours.

What saved Germany from the fate of Poland was the ability of one of the member states to defend the empire against aggression. For 200 years Austria acted as the bulwark of central Europe against French expansion. Its possessions, forming a chain of protective bases extending between the North Sea and the Danube, had time and again borne the brunt of attacks by Bourbon armies. The frontiers of France kept moving closer to the Rhine, but the Holy Roman Empire was at least spared the tragedy of partition that befell the Polish state. It was partly in recognition of the vital role that the Habsburgs played in the defense of Germany that the electors chose them emperors with such regularity. The Austrian monarchy, moreover, endowed with resources comparable to those of the western nations, was able to pursue a policy of political rationalization with greater success than the principalities. The rulers in Vienna succeeded in improving the administration, strengthening the economy, and centralizing the government. Until the middle of the 18th century Austria remained the only great power east of the Rhine.

Traditional role of the Austrian Habsburgs

Further rise of Prussia and the Hohenzollerns. The emergence of the Hohenzollerns as rivals of the Habsburgs and the beginning of the Austro-Prussian dualism created the possibility of reversing the process of civic

decentralization that had prevailed in central Europe since the late Middle Ages. The interests of the territorial princes of the Holy Roman Empire inclined them toward a policy of particularism, while the government of Austria, with its Flemish, Italian, Slavic, and Magyar territories, could not perforce become the instrument of German unification. Prussia, on the other hand, was militarily strong enough and ethnically homogeneous enough to make national consolidation the main object of statecraft. But though the creation of a united Germany became its mission, this mission was not from the outset one that it accepted willingly or even consciously. The intention of Frederick the Great and of his successors Frederick William II and Frederick William III was to pursue dynastic rather than national objectives. Like the lesser princes of central Europe, all they sought was to maintain and enlarge their authority against the claim of imperial supremacy. Far from wanting to end the disunity of Germany, they hoped to prolong and exploit it. The patriotic Prussophile historians who a hundred years later argued that what Bismarck had achieved was the consummation of what Frederick had sought were letting the present distort their understanding of the past. In fact, the greatest of the Hohenzollerns had remained as indifferent to the glaring political weaknesses of his nation as to its great cultural achievements. His attitude toward the constitutional system of the Holy Roman Empire was similar to that of the self-seeking princelings who were his neighbours and from whom he was distinguishable only by talent and power. He may have scorned their sybaritic way of life. But politically he wanted what they wanted: the freedom to seek the advantage of his dynasty without regard for the interests of Germany as a whole.

The interests of Frederick the Great and his successors

His preoccupation with the welfare of his state rather than with that of his nation is apparent in the strategy by which he tried to check Habsburg ambitions after the Seven Years' War (1756–63). During the first half of his reign he had relied primarily on military force to restrict and undermine imperial authority. In the second half he preferred to employ the weapons of diplomacy to achieve the same end. In 1777 the ruling dynasty of Bavaria came to an end with the death of Maximilian Joseph. The elector of the Palatinate, Charles Theodore, now became ruler over the territories of both branches of the House of Wittelsbach. Without legitimate offspring to whom to leave his state and without affection for his newly acquired eastern possessions, he agreed to a plan proposed by Emperor Joseph II according to which part of the Bavarian lands would be ceded to Austria. But any increase in the strength of the Habsburgs was unacceptable to Frederick the Great. With the tacit approval of most of the princes of the empire, he declared war against Austria in 1778, hoping that other states within and outside central Europe would join him. In this expectation he was disappointed. Yet Joseph also became discouraged by the difficulties encountered in what he had believed would be an easy success. The War of the Bavarian Succession dragged on from the summer of 1778 to the spring of 1779, with neither side enhancing its reputation for prowess on the field of battle. There was much marching back and forth, while hungry soldiers scrounged for food in what came to be called the "potato war." The upshot was the Treaty of Teschen (May 1779), by which the Austrian government abandoned all claims to Bavarian territory except for a small strip along the Inn River. The conflict had brought Frederick no significant military victories, but he had succeeded in frustrating Habsburg ambition.

War of the Bavarian Succession

Joseph II, however, was a stubborn adversary. In 1785 he once again advanced a plan for the acquisition of Wittelsbach lands, this time on an even more ambitious scale. He suggested to Charles Theodore nothing less than an outright exchange of the Austrian Netherlands for all of Bavaria. The Emperor, in other words, proposed to surrender his distant possessions on the North Sea, which were difficult to defend, for a territory that was contiguous and a population that was assimilable. The scheme went far beyond that which Prussia had defeated seven years before, and Frederick opposed it with equal determination. He hoped to enlist the diplomatic aid of France and Russia against what he regarded as an attempt to upset the balance of power in central Europe. But more than that, he succeeded in forming a Constitutional Association of Imperial Princes, which 17 of the more important rulers in Germany joined. The members pledged themselves to maintain the fundamental law of the empire and to defend the possessions of the governments included within its boundaries. The growing opposition to the absorption of Bavaria by Austria persuaded Joseph that the risks inherent in his plan outweighed its advantages. The proposed exchange of territories was dropped, and Frederick could celebrate yet another triumph of his statecraft, the last of an illustrious career. But the association of princes that he founded did not survive its author. Its sole purpose had been the protection of princely prerogative against imperial authority. Once the danger had passed, it lost the only justification for its existence. Those nationalists who later maintained that it foreshadowed the creation of the German Empire misunderstood its origins and objectives. It was never more than a weapon in the struggle for the preservation of a decentralized form of government in central Europe.

The subordination by the Hohenzollerns of national to dynastic interests was even more apparent in the role they played during the partitions of Poland. Frederick the Great was the chief architect of the first partition, that of 1772, by which the ill-starred kingdom lost about a fifth of its inhabitants and a fourth of its territory to Prussia, Russia, and Austria. His successor, Frederick William II, helped to complete the destruction of the Polish state by the partitions of 1793 (between Prussia and Russia) and 1795 (between Prussia, Russia, and Austria). The result was bound to be an enhancement of Prussia's role in Europe but also a diminution of its role in Germany. The Hohenzollerns willingly embarked on a course that would in time have transformed their kingdom into a binational state comparable to the Habsburg empire. The German population in the old provinces would have been counterbalanced by the Slavic population in the new; the Protestant faith of the Prussians would have had to share its influence with the Catholicism of the Poles; the capital city of Berlin would have found a competitor in the capital city of Warsaw. In short, the centre of gravity of the state would have shifted eastward, away from the problems and interests of the Holy Roman Empire. Yet the rulers of Prussia did not shrink from a policy that was likely to have such far-reaching consequences. They never contemplated sacrificing the advantage that their state would gain from an enlargement of its resources in order to assume the role of unifiers of their nation. Such a political attitude would have been an anachronism during the age of princely absolutism in central Europe. It was not design but accident that before long led to the abandonment by Prussia of most of its Polish possessions and that thereby allowed it to continue to play a vital part in the affairs of Germany.

Contingent results of the Polish partition

The cultural scene. Whereas in England the great literary epoch of Queen Elizabeth I had coincided with commercial and naval expansion, and in France the golden age of classicism had added lustre to the military glory of Louis XIV, German arts and letters flourished amid tiny principalities and somnolent towns that could only envy the powerful national monarchies west of the Rhine. Moreover, whereas in France and England, where public opinion could exert a significant influence on government, the debate over issues of state and society was conducted with a vigour that reflected its importance, in Germany the debate was bound to remain purely theoretical. No Voltaires, Rousseaus, or Burkes were likely to emerge out of such an environment. The thinkers of central Europe tended to emphasize introspection and spirituality. Culture became an escape from the narrow world of princely absolutism. Intellectual energies that could not reform the community fought to emancipate the individual through self-purification and self-perfection.

This was the background of German Idealism, a philosophic movement seeking to liberate ethics and aesthetics

German Idealism

from the shackles of empirical knowledge. Armed with the weapons of Kantian thought, it attempted to prove that there was a realm of experience lying beyond the categories of scientific investigation, the realm of the good, the true, and the beautiful. There were realities of the spirit and the mind, in other words, that were inaccessible to the practicality of the British Empiricists or the intellectualism of the French Rationalists. The disciples of Idealism hoped to transcend the barriers created by nation, class, and religion. They spoke in the name of humanity as a whole, which manifested its underlying harmony through the infinite variety of its political, social, and theological categories. Gotthold Ephraim Lessing pleaded for religious toleration on the basis of a common system of ethical values to which all men of good will could subscribe. Johann Gottfried Herder preached that the unique character and meaning of each culture contributed to the richness of common humanity that defied state boundaries. Johann Joachim Winckelmann deified the classical ideal of beauty that he found in Greek art as an eternal standard, immune to the vicissitudes of time and history. These were views that offered an escape from the narrowness of everyday life. Men who found no scope for their talents in the petty world of princely authority could turn to the liberating spirituality of the Idealist philosophy. Thought in central Europe gradually acquired a metaphysical coloration that distinguished it from the more robust Pragmatism of philosophy in the west. It was during the second half of the 18th century that the Germans began to consider their country "the land of thinkers and poets."

The introspection of arts and letters

The literary revival of the age displayed the same quality of introspective Idealism as the philosophic movement. Johann Wolfgang von Goethe, the greatest genius of German letters, willingly accepted the existing system of civic and social values. He regarded the disunity of his nation as an expression of its historic character, and he defended the authority of the petty princes as an instrument of good government. He urged his countrymen to seek greatness not in collective action but in individual perfectibility. After a period of youthful rebellion against traditional canons of literary propriety, he turned to a classicism in which a serene acceptance of life harmonized with his own sympathy for the established order. Friedrich Schiller, a man of more turbulent temperament, felt a sense of resentment against political injustice and weakness. In his plays and poems there are occasional outbursts of indignation and appeals for reform. Yet there is also a pessimistic mood of resignation induced by the burden of civic ineffectualness that history had imposed on his people. Ultimately, he too sought refuge from the world in the poet's private vision. The *Sturm und Drang* ("storm and stress") was a movement of literary innovation through which a group of young writers in the last decades of the 18th century sought to throw off the yoke of accepted standards of composition. But it remained confined to problems of prosody and taste, refusing to grapple with political or social issues.

The cultural achievements could not alter the harsh realities of national fragmentation and princely autocracy. They supported, however, the ideals of rational reform and social progress that the Enlightenment had introduced throughout the Continent. In Germany as elsewhere the 18th century became the age when the monarchical principle advanced the loftiest justification of its claim to power. The authority of the prince, so the argument went, was to be exercised not for his private advantage or gratification but for the greatness of his state and the welfare of his people. His power had to be unrestricted so that his benevolence might be unlimited. Absolute government was the only effective instrument for achieving the general good. Impressed by the scientific discoveries and material advances that they saw about them, men began to believe that the prejudices and injustices that had plagued society would gradually disappear before the steady march of reason.

Enlightened reform and benevolent despotism. The main source of enlightened reform was to be the crown, but many well-intentioned people of means and education also began to apply a new standard of conduct in their dealings with their fellow man. This change in attitude was apparent in the decline of religious resentments and discriminations. Never before had the relationship between Catholics and Protestants among the well-to-do classes of central Europe been as free of rancor as on the eve of the French Revolution. It was at this time also that the Jews first began to emerge from the isolation to which a deep-seated intolerance had consigned them. The idea of assimilation held out to them the prospect of escape from the ghetto on the condition that they identify themselves in thought, speech, and attitude with the Christian society in which they lived. That prospect was to attract the Jewish minority in Germany more and more during the next 150 years. Religious toleration, however, was not the only article of faith of the Enlightenment. Its vision of a happier future included the reform of education, the abolition of poverty, the alleviation of sickness, and the elimination of injustice. Men of good will established schools, founded orphanages, built hospitals, improved farming methods, modernized industrial techniques, and tried to raise the standard of living of the masses. While the hopes of the enlightened reformers of the 18th century far outstripped their accomplishments, the practical results of their efforts should not be underestimated.

Growing religious toleration

According to the doctrines of benevolent despotism, however, the chief instrumentality for the improvement of society was not private philanthropy but government action. The state had the primary responsibility for preparing the way for that golden age which, in the opinion of many intellectuals, awaited mankind. The extent to which official policy conformed to rationalist theory depended, in central Europe as elsewhere, on the personality and ability of the ruler. Both of the leading powers of the Holy Roman Empire followed the teachings of benevolent despotism, but with substantially different results. The emperor Joseph II, a well-meaning though doctrinaire reformer, attempted to initiate a revolution from above against the opposition of powerful forces that continued to cling to tradition. In the course of a single decade he tried to centralize the government of his far-flung domains, reduce the influence of the church, introduce religious toleration, and ease the burden of serfdom. His uncompromising program of innovation, however, alienated the landed aristocracy, whose support was essential for the effective operation of the government. The emperor encountered mounting unrest that did not end until his death in 1790 and the subsequent abandonment of most of the reforms that he had promulgated. Frederick the Great was more successful as an enlightened autocrat, but only because he was more cautious. His reorganization of the government was not as drastic, his belief in religious toleration remained less profound, and his assistance to the peasants did not go beyond a prohibition against the absorption of their holdings by the nobility. He invited settlers to cultivate reclaimed lands, and he encouraged entrepreneurs to increase the industrial capacity of Prussia. Among his most important accomplishments, although it was not completed until after his death, was the Prussian Legal Code, which defined the principles and practices of an absolute government and a corporative society. Yet Frederick was also convinced that the Prussian landed noblemen, the Junkers, were the backbone of the state, and he continued accordingly to uphold the alliance between crown and aristocracy on which his kingdom had been built.

Attempts at social and political reform

The achievements of benevolent despotism among the minor states of the Holy Roman Empire varied considerably. Some princes employed their inherited authority in a serious effort to improve the lot of their subjects. Charles Frederick of Baden, for example, devoted himself to the improvement of education in his margravate, and he even abolished serfdom, though without eliminating manorial obligations. Carl August of Saxe-Weimar-Eisenach was a hard-working administrator of his small Thuringian principality, whose capital he transformed into the cultural centre of Germany. Charles Eugene of Württemberg, on the other hand, led a life of profligacy and licentiousness in defiance of protests by the estates

of the duchy. Frederick II of Hesse-Kassel was another princely prodigal whose love of pleasure impoverished his subjects and forced his soldiers into mercenary service for England. The record of enlightened autocracy in central Europe was as uneven as in western Europe. Yet the ideas of the Enlightenment even at their best were unable to transform the basis of political life in the Holy Roman Empire. They could palliate, reform, and improve. But they could not alter a system of particularistic sovereignty and absolutistic authority resting on a hierarchical structure of society. They could not become an instrument of national consolidation or representative government. Only some great creative disruption of existing civic institutions could break through the crust of habit and tradition sanctified by history. Germany lacked the vital energies required for a process of political reconstruction. The galvanizing forces of rejuvenation and regeneration were to come from the outside.

The French Revolutionary and Napoleonic era. The French Revolution, transforming the Bourbon kingdom into a constitutional state, aroused intense excitement east of the Rhine. Most German intellectuals were at first in sympathy with the new order in France, hoping that the defeat of royal absolutism in western Europe would lead to its decline in central Europe as well. The princes, on the other hand, were from the outset fearful of the Revolution, which they regarded as a serious danger; for the example of unpunished insubordination by the French might encourage demands for reform among the Germans. The result was a growing hostility between the government in Paris and the rulers of the Holy Roman Empire, which led in the spring of 1792 to the outbreak of war (the War of the First Coalition, 1792–97). The immediate occasion of the conflict was a quarrel over the rights of German princes with holdings in France and over the propagandistic activities of French émigrés in Germany. But the underlying cause was the clash of two incompatible principles of authority divided by profound differences regarding the nature of political and social justice. The course of hostilities soon revealed that the civic ideals and military tactics of the Revolution were more than a match for the decrepit Holy Roman Empire. After 1793 the left bank of the Rhine remained under the control of France, and for the next 20 years its inhabitants were governed from Paris. Yet there is no evidence that they were dissatisfied with French rule or at least that they strongly opposed it. Devoid of a sense of nationalism and accustomed to submission to authority, they accepted their new status with the same equanimity with which they regarded a succession to the throne or a change in the dynasty. The Prussians, moreover, discouraged by defeats in the west and eager for Polish spoils in the east, concluded a separate peace at Basel in 1795 by which they in effect recognized the French acquisition of the Rhineland. The Austrians held out two years longer, but the brilliant successes of the young Napoleon Bonaparte forced them to accept the loss of the left bank in the Treaty of Campo Formio (October 17, 1797).

War of the First Coalition

War of the Second Coalition

End of the Holy Roman Empire. The peace proved short-lived, however, for at the end of 1798 a new coalition directed against France was formed (the War of the Second Coalition, 1798–1802). This time Prussia remained neutral. Frederick William III, a conscientious and modest but ineffectual ruler, was distinguishable from his father by private morality rather than political skill. The government in Berlin drifted back and forth, dabbling in minor economic and administrative reforms without introducing a significant improvement in the structure of the state. A decade of neutrality was frittered away, while the army commanders rested on the laurels of Frederick the Great. Austria, on the other hand, played the same leading role in the War of the Second Coalition as in the War of the First Coalition, and with the same unfortunate result. The French victories at Marengo (June 14, 1800) and Hohenlinden (December 3, 1800) forced Emperor Francis II to agree to the Treaty of Lunéville (February 9, 1801), which confirmed the cession of the Rhineland. More than that, those rulers who lost their possessions on the left bank under the terms of the peace were to receive compensation elsewhere in the empire. In order to carry out this redistribution of territory, the imperial Diet entrusted a committee of princes, the Reichsdeputation, with the task of drawing a new map of central Europe. The major influence over its deliberations, however, was exercised by France. Napoleon had resolved to utilize the settlement of territorial claims to achieve a fundamental alteration in the structure of the Holy Roman Empire. The result was that the Final Recess (*Hauptschluss*) of the Reichsdeputation of February 1803 marked the end of the old order in Germany. In their attempt to establish a chain of satellite states east of the Rhine, the French diplomats brought about the elimination of the smallest and least viable of the political components of central Europe. But thereby they also furthered the process of national consolidation, since the fragmentation of civic authority in the empire had been a mainstay of particularism. That it was not Napoleon's intention to encourage unity among his neighbours goes without saying. Yet he unwittingly prepared the way for a program of centralization in Germany that helped to frustrate his own plans for the future aggrandizement of France.

The chief victims of the Final Recess were the free cities and the ecclesiastical territories. They fell by the dozens. Too weak to be useful allies of Napoleon, they were destroyed by the ambition of their French conquerors and by the greed of their German neighbours. They could still boast of their ancient history as sovereign members of the Holy Roman Empire, but their continued existence had become incompatible with the establishment of effective government in central Europe. The principal heirs to their holdings were the larger secondary states. To be sure, Napoleon could not keep Austria and Prussia from making some gains in the general scramble for territory that they had helped make possible. But he was especially solicitous for the welfare of those German rulers, most of them in the south, who were strong enough to be valuable vassals but not strong enough to be potential threats. Bavaria, Württemberg, Baden, Hesse-Darmstadt, and Nassau were the big winners in the competition for booty that had been the main object of the negotiations. Napoleon's strategy had been in the classic tradition of French diplomacy, the tradition of Richelieu and Mazarin. The princes had been pitted against the emperor in order to enhance the role that Paris could play in the affairs of central Europe. Yet neither the states that gained nor those that lost territory felt any resentment at being used as pawns in a political game to promote the interests of a foreign power. Whatever objections they raised against the settlement of 1803 were based on expediency and opportunism. The most serious indictment of the old order was that in the hour of its imminent collapse none of the rulers made the least attempt to defend it in the name of the general welfare of Germany.

The Final Recess was the next to the last act in the fall of the Holy Roman Empire. The end came three years later. In 1805 Austria joined the third coalition of Great Powers determined to reduce the preponderance of France (the War of the Third Coalition, 1805–07). The outcome was even more disastrous than its participation in the first and the second coalitions. Napoleon forced the main Habsburg army in Germany to surrender at Ulm (October 17, 1805); then he descended on Vienna, occupying the proud capital of his enemy; and finally he inflicted a crushing defeat on the combined Russian and Austrian armies at Austerlitz in Moravia (December 2, 1805). Before the year was out Francis II had been forced to sign the humiliating Treaty of Pressburg (December 26), which signified the end of the dominant role his dynasty had played in the affairs of cental Europe. He had to surrender his possessions in western Germany to Württemberg and Baden, and the province of Tirol to Bavaria. Napoleon's strategy of playing off princely against imperial ambitions had proved a brilliant success. The rulers of the secondary states in the south had supported him in the war against Austria, and in the peace that ensued they were richly rewarded. Not only did they share in the booty seized from the Habsburgs but they

War of the Third Coalition

were permitted to absorb the remaining free cities, petty principalities, and ecclesiastical territories. Finally, in an assertion of the rights of full sovereignty, the rulers of Bavaria and Württemberg assumed the title of king, while the rulers of Baden and Hesse-Darmstadt contented themselves with the more modest rank of grand duke. The last vestiges of the imperial constitution had now been destroyed, and central Europe was ready to receive a new form of political organization reflecting the power relationship created by the force of arms.

Confederation of the Rhine

In the summer of 1806, 16 of the secondary states, encouraged and prodded by Paris, announced that they were forming a separate association to be known as the Confederation of the Rhine. Archbishop Karl Theodor von Dalberg was to preside over the new union as the "prince primate," while future deliberations among the members were to establish a college of kings and a college of princes as common legislative bodies. There was even talk of a "fundamental statute" that would serve as the constitution of a rejuvenated Germany. Yet all these brave plans were never more than a facade for the harsh reality of alien hegemony in central Europe. Napoleon was proclaimed the "protector" of the Confederation of the Rhine, and a permanent alliance between the member states and the French Empire obliged the former to maintain substantial military forces for the purpose of mutual defense. There could be no doubt whose interests these troops would serve. The secondary rulers of Germany were expected to pay a handsome tribute to Paris for their newly acquired sham sovereignty. On August 1 the confederated states proclaimed their secession from the empire, and a week later, on August 6, 1806, Francis II announced that he was laying down the imperial crown. The Holy Roman Empire thus came officially to an end after a history of a thousand years.

Period of French hegemony in Germany. The long conflict between emperors and princes in Germany had concluded with the triumph of the latter. Yet the victors soon discovered that instead of achieving independence they had merely exchanged one master for another. Indeed, they were more subordinate now to the wishes of Paris than they had been to those of Vienna. Napoleon gradually induced or forced all the states of Germany except Austria and Prussia, 36 states in all, to join the confederation. The inclusion of the two leading powers of central Europe might have proved troublesome for him or even dangerous. The participation of the secondary states, on the other hand, provided him with reliable mercenaries who owed him too much and feared him too much to oppose his wishes. He was free to do as he liked in the region between the Rhine and the Elbe. In order to enforce his embargo on continental trade with England, he annexed the entire coastline along the North Sea. When that was not enough, he added Lübeck on the Baltic to the French Empire. He carved out the Kingdom of Westphalia for his brother Jérôme and the Grand Duchy of Berg for his brother-in-law Joachim Murat. He was the undisputed master of all of middle Germany.

After the formation of the Confederation of the Rhine there was only one state in central Europe that had not yet been forced to submit to France. But the leaders of Prussia hesitated and wavered in their policy until they lost the opportunity of profiting from the War of the Third Coalition. Had they joined Austria and Russia against Napoleon, they might have kept him from gaining hegemony over Germany. Or had they become the allies of Napoleon, they might have established a sphere of influence in the region north of the Main River. As it was, they waited until they fell between two stools. They finally declared war against the French in October 1806, after Austria had been forced to surrender, Russia had decided to retreat, and the secondary states had become the vassals of Paris. Yet public opinion in the Prussian capital remained confident that the army of Frederick the Great would prove a match for the conqueror of Europe.

The subjection of Prussia

The result of such self-deception was a military disaster of unparalleled magnitude. In the two simultaneous battles of Jena and Auerstädt (October 14, 1806) the Hohenzollern armies were completely routed, and the road to Berlin lay open before the French invaders. The city was occupied on October 27.

More disastrous than the military defeat, however, was the moral collapse of a state that had taught its citizens that obedience to authority was the supreme political virtue. The civilian population never thought of offering resistance to the advancing enemy. Even many army officers were so disheartened by Napoleon's success that they surrendered one fortified position after another without a fight. Frederick William III had to pay a terrible price for the policy of his ancestors, who had built efficient government at the expense of civic initiative. He tried to hold out in East Prussia, hoping that the Russian armies, which were still at war with Napoleon, would help him regain the rest of his kingdom. But when in July 1807 Alexander I concluded peace with France at Tilsit, the unfortunate Hohenzollern had no choice but to follow suit. The treaty that he was forced to sign was a catastrophe. Prussia lost almost half its territory and population, including most of the Polish possessions in the east as well as all of the territories west of the Elbe. Subsequent agreements, moreover, imposed a heavy indemnity, a military occupation, and a reduction in the size of the army. The proud monarchy of Frederick the Great had been reduced to a secondary state in Germany.

Adapted from R. Treharne and H. Fullard (eds.), *Muir's Historical Atlas: Ancient, Medieval and Modern*, 9th ed. (1964); George Philip & Son Ltd., London



Germany in 1807 after reconstruction by Napoleon.

Napoleonic reforms and the birth of German nationalism

Central Europe remained under the dominant influence of France for more than a decade. That influence was at first limited and indirect, then pervasive and overpowering. Yet it was during this period of alien preponderance that Germany for the first time felt the stirrings of liberalism and nationalism. The regions that had become part of the French Empire experienced at first hand the advantages of efficient centralized government in which equality before the law and freedom of opportunity were accepted principles. Those states that retained a pseudo-independence as satellites of Napoleon, moreover, sought to imitate the example of their master, partly in order to gain his favour, partly in order to emulate his success.

One government after another began to remove religious disabilities, relax economic restrictions, eliminate servile obligations, and centralize administrative functions. Above all, constitutional rule and popular representation ceased to seem utopian to men of property and education who had witnessed the stirring events of the years since 1789. The French hegemony also led to the birth of nationalism in central Europe. For one thing, the achievement of political unity became a distinct possibility, once the territorial fragmentation of the Holy Roman Empire had come to an end. The presence of foreign invaders, furthermore—arrogant, overbearing, and avaricious—aroused among the Germans a sense of nationality that they had never felt in the tranquil days of the old order. Finally, an example of what great deeds the love of fatherland could inspire lay before all who admired or envied the triumphs of Napoleon. The ideal of cosmopolitan individualism that had been generally accepted in the 18th century began to give way before a growing consciousness of national identity. Yet the fact that the concepts of constitutional freedom and national unity were not indigenous but arose in response to foreign domination had an important effect on the form they assumed in central Europe.

Every German state felt the influence of the new principles of government and economy that the period of French hegemony had introduced. But nowhere was that influence more profound or fruitful than in Prussia. For only in the hour of deepest humiliation did the Hohenzollern kingdom finally make an effort to adapt its structure to the changing political and social conditions that it had stubbornly ignored during the years of greatness. Between 1806 and 1813 the statesmen in Berlin initiated a revolution from above in order to transform a rigid despotism into a popular monarchy supported by the loyalty of a free citizenry. Out of the disasters of Jena and Tilsit emerged a group of gifted reformers who sought to prepare the way for the regeneration of their country. The leading figures in this movement for civic reconstruction were the civil servants Karl vom Stein and Karl August von Hardenberg and the military commanders Gerhard von Scharnhorst and August Neithardt von Gneisenau. Among their most important achievements was the abolition of serfdom, a measure designed to create dedicated citizens out of human beasts of burden. Yet while it gave the peasant personal freedom, the government failed to provide him with economic independence. Most of the land remained in the hands of the aristocracy, which therefore continued to dominate the countryside politically as well as socially. More successful was the law establishing municipal self-government. Thereafter, the cities of the kingdom were to be administered by officials chosen not by the central bureaucracy but by the propertied inhabitants of the cities themselves. The autonomy of the urban communities, it was hoped, would help train a politically conscious and active middle class. The most effective reforms, however, were those introduced in the armed forces. After the officers who had shown themselves incompetent during the war were dismissed or retired, the high command carried out a thoroughgoing reorganization of the military system. Discipline became more humane, promotion was made a reward for merit, the method of recruitment was improved, and the training in tactics was modernized. Most important of all, the army's leaders sought to instill in the soldiers a new spirit rooted in inner conviction rather than unquestioning obedience. Defeat had changed Prussia from a garrison state into a centre of political and intellectual ferment.

The Prussian reformers: Stein and Hardenberg

Liberalism and nationalism became increasingly vehement in Germany as the burden of French domination grew progressively heavier. The financial sacrifices occasioned by the subordination to Napoleon reinforced the personal resentments aroused by his ruthless statecraft. Before long a network of secret organizations had sprung up in central Europe seeking the expulsion of the foreign invaders. Yet it would be a mistake to think that all Germans regarded the hegemony of France as an unmitigated evil. There were in fact wide differences of opinion among them. The rulers of the secondary states and their supporters in the army and the bureaucracy saw in Napoleon the instrumentality of their new importance. Many reformers in the south—the Bavarian statesman Maximilian von Montgelas, for example—believed that only French influence had made possible the modernization of government in Germany. Some publicists continued to argue, moreover, that the political disunity of central Europe was a natural result of its historic experience and reflected its essential character. To be sure, those who opposed alien domination outnumbered those who accepted it. But even among the former there was no agreement regarding the future political structure of the nation. Many of them dreamed of a liberal and united fatherland that would take its place among the great powers of Europe. Others were willing to settle for a loose association of governments, similar to the Confederation of the Rhine, which could safeguard the interests of the secondary states against Prussia and Austria. Still others hoped for a complete restoration of the old order in which they had grown up and to which they longed to return. And then there were the broad masses of the population of central Europe, exploited, illiterate, and uninformed. They remained by and large indifferent to the crosscurrents of political thought, seeking nothing more than an improvement in their standard of living and the preservation of their way of life. Germany was beginning to move toward new norms of civic and social value, but the transformation of political attitudes was gradual and intermittent.

The defeat of Austria

That the growth of the ideals of unity and freedom was slow became apparent during the first serious effort to throw off the yoke of foreign domination in central Europe. The Austrian government concluded in 1809 that the reverses that Napoleon had been encountering in Spain presaged a general uprising against the French hegemony on the Continent. The result was an ill-fated attempt at a war of liberation, in which the Habsburg troops challenged Napoleon for the fourth time, only to go down in defeat once again. Appeals from Vienna to the people of Germany found little response except in the Tirol and among a few nationalist hotspurs in the north. The princes refused to risk French wrath until they could be sure of ultimate victory, while their subjects refused to rise against French oppression without princely approval. The result was that the war in central Europe, unlike the one in the Iberian Peninsula, was waged primarily by regular forces rather than by guerrilla bands. Archduke Charles gained important successes for the Austrian Army at Aspern and Essling (May 21–22, 1809), an indication that the strategic mastery of the French was drawing to a close. But at Wagram (July 5–6) Napoleon was able to work the last of his military miracles. Vienna had to sue for peace once more, the Treaty of Schönbrunn (October 14) ceding Salzburg to Bavaria, West Galicia to the Grand Duchy of Warsaw, and the Adriatic coastland to France. The defeat finally persuaded the Emperor, who had exchanged the title Francis II of the Holy Roman Empire for Francis I of the Austrian Empire, that resistance would be as futile in the future as it had been in the past. He therefore adopted a policy of collaboration with France signallized by the marriage of his daughter Marie-Louise to Napoleon. Germany continued to languish in the grip of foreign domination.

The wars of liberation. A new struggle for liberation opened three years later with the defeat of Napoleon's *grande armée* in Russia. As the tsarist armies began to cross their western frontiers in December 1812, the crucial question became what reception they would find among the rulers and the inhabitants of central Europe. The first state to cut its ties to Paris was Prussia. It was not the King, however, but one of his generals, Ludwig Yorck von Wartenburg, who decided on his own initiative to cooperate with the Russians. Only hesitatingly and fearfully did Frederick William III then agree in February 1813 to a war against France, although public opinion in his kingdom greeted the outbreak of the conflict with enthusiasm. The other rulers of central Europe refused initially to follow the Prussian example. The members of the Confederation of the Rhine were still convinced of Napoleon's invincibility, while Austria pre-

ferred to see the combatants exhaust each other to the point at which it could play the role of mediator and arbiter. The foreign minister in Vienna, Clemens Lothar von Metternich, was afraid that the hegemony of France in central Europe might be replaced by that of Russia. He tried, therefore, to pursue a strategy of armed neutrality, hoping that he could persuade the opposing sides to accept a compromise by which an equilibrium would be maintained between Alexander I and Napoleon. This plan failed because of the obstinacy of the latter, who feared that concessions in foreign affairs would weaken his control over internal politics in France. The upshot was that in August 1813 Austria entered the conflict on the side of Russia and Prussia, and the balance of military power shifted in favour of the anti-French coalition. The faith of the secondary states in Napoleon's star began to weaken and Bavaria became the first member to secede from the Confederation of the Rhine (October 8). One great allied victory would now suffice to bring all of Germany into the struggle against France.

The anti-French alliance of Russia, Prussia, and Austria

That victory came on October 16–19, 1813, at the Battle of Leipzig. After four days of bitter fighting the French Army was forced to retreat, and its domination of central Europe was finally at an end. Before the year was out, Napoleon had withdrawn across the Rhine. Of all his conquests in Germany, only the left bank was still under the effective control of Paris. The Confederation of the Rhine promptly collapsed, as its members rushed to go over to the winning side before it was too late. The Rhineland was also reconquered early in 1814, after the allies had launched their invasion of France. In the course of the spring the capture of Paris, the restoration of the Bourbons, and the conclusion of peace in the first Treaty of Paris (May 30) ended the war of liberation except for the episode of the Hundred Days, when Napoleon briefly returned to power and was ultimately and finally beaten at Waterloo. The western frontier of central Europe was to remain essentially the same as at the time of the initial outbreak of hostilities more than 20 years before. New state boundaries within Germany would still have to be determined, to be sure, and the problem of a new political organization of the nation awaited the victorious statesmen. But the period of foreign hegemony was over at last. The rulers of central Europe, relying partly on the forces of innovation, partly on those of tradition, had succeeded in freeing themselves from alien domination. Now they had to decide what use they would make of their freedom. Would they create a new polity of unity and liberty, which many reformers demanded, or would they re-establish the old order of absolutism and particularism, which the conservatives advocated? As the statesmen began to gather in Vienna in the fall of 1814 to restore peace to a continent ravaged by two decades of war, they pondered the problem of devising an enduring form of government for Germany.

Results of the Congress of Vienna. The men who, in the course of the nine months from September 1814 to June 1815, redrew the map of Europe were diplomats of the old school. Francis I and Prince Metternich of Austria, Frederick William III and Prince Hardenberg of Prussia, Alexander I of Russia, Viscount Castlereagh of England, Prince Talleyrand of France, and the representatives of the secondary states were all intellectual heirs of the 18th century. They feared the principles of the French Revolution, they scorned the theories of democratic government, and they opposed the doctrines of national self-determination. But they also recognized that the boundaries and governments of 1789 could not be restored without modification or compromise. There had been too many changes in attitudes and loyalties that the rigid dogmas of legitimism were powerless to undo. The task before the peacemakers was thus the establishment of a sound balance between necessary reform and valid tradition capable of preserving the tranquillity that the Continent desperately needed. The decisions regarding Germany reached during the deliberations in Vienna followed a middle course between innovation and reaction, avoiding extreme fragmentation as well as rigid centralization. The Confederation of the Rhine was not maintained, but neither was the Holy Roman Empire restored. Although the reforms introduced during the period of foreign domination were partly revoked, the practices of enlightened despotism were not entirely re-established. Despite the bitter complaints of unbending legitimists and the dire predictions of disappointed reformers, the peacemakers succeeded in creating a new political order in central Europe that endured for half a century. The long years of war and unrest that had convulsed the Continent during the era of the French Revolution and Napoleon were followed by even longer years of stability and tranquillity.

Adapted from R. Treharne and H. Fullard (eds.) *Muir's Historical Atlas: Ancient, Medieval and Modern*, 9th ed. (1964); George Philip & Son Ltd., London



The German Confederation, 1815.

The Germany that emerged in 1815 from the Congress of Vienna included 39 states ranging in size from the two great powers, Austria and Prussia, through the minor kingdoms Bavaria, Württemberg, Saxony, and Hanover; through smaller duchies such as Baden, Nassau, Oldenburg, and Hesse-Darmstadt; through tiny principalities such as Schaumburg-Lippe, Schwarzburg-Sondershausen, and Reuss-Schleiz-Gera; to the free cities Hamburg, Bremen, Lübeck, and Frankfurt am Main. The new boundaries in central Europe bore little resemblance to the bewildering territorial mosaic that had been maintained under the Holy Roman Empire, but there were still many fragments, subdivisions, enclaves, and exclaves, too many for the taste of ardent nationalists. Yet the overall pattern of state frontiers represented a significant improvement over the chaotic patchwork of sovereignties and jurisdictions that had characterized the old order. The peacemakers not only created more integrated and viable political entities but also altered the role that these entities were to play in the affairs of the nation. Without design or even awareness on the part of Frederick William III, his kingdom of Prussia assumed a pivotal position in Germany. The victorious powers, on guard against a revival of French aggression, decided to make Berlin the defender of the western boundary of central Europe. The Rhineland and Westphalia, a region destined to develop into the greatest industrial centre on the Continent, became Hohenzollern provinces. More than that, the King agreed at the urging of Alexander I to cede the bulk of his Polish possessions to Russia in return for a substantial part of Saxony. Prussia, which at the end of the 18th century had

The new Germany of 39 states

been in the process of becoming a binational state, was thrust back into Germany and given a strategic position on both frontiers of the nation. The centre of gravity of Austria, on the other hand, shifted eastward. Francis I had decided to abandon the historic role of his state as protector of the Holy Roman Empire against the Bourbons for the sake of greater geographic compactness and military defensibility. The possessions in southern and western Germany were surrendered along with the Austrian Netherlands in return for Venetian territory on the Adriatic. The Habsburg empire thus became less German in composition and outlook, as its focus shifted in the direction of Italy and the Balkans. The consequences of this territorial rearrangement were to be far-reaching.

Germanic Prussia and the polyglot Austrian Empire

THE AGE OF METTERNICH AND THE ERA OF UNIFICATION: 1815–71

Reform and reaction. In place of the Holy Roman Empire the peacemakers of the Congress of Vienna had established a new organization of states of central Europe, the German Confederation. This was a loose political association in which most of the rights of sovereignty remained in the hands of the member governments. There was no central executive or judiciary, only a federal Diet meeting in Frankfurt am Main to consider common legislation. The delegates who participated in its deliberations were representatives appointed by and responsible to the rulers whom they served. The confederation was in theory empowered to adopt measures strengthening the political and economic bonds of the nation. In fact it remained a stronghold of particularism, unwilling to sacrifice local autonomy in order to establish centralized authority. It was designed essentially to defend the interests of the secondary states and the Habsburgs. The former, jealously guarding the independence and importance they had gained during the period of French hegemony, were opposed to any reform that might limit their sovereignty. The latter believed that only a decentralized form of political union in Germany would give them enough freedom of action to pursue their non-German objectives. The confederation was thus from the outset an ally of localism and traditionalism. To the nationalists, whose hopes had risen so high during the war of liberation, it seemed to be an instrument of blind reaction. Yet the truth is that the confederal system established in 1815 accurately reflected the slow development of civic consciousness and economic integration in central Europe. The militant reformers who demanded the centralization of government were a vocal but small minority. The lower classes accepted the territorial and constitutional decisions of the Congress of Vienna without a murmur of protest. The weakness of the peace settlement was not its failure to embody present realities but its inability to adjust to future changes. What had been a reasonable adaptation to the political needs of an agrarian and rural society became a hopeless anachronism 50 years later in the age of factories and railroads. This was the fatal flaw in the German Confederation.

Yet the reform movement that had begun under the impact of the French hegemony did not end with the downfall of Napoleon. It continued to exert influence over affairs of state for another few years, before the forces of authoritarianism and particularism crushed it. That influence was strongest in southern Germany, where the political example of western Europe had made the deepest impression. There, many civil servants, court officials, army officers, and even aristocratic landowners came to believe that the future of the state depended on its readiness to reform civic institutions in accordance with liberal theories. In the years following Waterloo one government in the south after another promulgated a constitution, Bavaria and Baden in 1818, Württemberg in 1819, and Hesse-Darmstadt in 1820. These constitutions established representative assemblies, elected by the propertied citizens, whose assent was required for the enactment of legislation. Their purpose was not only to win for the crown the support of the educated classes of society but also to engender a sense of unity in a heterogeneous population that still had diverse allegiances and traditions. To the north there were also persistent echoes of the reform movement.

The followers of Karl vom Stein were still influential in the councils of state, and Frederick William III of Prussia at first seriously considered ways of fulfilling the promise he had made in 1815 to establish constitutional government. The agitation for political reorganization was loudest, however, among university students, who formed patriotic groups known as *Burschenschaften*. They demanded the abandonment of the confederal system, the establishment of greater unity, and the achievement of national power. Gathering in 1817 at the Wartburg, a castle near Eisenach, they listened to veiled denunciations of the existing order and consigned to flames various symbols of traditional authority. The rulers of Germany began to stir uneasily at this bold display of defiance of legitimate government.

The chief strategist of the forces hostile to reform was Metternich. Not only did he reject the teachings of liberalism and nationalism in principle, but, as the leading statesman of the Habsburg empire, he recognized that the establishment of centralized authority in Germany would seriously impede the policies his government was pursuing in Hungary, Italy, and the Balkans. When on March 23, 1819, an unbalanced student, Karl Ludwig Sand, assassinated the conservative playwright and publicist August von Kotzebue, Vienna persuaded the princes of the German Confederation that they were facing a dangerous attempt to overthrow the established order in central Europe. The result was a series of repressive measures called the Carlsbad Decrees, which the federal Diet adopted on September 20, 1819. General censorship was introduced, and the *Burschenschaften* were outlawed. This first major success of the conservative counteroffensive had an important effect on the struggle within the state governments between the advocates and the opponents of reform. In Prussia, the liberal members of the ministry were forced to resign, and the plan to promulgate a constitution for the kingdom was rejected. This shift to the right by Berlin encouraged authoritarian tendencies among the secondary states of the north, which soon abandoned their own constitutional projects. By the end of 1820 the reform movement, which had begun some 15 years before, came to a complete halt. It had succeeded in altering the political and economic structure of society, but it had been unable to establish a tradition of liberal government and national loyalty in central Europe. The forces of particularism and legitimism, deriving their chief support from the landowning nobility and the conservative peasantry, remained strong. The foundation of bourgeois civic consciousness and material prosperity on which England and France had built their representative institutions was still lacking beyond the Rhine. The ideal of free government was introduced in Germany not as the fruit of an industrial or a political revolution but as the imitation of a foreign example and the reaction against a foreign oppression.

Repression and the Carlsbad Decrees

The established order was once again threatened briefly in the wake of the July days of 1830 in France. The news that there had been a successful insurrection against the Bourbons in Paris had an electrifying effect throughout the Continent. In central Europe there were sympathetic uprisings in some of the secondary states of the north. The rulers of Brunswick, Saxony, Hanover, and Hesse-Kassel, seeking to forestall more extreme demands, agreed to promulgate liberal constitutions. A mass meeting of southern radicals at Hambach Castle in the Palatinate (May 1832), moreover, expressed its approval of national unification, republican government, and popular sovereignty. A group of militant students even launched a foolhardy attempt to seize the city of Frankfurt am Main, dissolve the federal Diet, and proclaim a German republic. The effect of such harebrained schemes was predictable. As the princes of the German Confederation gradually recovered from their initial fear of the revolutionary movement, they began to oppose with increasing vigour plans to alter the existing system of government. Again, Metternich took the lead in the effort to crush liberalism and nationalism. Under his direction the fed-

Effects of the July Revolution of 1830

eral Diet adopted additional repressive measures reinforcing the position of the crown in state politics, limiting the power of the legislature, restricting the right of assembly, enlarging the authority of the police, and intensifying the censorship. Within a few years the opposition had been subdued, and the German Confederation could continue to vegetate in its cozy provincialism. Not until the middle years of the century did a new and more violent outburst of political disaffection shake the foundations of the system of authority that the Congress of Vienna had erected.

Evolution of parties and ideologies. Although the critics of the established order could be defeated, they could not be silenced. The struggle between the supporters and the adversaries of the existing form of government led to the emergence of a rudimentary party system in the German Confederation. In the legislative assemblies of the secondary states the proponents of reform began to meet, plan, organize, and propagandize. The defenders of legitimism were thereby forced in turn to concert their strategy and to publicize their program. Even in Prussia and Austria, where there were as yet no constitutions or parliaments, political criticism could be expressed obliquely through clubs, meetings, newspapers, pamphlets, and petitions. The result was the gradual development of amorphous civic associations held together by common convictions regarding the nature of state and society. These primitive groupings were only the raw material out of which disciplined political parties were slowly fashioned in the course of the century. They still lacked a clear sense of purpose and the systematic propagation of belief characteristic of a fully mature system of parliamentary politics. Yet they became the instrumentalities by which disaffected groups in the community could express their opposition to the established order. They reflected the fact that the civic attitudes of the period of the Restoration were no longer those of the age of enlightened despotism. There were men in central Europe now who refused to submit without question to princely authority, seeking freedom only in the inner recesses of the soul. The change in the form of economy and the structure of society produced by the beginnings of industrialization led to an alternation in the system and organization of politics.

The liberals and moderates

The most important opponents of legitimism and particularism were the liberals and moderates. Deriving their support primarily from industrialists, merchants, financiers, mineowners, railroad promoters, civil servants, and university professors, they represented the opposition of the well-to-do bourgeoisie to a form of government in which an aristocracy of birth rather than of talent predominated. Their political principles favoured a monarchical system of authority, but the crown was to share its powers with a parliament elected by the propertied classes. Influence in public affairs should be accessible to all citizens who had demonstrated through the acquisition of wealth and education that they were capable of exercising the franchise intelligently. While the liberals resented the inherited privileges of the nobility, they also feared the wild passions of the proletariat. The man who lived in poverty and ignorance, they reasoned, was ripe for demagoguery and insurrection. The path of civic wisdom was therefore the *juste-milieu* between royal absolutism and mob rule, which had been established in England by the Reform Bill of 1832 and in France by the regime of Louis-Philippe. In economics, the teachings of liberalism advocated a policy of unrestrained competition by which wealth would become the reward of business acumen rather than the perquisite of corporative privilege. Guild monopolies, government regulations, and entailed estates were to be abolished as violations of that freedom of enterprise that alone could ensure the well-being of all society. As for the problem of unification, the liberals favoured the transformation of the German Confederation into a national monarchy in which the states' rights would be curtailed but not destroyed by a central government and a federal parliament.

Farther to the left stood the democrats or radicals, whose following was made up largely of small businessmen, petty shopkeepers, skilled workers, independent farmers, publicists, journalists, lawyers, and physicians. They looked with scorn on the golden mean between autocracy and anarchy that the liberals were seeking. They preferred an egalitarian form of authority in which not parliamentary plutocracy but popular sovereignty would be the underlying principle of government. Their supporters drew inspiration from the French rather than the English or the American Revolution. Their ideal of petty bourgeois democracy was the Jacobin republic of 1793 as an instrument to shape the energies and aspirations of the people into a disciplined force for political and social reform. The spokesmen of this ideal could not openly demand the overthrow of monarchical institutions without risking imprisonment. Yet while they were forced to accept the crown as a political institution, they sought to transfer its power to a parliament elected by equal manhood suffrage. The masses would thereby become the ultimate arbiter of politics. The democrats were also willing to accept government regulation of business activity as a means of improving the economic position of the lower classes, although their belief in the sanctity of private property was as firm as that of the liberals. In their advocacy of national unification, however, they were less solicitous about royal prerogatives and state rights. While not as numerous as the moderates, the radicals remained an important source of opposition to the established order.

The democrats and radicals

The growth of criticism directed against the political system of the Restoration forced its supporters to define their ideological position with greater precision. The old theories of monarchy by divine right or despotic benevolence offered little protection against the assaults of liberalism and democracy. Legitimism, whose defenders came mostly from the landed nobility, the court aristocracy, the officer corps, the upper bureaucracy, and the established church, began therefore to advance new arguments based on conservative assumptions regarding the nature of man and society. The relationship between the individual and his government, so the reasoning went, cannot be determined by paper constitutions founded on a doctrinaire individualism. Human actions are not motivated solely by rational considerations, but by habit, feeling, instinct, and tradition as well. The impractical theories of visionary reformers fail to take into account the historic forces of organic development by which the past and the present shape the future. To assert that all men are equal is to ignore differences in rights and duties expressing differences in birth, class, background, education, and tradition. The dogmas of constitutional authority and parliamentary government are merely a facade behind which a self-seeking bourgeoisie seeks to disguise its lust for power. An enduring form of government can be built only on the traditional institutions of society that men have learned in the course of time to accept: the throne, the church, the nobility, and the army. Only a system of authority legitimated by law and history can protect the worker against exploitation, the believer against godlessness, and the citizen against revolution. According to these tenets, the political institutions of the German Confederation were valid, because they represented fundamental convictions deeply embedded in the spirit of the nation.

The conservatives or legitimists

Economic changes and the Zollverein. The struggle of parties and ideologies during the Restoration reflected far-reaching changes in the structure of the economy and the community. The most significant of these changes was the rise of large-scale industry in central Europe. Techniques of mechanization, introduced in textile mills and coal mines, spread to other branches of manufacture and exerted pressures that influenced the entire economic life of the nation. The transportation network improved with the construction of railroads, steamships, bigger highways, and better canals. Banking institutions and private investors began to transfer their funds from government bonds and commercial ventures to manufacturing enterprises. Millowners, ironmasters, railroaders, financiers, and stockbrokers gradually formed a new middle class whose wealth was derived primarily from

The growth of large-scale industry

industrial activity and whose growing economic importance encouraged among its members a demand for greater political influence. Skilled handicraftsmen, constituting the bulk of the urban working class, could not compete successfully with the factories. The conflict between industrial and pre-industrial forms of output was aggravated, moreover, by important demographic changes in the period following the Congress of Vienna. Population began to shift from country to city, although a majority of the inhabitants of the German Confederation continued to live in rural communities.

The reorganization of agriculture

Agriculture also went through as difficult a period of reorganization and rationalization as industry. The end of serfdom had led in the regions east of the Elbe to the establishment of large estates belonging to aristocratic landowners but cultivated by a propertyless rural proletariat. Peasant emancipation in Prussia allowed the Junkers to enlarge their lands by absorbing the holdings of small farmers. The result was the continuing economic, social, and political domination of the village by the nobility in the eastern provinces of the Hohenzollern kingdom. The squirearchy entrenched in Pomerania, Brandenburg, Silesia, and East Prussia controlled agriculture, commanded the army, directed the bureaucracy, and influenced the court. It constituted a powerful force for conservatism and particularism.

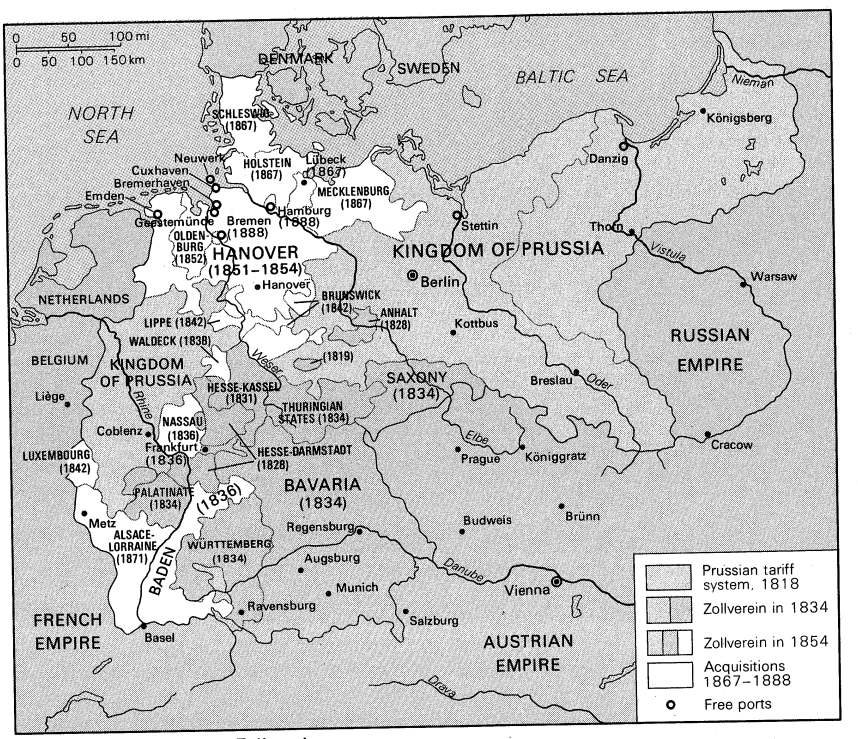
West of the Elbe the basic problem was not landlessness but overpopulation. The aristocracy along the Rhine and the Danube was often willing to give the peasantry possession of the soil in return for a substantial payment. The farmer was thereby saddled with heavy financial obligations. Many rustics tried to escape poverty by emigrating to the New World; those who remained faced swift demographic expansion and often had to subdivide small holdings until they yielded no profit. Civil discontent mounted among impoverished villagers who lacked employment in industry.

The system of authority in the German Confederation was thus being undermined by the struggle of artisans against industrial mechanization, by the disaffection of peasants hungry for land, and above all by the criticism of businessmen groaning under the shackles of particularism. Industrialists and financiers had to overcome the barriers created by a variety of monetary systems, commercial regulations, excise taxes, and state boundaries. It is little wonder that the bourgeoisie of central Europe turned increasingly to the teachings of liberalism and nationalism. Yet the established order did make a major attempt to meet the needs of the business community. Before long, several of the more important secondary governments also concluded agreements with Berlin by which a sizable free-trade area was established in the heart of central Europe.

The Zollverein

In 1834 the Zollverein, or "customs union," including most of the states of the German Confederation, came into existence. Only Austria and the northwest coastland remained aloof. The industrialists of the Habsburg empire, who wanted their products protected against outside competition, felt that the tariff rates of the new association were too low for their needs; whereas the merchants and bankers of the coastal region, who depended on imports, thought that they were too high. Yet for some 25,000,000 Germans the Zollverein meant in effect the achievement of commercial unification without the aid of political unification. The Prussian government, moreover, acquired a powerful new weapon in the struggle against Austria for the dominant position in central Europe. Still, although the customs union helped meet the most pressing demands of the middle class for economic consolidation, it could not surmount all the material disadvantages of a particularistic form of government. Men of means and education continued to grumble about the confederal system under which they lived, while the masses became increasingly restless under the pressure of social dislocation.

The revolutions of 1848–49. The hard times that swept over the Continent in the late 1840s transformed widespread popular discontent in the German Confederation into a full-blown revolution. After the middle of the decade a severe business depression halted industrial expansion and aggravated urban unemployment. At the same time, serious crop failures led to a major famine in the entire area from the Irish Sea to Russian Poland. In central Europe the hungry '40s drove the lower classes, which had long been suffering from the economic effects of industrial and agricultural rationalization, to the point of open rebellion. There were sporadic hunger riots and violent disturbances in several of the states, but the signal for a concerted uprising did not come until early in 1848 with the exciting news that the regime of the bourgeois king Louis-Philippe had been overthrown by an insurrection in Paris (February 22–24). The result was



The growth of the German Zollverein.

a series of sympathetic revolutions against the governments of the German Confederation, most of them mild but a few, as in the case of the fighting in Berlin, bitter and bloody. When on March 13 Metternich, the proud symbol of the established order, was forced to resign his position in the Austrian cabinet, the princes hastened to make peace with the opposition in order to forestall republican and socialistic experiments like those in France. Prominent liberals were appointed to the state ministries, and civic reforms were introduced safeguarding the rights of the citizen and the powers of the legislature. But even more important was the attempt to achieve political unification through a national assembly representing all of Germany. Elections were held soon after the spring uprising had subsided, and on May 18 the Frankfurt National Assembly met to prepare the constitution for a free and united fatherland. Its convocation represented the realization of the hopes that nationalists had cherished for more than a generation. Within the space of a few weeks, those who had fought against the particularistic system of the Restoration for so long suddenly found themselves in power with a popular mandate to rebuild the foundations of political and social life in central Europe. It was an intoxicating moment.

The Frankfurt National Assembly

The forces that had defeated the establishment order, however, soon discovered that they were in disagreement regarding the use to be made of their common victory. There were, first of all, sharp differences between the liberals and the democrats. While the former had comfortable majorities in most of the state legislatures as well as in the Frankfurt parliament, the latter continued to plead, agitate, and conspire for a more radical course of action. There were also bitter disputes over the form that national unification should assume. The Grossdeutsch (Great German) movement maintained that only Austria, the state whose rulers had worn the crown of the Holy Roman Empire for 400 years, could successfully guide the united fatherland. The Kleindeutsch (Little German) party, on the other hand, argued that the Habsburgs had too many Slavic, Magyar, and Italian interests to work single-mindedly for the greatness of Germany. The natural leader of the nation was Prussia, whose political vigour and geographic position would provide efficient government and military security for central Europe. Finally, there was a basic conflict between the proletariat, which wanted protection against mechanized production and rural impoverishment, and the bourgeoisie, which sought to use the political power it had gained in order to promote an industrial capitalism based on freedom of enterprise. Once the spring uprising was over, the parties and classes that had participated in it began to quarrel about the nature of the new order that was to take the place of the old. Popular support for the revolution, which had made the defeat of legitimism during the March days possible, began to dwindle, discouraged by the realization that the liberals would do no more to solve the problems of the masses than the conservatives had done. While the Frankfurt parliament was debating the constitution under which Germany would be governed, its following diminished and its authority declined. The forces of the right, recovering from the demoralization into which they had been plunged by the revolution, began to plan a counterrevolution.

Suppression of the revolutions

Their first major victory came in Austria, where the young emperor Francis Joseph found an able successor to Metternich in his prime minister Prince Felix zu Schwarzenberg. In the course of the summer of 1848 the Habsburg armies crushed the uprising in Bohemia and checked the insurrection in Italy. By the end of October they subjugated Vienna itself, the centre of the revolutionary movement, and now Hungary was still in arms against the imperial government. At the same time in Prussia, the irresolute Frederick William IV had been gradually persuaded by the conservatives to embark on a course of piecemeal reaction. Early in December he dissolved the constituent assembly that had been meeting in Berlin, promulgated by his own authority a middle-of-the-road constitution for the kingdom, and proceeded little by little to reassert the prerogatives of the crown. Among the secondary states there was also a noticeable shift to the right, as particularist princes and legitimist aristocrats began to regain their courage. By the time the Frankfurt parliament completed its deliberations in the spring of 1849, the revolution was everywhere at ebb tide. The constitution that the National Assembly had drafted provided for the creation of a federal union at the head of which was to stand a hereditary emperor with powers limited by a popularly elected legislature. Since the Austrian government had already indicated that it would oppose the establishment of a centralized system in Germany, the imperial crown was offered to the King of Prussia. Frederick William IV hesitated, brooded, and agonized, but in the end he refused the dignity whose authority was in his opinion too restricted. This rejection of political consolidation under a liberal constitution destroyed the last chance of the revolutionary movement for success. The moderates, admitting failure, went home to mourn the defeat of their hopes and labours. The radicals, on the other hand, sought to attain their objectives by inciting a new wave of insurrections. Their appeals for a mass uprising, however, were answered mostly by visionary intellectuals, enthusiastic students, radical politicians, and professional revolutionaries. The lower classes remained by and large indifferent. There was sporadic violence, especially in the southwest, but troops loyal to princely authority had little difficulty in defeating petty bourgeois democracy. By the summer of 1849 the revolution, which had begun a year earlier amid such extravagant expectations, was completely crushed.

The 1850s: years of political reaction and economic growth. The attempt to achieve national unification through liberal reform was followed by an attempt to achieve it through conservative statesmanship. Frederick William IV had refused to accept an imperial crown vitiated by parliamentary government, but he was willing to become the head of a national federation in which the royal prerogative remained unimpaired. While the Austrian armies were still engaged in the campaign against the revolution in Hungary, Berlin began to exert diplomatic pressure on the secondary governments to join in the formation of a new federal league known as the Prussian Union. If Frederick William IV had acted with enough determination, he might have been able to reach his goal before Francis Joseph could intervene effectively in the affairs of Germany. But he allowed his opportunity to slip away. Though he succeeded through threats and promises in persuading most of the princes to accept his proposals, no irrevocable commitments had been made by the time the Hungarians were defeated in August 1849. Vienna could now proceed to woo the secondary governments, which had in most cases submitted to Prussia only out of weakness and fear. Basically they remained opposed to sacrificing their sovereignty in order to exalt the Hohenzollern dynasty. When Schwarzenberg suggested the re-establishment of the old federal Diet, he won the support of many rulers who had agreed to follow Berlin against their will. The nation was now divided into two camps, the Prussian Union on one side and the revived German Confederation on the other. It was only a question of time before they would clash. When both Austria and Prussia decided to intervene in Hesse-Kassel, where there was a conflict between the supporters and the opponents of the prince, Germany stood on the brink of civil war. But Frederick William IV decided at the last moment to back down. His fear overcame his pride, especially after Nicholas I of Russia indicated that he supported Vienna in the controversy. By the Punctation of Olmütz of November 29, 1850, the Prussians agreed to the restoration of the German Confederation, and the old order was fully re-established in all its weakness and inadequacy.

Revival of the German Confederation

The years that followed were a period of unmitigated reaction. Those who had dared to defy royal authority were forced to pay the penalty of harassment, exile, imprisonment, or even death. Many of the political concessions made earlier, under the pressure of popular turmoil, were now restricted or abrogated. In Austria, for example, the constitution that had been promulgated

in 1849 was revoked, and legitimacy, centralization, and clericalism became the guiding principles of government. While in Prussia the constitution granted by the King remained in force, its effectiveness was reduced through the introduction of a complicated system of election by which the ballots were weighted in accordance with the income of the voters. The consequence was that well-to-do conservatives controlled the legislature. The secondary states returned to the policies of legitimism and particularism that they had pursued before the revolution. In Frankfurt am Main, where the federal Diet now resumed its sessions, diplomats continued to guard the prerogatives of princely authority and state sovereignty. The restoration of the confederal system also served the interests of the Habsburgs, who stood at the pinnacle of their prestige as the saviours of the established order. In Berlin, on the other hand, the prevailing mood was one of confusion and discouragement. The King, increasingly gloomy and withdrawn, came under the influence of ultraconservative advisers who preached legitimism in politics and orthodoxy in religion. The government, smarting under the humiliation suffered at the hands of Austria, was as timid in foreign as it was oppressive in domestic affairs. The people, tired of insurrection and cowed by repression, were politically apathetic. The German Confederation as a whole, rigid and unyielding, remained during these last years of its existence blind to the need for reform that the revolution had made clear.

Growth of industrial capitalism in central Europe in the 1850s

Yet the 1850s, so barren in politics, were of the highest importance in economics. For it was during this period that the great breakthrough of industrial capitalism occurred in central Europe. The national energies, frustrated in the effort to achieve civic reform, turned to the attainment of material progress. The victory of the reaction was followed by an economic expansion as the business community began to recover from its fear of mob violence and social upheaval. The influx of gold from America and Australia, moreover, generated an inflationary tendency, which in turn encouraged a speculative boom. Not only did the value of industrial production and foreign trade in the Zollverein more than double in the course of the decade, but new investment banks based on the joint-stock principle were founded to provide risk capital for factories and railroads. The bubble burst in 1857 in a financial crash that affected the entire Continent. For many investors the price of over-optimism and speculation turned out to be misfortune and bankruptcy. Yet Germany had now crossed the dividing line between a pre-industrial and an industrial form of economy. Although the rural population still outnumbered the urban, the tendency toward industrialization and urbanization had become irreversible. And this in turn had a profound effect on the direction of politics. For as wealth continued to shift from farming to manufacturing, from the country to the city, and from the aristocracy to the bourgeoisie, the pressure for a redistribution of political power also gained strength. While the reactionaries were solemnly proclaiming the sanctity of traditional institutions, economic change was undermining the foundation of those institutions. By the end of the decade a new struggle between the forces of liberalism and conservatism was in the making.

The 1860s and the triumphs of Bismarck. The revival of the movement for liberal reform and national unification at the end of the 1850s came to be known as the "new era." Its coming was heralded by scattered but distinct indications that the days of the reaction were numbered. In 1859 the defeat of Austria in the war against France and Piedmont had a profound effect on central Europe. For one thing, the maintenance of the authoritarian regime in Vienna depended on respect for its military strength. Now that the Habsburg armies had shown themselves to be vulnerable, popular unrest in the empire began to increase. Since autocracy was no longer an effective principle of government, Francis Joseph decided to experiment with a parliamentary form of authority. On October 20, 1860, he promulgated a constitution (the October Diploma) for his domains, setting up a bicameral legislature with an electoral system favouring the bourgeoisie; and Austria ceased to be an absolutist state. The achievement of political consolidation in Italy, moreover, aroused hope and envy north of the Alps. If the Italians could overcome the obstacles of conservatism and particularism, why not the Germans? National sentiment in central Europe, dormant since the revolution, suddenly awoke. Patriotic organizations like the Nationalverein (National Union) and the Reformverein (Reform Union) initiated an agitation for a new federal union, the former advocating Prussian, the latter Austrian, leadership. Liberal publicists and politicians began to advance plans for the reconstruction of the German Confederation. Some of the states, detecting a shift in the trend of public opinion, decided to change their course accordingly. Here and there the conservative ministers of the reaction were retired or dismissed, and their place was taken by statesmen with more moderate views.

The most significant portent of a new age in politics, however, appeared in Prussia. In 1857 Frederick William IV, crushed by memories of the mass insurrections and diplomatic defeats that he had been forced to endure, had suffered a mental breakdown. A year later his brother had become regent, and the government in Berlin immediately began altering the direction of its policy. Prince William, although a man of conservative inclination, had little sympathy with the mystical visions and pious dogmas prevailing at the court during the period of reaction. He dismissed the cabinet that had served his predecessor, announced a program of cautious reform in Prussian as well as German affairs, and won a popular endorsement of his course in the elections that gave the liberals control of the legislature. After a long period of discouragement, the advocates of civic reconstruction could once again look to the future with hope and expectation.

Yet there was an important difference between the political attitude of the liberals in 1858–59 from that of 1848–49. They came to feel during the new era that their defeat ten years before had been due to an excess of idealism and exuberance. The fatal mistake of the revolution, they reasoned, had been the assumption that enthusiasm and selflessness could be translated into power and substituted for statesmanship. Now a more calculating policy, a policy of *Realpolitik*, must be adopted. Not theory and rhetoric but negotiation and compromise would lead to the attainment of unity and freedom. The liberals therefore pursued at first a strategy of conciliation, anxious not to frighten the established order into blind resistance against all reform. In Prussia, for example, they waited patiently for the regent to move against the forces of disunity and oppression, confident that if they only gave him enough time he would obtain for them by royal authority what they could not seize through revolutionary violence. Yet it gradually became apparent that their hopes would not be realized. Prince William, who in 1861 became king in his own right, was a moderate conservative, but a conservative nevertheless. As the advocates of reform grew increasingly restless, the more militant among them formed the Progressive Party, which sought to hasten the introduction of liberal legislation by exerting pressure on the government. The monarch, afraid that he was being pushed farther to the left than he wanted to go, became more adamant and uncompromising. Sooner or later a conflict between crown and parliament was bound to arise.

The Prussian policy of *Realpolitik*

It came in connection with the question of army reform, although if that issue had not developed there would undoubtedly have been another. The King wanted to strengthen the armed forces by increasing the number of line regiments and decreasing the size of the popular militia. The legislature, reluctant to enhance the power of the conservative officer corps, demanded a modification of the plan. The King refused, convinced that the politicians were attempting to gain control of the army, which he considered subject only to royal authority. The legislature responded by withholding approval of the budget to pay for the cost of army reform. A complete deadlock ensued. In the spring of 1862 the liberal minis-

ters who had been appointed after the establishment of the regency were dismissed, and a conservative cabinet took office. But the new leaders of the government were as unsuccessful as the old in resolving the crisis. William I began to think about abdicating in favour of his son, who was believed to have political views close to those of the parliamentary opposition. He was persuaded, however, to consider first the possibility of naming a new ministry led by Otto von Bismarck, the Prussian ambassador to Paris. There was a momentous interview between the monarch and the envoy, as a result of which the former abandoned all thought of retirement, and the latter became head of a cabinet pledged to continue the struggle against the legislature. The battle between crown and parliament, which the liberals had been on the point of winning, was now to be waged without regard for constitutional provisions concerning the budget. On September 23, 1862, the nation was startled by the news that a statesman with a reputation for unyielding conservatism had become the prime minister of Prussia. It meant that the established order, having successfully defended its interests against the forces of reform after 1815 and after 1848, was determined to fight to the bitter end against the new challenge to its predominance.

Advent of Otto von Bismarck

The constitutional conflict in the Hohenzollern kingdom continued for another four years. The legislature refused to approve the budget until its wishes concerning military reform had been met. Bismarck's government, after carrying out the controversial reorganization of the army, continued to collect taxes and disburse funds without regard for parliamentary authorization. The liberals condemned the prime minister as a violator of the law, while the prime minister denounced the liberals as blindly doctrinaire. Although the electorate remained on the side of the opposition, the cabinet declared that it would not be swayed by party politics or parliamentary majorities. The broad masses of the population, it maintained, were still loyal to the crown. And so the struggle went on without prospect of alleviation or resolution. There were even dark prophecies of a violent uprising against a regime that was so indifferent to its constitutional obligations. Yet in fact Bismarck was not blind to the need for a reconciliation between crown and bourgeoisie. Despite his reputation as a fire-eating legitimist, he had a supple mind and recognized that the political principles of Frederick the Great could not solve the problems facing William I. He hoped, therefore, for an eventual reconciliation between the government and the legislature, but a reconciliation in which the prerogative of the monarch and the influence of the nobility would remain undiminished.

What Bismarck sought in essence was an alteration in the form of government that would create a facade of parliamentary institutions disguising the continuation of authoritarian policies. The middle class wanted to end the domination of the traditional forces in society, he calculated, but it also wanted to achieve national unification in central Europe. Here was the key to a solution of the constitutional conflict. Unity could be used to restrict freedom; nationalism could become the means of taming liberalism. Bismarck had concluded that the political integration of Germany was, in the long run, inevitable. If the established order did not effect it, the reformers, democrats, and revolutionaries would. Thus, it was in the interest of conservatism to take the task of centralization in hand, bring it to a successful conclusion, and create a new system of authority compatible with the preservation of royal and aristocratic preponderance. Such a policy would make possible a compromise between crown and bourgeoisie by which the latter obtained the benefits of economic consolidation, while the former retained the advantages of political domination. Through this strategy the prime minister hoped to end the constitutional conflict.

The defeat of Austria. The international situation was favourable to a program of unification in the German Confederation. Since its defeat in the Crimean War (1853–56), Russia had ceased to play a decisive role in the affairs of the Continent. England remained preoccupied with the problems of domestic reform. And Napoleon III was not unwilling to see a civil war east of the Rhine that he might eventually use to enlarge the boundaries of France. Bismarck could thus prepare for a struggle against Austria without the imminent danger of foreign intervention that had faced Frederick William IV. His first great opportunity came in connection with the duchies of Schleswig and Holstein, which were ruled by the king of Denmark but which were politically and ethnically tied to Germany. When the government in Copenhagen sought to make Schleswig an integral part of the Danish state in 1863, nationalist sentiment in central Europe was outraged. William I proposed to Francis Joseph that the two leading powers of the German Confederation should occupy the duchies in order to prevent the violation of an international agreement that had guaranteed their separate status. Afraid to let the Prussians act on their own, the Emperor agreed, and in 1864 there was a brief war against Denmark that demonstrated the strength of the reorganized army of the Hohenzollerns. Danish hopes for foreign assistance proved illusory, and by the Peace of Vienna (October 30) the duchies became the joint possession of Prussia and Austria.

The Danish War and the issue of Schleswig-Holstein

The easy victory of the allies, however, was only the prelude to a bitter conflict between them. Vienna would have liked to see Schleswig-Holstein become an independent secondary state in the German Confederation, committed to a policy of particularism. Berlin, on the other hand, hoped for the outright annexation of the duchies or at least the indirect control of their government. Even more important than the disposition of the spoils of war, however, was the mounting rivalry between the two great powers for hegemony in central Europe. Bismarck would probably have been willing to collaborate with Austria in the division of Germany into two spheres of influence, a northern under the Hohenzollerns and a southern under the Habsburgs. But Francis Joseph was reluctant to restrict his authority and ambition to the region below the Main River. The result was a steady deterioration of relations between Vienna and Berlin. In 1865 their differences were papered over by the Convention of Gastein, which placed Schleswig under Prussian and Holstein under Austrian administration, but which also reaffirmed the joint sovereignty of the two governments over the duchies. Still, this was only a temporary solution, and before long the danger of civil war in the German Confederation began to grow once again.

In the course of the spring of 1866 both sides stepped up their preparations for a military solution to the Austro-Prussian dualism. Bismarck concluded an alliance with Italy by which the latter was to receive Venetia as a reward for participating in a war against the Habsburg empire. He also sought to gain the support of public opinion in the German Confederation by introducing a motion in the federal Diet for the convocation of a national parliament elected by equal manhood suffrage. The Austrians in the meantime secured a promise of French neutrality in the event of hostilities and tried to win the adherence of the secondary states in the impending struggle. The last desperate attempts to preserve peace collapsed in June. Vienna announced that it would submit the question of the duchies to the federal Diet. Berlin, condemning this step as a violation of the Convention of Gastein, ordered its troops in Schleswig to expel the Austrians from Holstein. Francis Joseph in reply called on the other states of the confederation to mobilize their armies against the Prussian threat to domestic tranquillity, and central Europe trembled on the verge of civil war. The only question now was what the position of the secondary states would be. Most of them lined up behind Austria, which they regarded as the defender of their independence against the ambitions of Berlin. Bismarck's attempt to enlist the aid of the national movement by advocating reform of the confederal system thus failed. It alarmed the particularists without propitiating the centralists. Public opinion remained frightened and confused, distrusting one side and fearing the other. The future of the nation was decided not by popular insurrec-

Prussia's defeat of Austria in the Seven Weeks' War

tions or parliamentary deliberations but by the force of arms.

The Seven Weeks' War between Prussia and Austria (June–August) produced a diplomatic revolution in Europe, destroying the balance of power that had been established 50 years before by the Congress of Vienna. Yet this momentous alteration in the international equilibrium was accomplished so swiftly that foreign diplomats had barely begun to grasp its implications before the struggle for hegemony in Germany ended. The Hohenzollern armies had a brilliant strategist in Helmuth von Moltke and a deadly weapon in the breech-loading needle gun. The Austrian high command, on the other hand, became irresolute and demoralized before a decisive encounter had even taken place. The Prussians succeeded in dividing and defeating the forces of the secondary states, and on July 3 they routed the Habsburg troops as well at the Battle of Sadowa (Königgrätz). The war was thus decided within a few weeks after its outbreak. Bismarck, refusing to be dazzled by the brilliance of the victory, urged the swift conclusion of an honourable peace. Not only did he feel that the preservation of a strong Austria was essential for the maintenance of stability on the Continent, but he feared that a prolongation of hostilities would enable Napoleon III to intervene in the affairs of Germany. By the preliminary Peace of Nikolsburg (July 26) and the definitive Treaty of Prague (August 23), Francis Joseph was permitted to retain all of his possessions except Venetia, which had been promised to the Italians. There was to be no occupation and only a modest indemnity. The Emperor had to acquiesce, however, in the Prussian annexation of Hanover, Nassau, Hesse-Kassel, Schleswig-Holstein, and Frankfurt am Main, in the dissolution of the German Confederation, and in the formation under Hohenzollern leadership of a new federal union north of the Main River. The contest between Berlin and Vienna that had determined the history of central Europe for more than a century was now over.

Bismarck's national policies: the restriction of liberalism. Bismarck's triumph in the military struggle led directly to his victory in the constitutional conflict. Before the outbreak of hostilities he had tried to reach an understanding with the liberal opposition, but the liberals hesitated to make peace with a statesman who had so flagrantly violated the fundamental law of the kingdom. The defeat of Austria changed all that. While the war was still in progress, general elections resulted in important gains for the right. Many voters, elated over the successes of the Prussian armies, expressed their confidence in the government by supporting its adherents at the polls. Some of the ultraconservatives hoped that the Cabinet would now capitalize on its triumph by suspending the constitution and establishing an authoritarian regime. Yet the prime minister recognized that such reactionary schemes would prove futile in the long run. What he wanted was not the suppression of liberalism but an accommodation with it. As soon as peace was concluded, he introduced in the legislature a bill of indemnity granting the government retroactive approval for its conduct of affairs of state without a legal budget. The consequence, as Bismarck had foreseen, was a split in the ranks of his adversaries. Those who argued that there could be no compromise with the principle of constitutional government rejected a reconciliation between crown and parliament based on mutual concessions. But many members of the opposition, who eventually formed the National Liberal Party, decided to accept the settlement offered by the prime minister. Their reasoning was that an obstinate resistance against the cabinet would only condemn them to sterile dogmatism, whereas a willingness to accept what could not be prevented would enable them to influence official policy in the direction of greater freedom. This view prevailed, and on September 3, 1866, the legislature approved the Bill of Indemnity, 230 to 75. By dividing the forces of reform and weakening their sense of purpose, Bismarck won as important a success in domestic affairs as the victory on the field of battle.

His subjugation of liberalism was made possible by the triumph of nationalism. It had long been the prime minister's belief that the achievement of unity could appease the demand for freedom. The success of the Prussian armies provided him with an opportunity to test this assumption. Once Austria had been subdued, he cajoled and bullied the secondary governments above the Main into joining Berlin in the establishment of the North German Confederation. It was the union of a giant and 21 pygmies, for the Hohenzollern kingdom comprised about four-fifths of the territory and population of the confederation. Executive authority was vested in a presidency held in accordance with hereditary right by the rulers of Prussia, who were to exercise the powers of their office with the assistance of a chancellor responsible only to them. The legislature was composed of a federal council or Bundesrat, whose members were appointed by the state governments, and a lower house or Reichstag, elected by equal manhood suffrage. Since Berlin had 17 votes in the Bundesrat out of 43, it could easily control the proceedings with the support of a few of its satellites among the small principalities. Although the Reichstag could theoretically exercise considerable influence over legislation by granting or withholding its approval, parliamentary authority and party initiative were weak and untested, and Bismarck had little difficulty in piecing together a workable majority for his program by a strategy of divide and rule. His support came largely from a combination of moderate liberals and moderate conservatives willing to sacrifice theory for expediency. The federal constitution provided no bill of rights, no ministerial responsibility, and no civilian supervision over military affairs. But it introduced uniformity in currency, weights, measures, commercial practices, industrial laws, and financial regulations. In short, it created the economic unity that the middle class had long demanded and that helped reconcile it to the defeat of its hopes for greater political freedom.

Establishment of the North German Confederation

Franco-Prussian conflict and the new German Reich. The Seven Weeks' War, by creating a powerful new state in the heart of central Europe, abruptly altered the system of international relations on the Continent. Every government now had to re-examine its diplomatic and military position in the light of the establishment of the North German Confederation. No nation, however, was affected by the victory of the Hohenzollern armies as directly as France. Emperor Napoleon III had encouraged the outbreak of hostilities between Austria and Prussia on the assumption that both combatants would emerge from the struggle exhausted and that the Second Empire of France could then expand eastward against little resistance. The outcome of the war revealed how shortsighted such calculations had been. Instead of profiting from the conflict between Francis Joseph and William I, Paris suddenly confronted a strong and united German state that presented a serious threat to French interests. The imperial government was bound to regard this turn of events with suspicion and hostility. It sought to mitigate its discomfiture by seeking compensation in the Rhineland, Luxembourg, or Belgium. But Berlin succeeded in frustrating these plans, and the conviction began to grow in France that sooner or later a struggle with Germany would be unavoidable. The prospect of a new armed conflict was not unwelcome to Bismarck. He wanted to see national unification consummated by the entry of the southern states into the North German Confederation. Yet public opinion below the Main remained distrustful. Only a common patriotic struggle against foreign aggression might overcome the reluctance of the south to unite politically with the north. Thus in Berlin as well as in Paris there were reasons for seeking a test of strength. The immediate occasion came in the spring of 1870 with the candidacy of Prince Leopold, a relative of William I, for the throne of Spain. The ensuing controversy was cleverly exploited by Bismarck to provoke the French into initiating hostilities which both sides had wanted but for which the French government received the blame.

When France learned of Leopold's acceptance of the offer of the Spanish crown, there were wild protests in Paris and an immediate demand that Leopold be ordered to

withdraw. On July 12 Leopold's father renounced the Spanish candidature on his behalf. This was not enough for the French government; it insisted that William I, as head of the Hohenzollern family, should promise that the candidature would never be renewed. This demand was presented to William at Ems by the French ambassador. Though William refused to give a promise, he dismissed the ambassador in a friendly enough way. But when the "Ems telegram," a report of what was happening, reached Bismarck, he shortened it for publication in such a way as to imply that William had refused to see the French ambassador again. The French used this as an excuse to declare war on Prussia on July 19.

Franco-German War

Bismarck's calculation that a struggle waged ostensibly against the expansionist designs of Napoleon III would overcome particularism below the Main proved correct. The southern states joined the north in the Franco-Prussian War, and the sense of unity engendered by the brotherhood of arms was soon enhanced by the intoxication of victory. The German troops won one battle after another in hard fighting along the frontier, until on September 1 they forced a large French Army, headed by the Emperor himself, to surrender at Sedan. The result was the establishment of a republican government in France, which continued to wage the struggle in the name of the old revolutionary ideals of 1793. But the generalship of Moltke was too much for the fierce determination of the new regime. Paris capitulated on January 28, 1871, after a long and bitter siege, and on May 10 the Treaty of Frankfurt brought the war officially to a close. The Third Republic had to cede Alsace-Lorraine, pay an indemnity of 5,000,000,000 francs, and accept an army of occupation. It was a Carthaginian peace designed to crush a dangerous rival. The work of national unification in Germany, in the meantime, was successfully completed even before hostilities had ended. Bismarck had entered into negotiations with the southern states soon after the outbreak of war, determined to use patriotic fervour as an instrument for achieving political consolidation. The enthusiasm aroused in central Europe by the victory over France proved too much for the defenders of particularism. On January 18, while Prussian guns bombarded Paris, William I was proclaimed emperor of a united nation at military headquarters in Versailles. The governments below the Main joined the North German Confederation to form a powerful new *Reich* under the Hohenzollerns. Within a single lifetime central Europe had completed the transition from cosmopolitanism to nationalism, from serfdom to industrialization, from division to union, from weakness to preponderance, from the Holy Roman Empire to the German Empire. (T.S.H.)

V. Germany from 1871

THE SECOND REICH, 1871–1918

Bismarck as imperial chancellor, 1871–90. *The making of the German Empire.* During the Franco-Prussian War, negotiations were pushed on for the uniting of all Germany outside Austria. Bismarck gave some political concessions to the southern German states, and eventually both Bavaria and Württemberg made separate treaties of union with Prussia. The princes were persuaded to offer the imperial crown to King William I of Prussia, and he was proclaimed German emperor on January 18, 1871. A Reichstag elected from all Germany accepted as the imperial constitution that of 1867, with certain minor concessions allowed to Bavaria. The new *Reich* consisted of four kingdoms, five grand duchies, 12 duchies and principalities, and three free cities (Hamburg, Lübeck, and Bremen). Alsace-Lorraine was treated as a conquered province, ruled by an imperial governor. The weakness of the constitution was its inadequate definition of the powers of the Reichstag over the executive. The chancellor was defined as "responsible," but it was not stated to whom; Bismarck contended that his responsibility was to the emperor, while the politicians asserted that it was to the Reichstag. The question of effective control of the army through authorization of expenditure was also not clear. Bismarck failed to get provision for a permanent grant to be written into the constitution and had to agree to a compromise, the Septennat, by which finance for the army was to be voted every seven years. As a result, artificial alarm had to be created every seven years in order to get the army grant renewed.

Bismarck's liberal period and the Kulturkampf. From 1871 to 1879, the National Liberals acted almost as a government party in the Reichstag, and their predominance made the first period of the empire an age of great liberal reform. Germany acquired a uniform legal procedure, uniform coinage, and uniform administration. An imperial bank was created, most restrictions on freedom of enterprise and of movement were removed, and limited companies and trade combinations were allowed. Freedom of the press was secured in 1874, and work was begun on an imperial civil code that was in universal operation by 1900. Particularly important was the establishment of municipal autonomy in 1873; towns were freed from control of the *Landrat* (usually a large landowner), and the way was clear for the development of local government, in which Germany led the world.

Distrust of political Roman Catholicism, which Bismarck shared with liberals, grew stronger in 1871 when a confessional party, the Centre, gained 58 seats in the Reichstag and drew support from all the elements that had opposed Bismarck's work. Believing that unity was more easily created when there was some object to attack, Bismarck vigorously denounced the party, and a dispute over the clerical control of education extended his policy into a general attack, known as the *Kulturkampf*, on the independence of the Roman Catholic Church. The recent defection of the so-called Old Catholics, who could not accept the doctrine of papal infallibility proclaimed in 1870, caused the church to seek to expel all Old Catholics from teaching positions. Bismarck in turn attacked the Catholic teaching orders, expelling the Jesuits and insisting that the state should train and license priests. Priests and bishops who would not conform were imprisoned, and sees were left vacant. These penal measures were expressed in the May Laws passed by the Prussian Landtag (state assembly) in 1873. But, by about 1875, it was clear that Bismarck would not achieve victory; the Old Catholics carried no weight, and even many Protestants disliked the attack on religious teaching.

The *Kulturkampf*

Bismarck's breach with the National Liberals. From 1877 fiscal, economic, and political factors turned Bismarck against the National Liberals. The question of raising money to meet deficits in the armed forces' budgets created dissension, the Liberals wanting it to be done by direct taxation that would be controlled by the Reichstag but Bismarck obtaining some grants of indirect taxes voted for an indefinite period. His attempt to silence the Liberals' opposition on this issue by offering to make their leader virtually his deputy was foiled. Soon afterward the breach was widened, when Bismarck introduced protectionist measures designed to safeguard German agriculture from the new competition provided by American wheat. Regarding agricultural workers as potentially good soldiers, Bismarck wished to keep German agriculture prosperous. He also took steps to protect the German iron and steel industry. The Liberals, traditionally supporters of free trade and laissez-faire policies, were further alienated. By 1879 Bismarck was ready to form an alliance with the Centre. Gradually he reversed the measures taken against the Catholic Church, in 1880 suspending the May Laws for individual cases and abolishing the secular examinations that had been required for candidates for the priesthood. By 1887, when peace was finally restored, the church had regained complete control of seminary education; in return, the Centre Party agreed to support all of Bismarck's nonreligious policies.

In the early 1870s, the Roman Catholic Church had been Bismarck's scapegoat; from the time of his alliance with the Centre, his attacks became directed against the Social Democrats, particularly after an attempted assassination of the Emperor in 1878. The Social Democratic Party was declared illegal, although its members still sat in the Reichstag; but the uproar produced a conservative majority in an election later in 1878, and in 1879 the Septennat was renewed almost without opposition.

Social security measures

During the early 1880s, Bismarck promoted measures of social security in order to attach the workers to the state. A Sickness Insurance Law was passed in 1883, a scheme of compulsory accident insurance in 1884, and old age pensions were also introduced.

Nevertheless, Bismarck's alliance with the Centre was never completely firm. At any time in the period from 1879 to 1890, they could have defeated him by alliance with the Radicals and the Social Democrats. So he continued to create artificial panics, in 1884 colonial disputes and in 1887 foreign dangers, in order to stampede the electorate to his support.

Bismarck's foreign policy, 1871–90. Once the empire was founded, Bismarck's sole aim was peace and security, pursued at first through neutrality and later through a complicated network of alliances. Resolutely impartial during the great Near Eastern crisis of 1875–78, when war nearly broke out between Great Britain and Russia, he decided thereafter to pursue a more active role, concluding, in 1879, a defensive alliance with Austria-Hungary. This, while guaranteeing Austria-Hungary's survival as a great power, specifically gave it no support for its Balkan ambitions. Russia, which might have been antagonized by the alliance, felt that there was now less danger of Austria working with Britain and was persuaded to renew (1881) the Dreikaiserbund, or League of the Three Emperors (of Russia, Austria-Hungary, and Germany), a vague alliance that Bismarck had first established in 1873. Bismarck also concluded (1882) the Triple Alliance of Germany, Austria, and Italy; Italy was guaranteed support against France in return for a promise of neutrality in the event of a war between Austria and Russia. A crisis in which Russo-Austrian rivalries were aroused over Bulgaria in 1885 caused Bismarck to strengthen his ties with both sides in order to maintain a balance between them. By the Reinsurance Treaty (1887) he promised Russia diplomatic support in the Near East and German neutrality in any Russian war except one with Austria; by promoting a Mediterranean naval agreement between Italy, Britain, and Austria, he established virtually the Triple Entente opposing Russia in the Near East. These apparently contradictory policies have caused Bismarck to be charged with duplicity; in fact, Germany's position in the centre of Europe necessitated a two-faced policy. On this occasion, at least, a Balkan war was averted.

Bismarck occasionally pursued other issues in order to satisfy the aspirations of German nationalism. Thus, between 1883 and 1885, he deliberately encouraged colonial rivalry with Great Britain; and, in the 1870s and again after 1886, he encouraged resentment against the Poles. In 1887, when a new army grant was needed, he raised the alarm that the French would try to regain Alsace-Lorraine. Although the Reichstag threw out his army bill, he was able in the subsequent general election to win sufficient support to get it passed (March 1887) in the new house.

The fall of Bismarck. Bismarck's position had always depended on the emperor's support; during the reigns of William I (died March 1888) and Frederick III (died June 1888) this was assured, but the position was entirely changed with the accession of William II. The new emperor had no reasons, as had his predecessors, for gratitude to Bismarck; moreover, he represented a new, self-confident Germany and was impatient with Bismarck's social conservatism, which seemed to prevent the emperor from establishing a demagogic relation with his subjects. Lacking imperial support, Bismarck's position became untenable after his political allies lost the general election of 1890. His attempt to engineer a strike of Prussian ministers failed, and he was forced to resign on March 18, 1890.

Accession of William II

Chancellors Caprivi and Hohenlohe, 1890–1900. The chancellorships of Leo, Graf von Caprivi (1890–94), and of Chlodwig Karl Victor, Fürst zu Hohenlohe-Schillingsfürst (1894–1900), represented, respectively, an official attempt at alliance with left-wing elements and a disillusioned return to conservatism. Caprivi's abandonment of old conservative policies was shown in every field. In foreign policy, he refused to renew the Reinsurance Treaty with Russia, thus breaking the traditional partnership between tsardom and the Prussian monarchy. He gave encouragement to Austrian ambitions in the Balkans and hoped to persuade Britain to join the Triple Alliance (of Germany, Austria-Hungary, and Italy). A step in this direction was the agreement (1890) by which Britain gave the North Sea island of Heligoland to Germany in return for colonial concessions in East Africa. At home, Caprivi promoted social security measures and wooed the parties of the left. But, like Bismarck, he was confronted with the perennial problem of achieving the passage of the Septennat. He eventually got the army grant passed in 1893, after having conceded that its renewal should in future occur every five instead of seven years. By representing Russia as the "enemy," he drew his support for the renewal from the left rather than the right. But the political manoeuvres involved split the German Radical Party, a serious event of decisive future importance.

Meanwhile, the Emperor had become disillusioned with Caprivi's social policies. Failure of the Chancellor's pro-British policy, occurring when a colonial dispute (1894) with the British caused them to repudiate their promises of support for Austria-Hungary, further weakened his position. A quarrel between Caprivi and the Prussian prime minister (Caprivi had given up the Prussian premiership in 1892, in order to appear less conservative) gave William II the opportunity to dismiss both ministers (October 1894).

Caprivi's successor, Hohenlohe-Schillingsfürst, an old man (75), was intended to revive the glories of the Bismarckian era without engendering its troubles. He got on well with the Conservatives and, as a former prime minister of Catholic Bavaria, was acceptable to the Centre Party and agreed to many of their confessional demands. Hohenlohe returned to a policy of alliance with Russia and refused to support Austria-Hungary in the Balkans. One of the most striking events of his chancellorship was the dispute with Britain over the Boer republics in South Africa, which culminated in the telegram sent (January 3, 1896) by the Emperor, congratulating Paul Kruger, president of the Transvaal, on the defeat of the British Jameson raid. The telegram did not affect British policy, but it had a lasting effect on German national sentiment, which, for the first time, came to regard Britain as Germany's principal rival in imperial greatness.

The Kruger telegram

Chancellorship of Bülow, 1900–09. Appointed secretary of state in June 1897, Bernhard, Fürst von Bülow, wielded the real political power in Germany during the last years when Hohenlohe was chancellor and succeeded him in that office in October 1900. He founded his power on the support of the Conservative Prussian landowners, bought, with the aid of the Prussian finance minister, Johannes von Miquel, by granting easy credit facilities and high tariffs to protect agricultural produce. Thus the landowners became economically dependent on the *Reich* as a whole, obliged to support Pan-German policies.

With this assurance behind him, Bülow launched on a policy of achieving grandeur abroad in order to stave off reform or conflict at home. The popular desire for imperial greatness was ripe for exploitation, and, indeed, the strength of Germany's iron and steel industry in the Ruhr had already made Germany Europe's greatest industrial power. At a time when Britain, France, and Russia appeared occupied with their own rivalries in Africa and in the Far East, it seemed desirable to Bülow to keep Germany free from involvements, apart from a protective relationship with Austria-Hungary. He rejected alliances offered by Britain in 1898 and 1901, but he retained a friendly relationship by participating in a hypothetical plan to carve up the Portuguese colonial empire and by maintaining official neutrality during the war between Britain and the Boers in South Africa.

A major cause of the decline in good relations with Britain was the rapid development of the German Navy as a result of the Navy laws of 1898 and 1900, promulgated by Alfred von Tirpitz, secretary for the navy from 1897 to 1916. The supporters of the project felt that a great navy was essential for a great power; moreover, its

construction would provide steady work for the iron and steel industry. The plans were grandiose in scale, and the Navy Law of 1900 laid down details of annual development up to the year 1917. The naval program had two main effects. First, by 1909, when parity with the Royal Navy was almost achieved, Britain became seriously alarmed; but, even had it been desired, it would have been difficult then to reverse such carefully laid plans. Second, the problem of paying for the navy caused serious splits between the Conservatives, who supported indirect taxation, and the Centre Party, which wanted direct taxation. The Centre was not successful, and the navy, in fact, was paid for largely by state borrowing, thus inaugurating the inflationary trends that later notably characterized German finance during World War I.

The policy of international detachment received a blow in 1902 when the Anglo-Japanese alliance and the subsequent Russo-Japanese War of 1904–05 enabled Britain to check the Russians in the Far East without becoming involved itself. Germany then began to seek to effect a continental bloc opposed to Britain; France, which in 1904 had concluded the entente with Britain, seemed to need coercion in order to join it. Germany's attempt at such coercion resulted in the first Moroccan Crisis.

William II landed at Tangier on March 31, 1905, and announced German support for Moroccan independence of France. The French sought to negotiate but, under the threat of war, gave in to Germany's demand that Théophile Delcassé, the foreign minister, resign. Germany's stand seemed quite justified, and Bülow was created a prince. But the sequel was less satisfactory. Britain rallied to France's support, and, at the Algeciras Conference (January–April 1906), Germany was obliged to acquiesce in the continuation of French influence in Morocco. The Tangier triumph had not appealed overmuch to the German public; and an important effect of the Algeciras reversal was the resignation from the German foreign ministry of the undersecretary Friedrich von Holstein, who had largely influenced policy since the 1870s.

A change in the political basis of Bülow's power occurred in 1906, when the Centre Party, on which, with the Conservatives, he had hitherto relied, withdrew its support when its ever increasing demands for concessions to the Catholic Church were resisted. The Centre members of the Reichstag voted against a resolution providing finance for the suppression of a native revolt in southwest Africa, and Bülow at once sought other political allies. He gained the support of a variety of non-Socialist left-wing parties, including the two branches of the old Radical Party. His allies won a handsome majority in the election of 1907 but were still not strong enough in the Reichstag to outvote Conservative opposition to the introduction of direct taxation. Nevertheless, his new alliances enabled Bülow to pose as a liberal statesman; and the zenith of this phase of his chancellorship was reached in October 1908 when, after an indiscreet and bombastic interview granted by William II to the English newspaper *The Daily Telegraph*, Bülow was able grandly to announce that in future the Emperor would "respect his constitutional obligations."

William II's *Daily Telegraph* interview

The logical consequence of the swing toward liberalism in home affairs should have been a rapprochement with Britain and an estrangement from Russia, as in the days of Caprivi. But Anglo-German relations worsened and reached a crisis in 1909 when Britain, in reply to Germany's expanded naval program, inaugurated a hasty increase in its naval building schedules. Bülow, however, did fall out with Russia when, in 1908, he supported Austria-Hungary's annexation of Bosnia-Hercegovina.

Bülow, like Bismarck, could survive in power only with imperial or adequate parliamentary support. The Conservatives resented his quarrel with Russia and resented still more a proposed introduction of death duties on landed estates. Tiring of their strange alliance with the left, they made common cause with the Centre Party, defeated the death-duties proposal by a narrow majority, and caused him to resign. The Emperor, bitterly inimical to him since the *Daily Telegraph* affair, accepted his resignation (July 1909). Bülow was the last effective chancellor of the Second Reich; his successors were mere administrators.

The prewar years, 1909–14. Theobald von Bethmann Hollweg, Bülow's successor as chancellor, was the first to be appointed to that office from the purely administrative grade of the civil service. Essentially a bureaucrat, he lacked the force to pursue any policies of his own and, although of high character, constantly gave in to the dubious counsels of Tirpitz, the Emperor, and the secretary of state, Alfred von Kiderlen-Wächter.

The second Moroccan Crisis, of 1911, was almost entirely Kiderlen-Wächter's responsibility. He believed that the recent French occupation of Rabat and Fez meant the end of Moroccan independence and decided to try to coerce France into making concessions to Germany. The gunboat "Panther" was sent to the port of Agadir, and Germany demanded part of the French Congo in return for alleged (but nonexistent) German "rights" in Morocco. On this occasion, German public opinion was more belligerent than the government had expected; Germany had to create an artificial war crisis but was outmanoeuvred by the British and French, acting in close cooperation. Although Germany acquired a portion of the Congo to add to its territory in the Cameroons, Bethmann and Kiderlen were furiously attacked in the Reichstag for their timidity.

This crisis had seriously aggravated international tension. In the remaining years before the outbreak of World War I, Bethmann made sincere efforts to reduce it, but he was always pushed by ministers such as Tirpitz and the German military general staff and was influenced by an aggressive public opinion. He could postpone but not avert war.

In his home policy Bethmann Hollweg was even less effectual than in foreign affairs, and the years immediately preceding the war saw a complete stalemate in German domestic politics, culminating in near farce over the affair of the arrogant officers garrisoning Saverne in Alsace. It was unfortunate for German politics at this time that no genuine democratic majority could be arrived at, because the Social Democrats clung to revolutionary theories that prevented other left-wing parties, which otherwise would have been their natural allies, from joining with them. In fact, the Social Democrats became in 1912 the largest single party in the Reichstag; but in the face of their victory, Bethmann maintained his customary inertia, neither attacking nor wooing them. When, in November 1913, the officers at Saverne provoked, and then illegally arrested, a number of the townsfolk, Bethmann, although privately disapproving of their action, defended them in the Reichstag. For so doing, he was defeated in a vote of censure, by 293 votes to 54. German politics had reached such a low ebb that he neither resigned after this defeat nor were steps taken to punish the soldiers whose action the Reichstag had so vigorously condemned. In fact, the same assembly shortly afterward voted a capital levy in order to increase and further equip the army. Thus, right up to 1914, the German people tried to combine the rule of law at home with the rule of German military power abroad.

The Saverne affair

World War I. There is little or no evidence that the Germans deliberately planned war in the summer of 1914. The crisis of July 1914, caused by the assassination of the Austrian archduke Francis Ferdinand, caught the German statesmen unawares, forcing on them a decision whether or not to stand by Austria-Hungary. In authorizing Austria to act against Serbia and promising support if Russia tried to intervene (July 5), they did not at first realize that they had chosen war; they supposed that a firm line would lead the other powers to give way. Three weeks later Germany vainly warned Russia against mobilization. At this point, Helmuth von Moltke, chief of the general staff, believing that Germany's only chance of victory lay in defeating France before Russia was ready, insisted that, in view of Russia's mobilization, Germany must immediately declare war on both France and Russia. Bethmann could not effectively answer this military argument and gave in, finally himself defending the German march through Belgium, which brought

Great Britain into the war against Germany.

The outbreak of war achieved something that social concessions had failed to do; it brought the Social Democrats over to the support of the imperial government. The German Socialists had always been the leading spokesmen in the Socialist International of the general strike against war. When it came to the point, they were won over by the argument that Germany was being attacked by reactionary tsarist Russia. At the meeting of Socialist members, a minority opposed the war; but when the Reichstag met, the entire Socialist Party voted for war credits, in the name of party unity. The Socialists went further. They joined the other parties in declaring *Burgfrieden*, a civil truce, by which they agreed to criticize neither each other nor the government. Thus the members of the Reichstag abdicated to the imperial government, though it remained unchanged and beyond their control.

Truce among political parties

For Germany, as for other belligerent countries, World War I fell into two distinct phases: the first, that of conventional warfare, which lasted until 1916; the second, a war of desperate expedients when both sides struggled for their very existence. The German High Command had envisaged a short war in which France would be overrun in six weeks and Russia in six months. In the west, this plan was disrupted by the First Battle of the Marne (September 1914), as a result of which the Germans failed to capture Paris and were obliged to occupy long miles of trenches across Belgium and northern France. Yet the defeat of the Russians at Tannenberg had given Germany the security that was its ostensible war aim, and, at any time up to the summer of 1917, it could have negotiated peace on a status quo basis. But such a peace was not then acceptable to Germany, since it would have arrested the expansion of German industry and brought about political revolution at home.

After the Marne, Moltke was succeeded as chief of the general staff by Erich von Falkenhayn, whose plan to remain defensive on the western front, while attacking in the east, achieved some success. Anglo-French offensives on the west achieved nothing, while Germany drove the Russians from Galicia, overran Serbia (autumn 1915), and, by the entry of Bulgaria into the war as its ally, secured a land route to Turkey and beyond to the Persian Gulf.

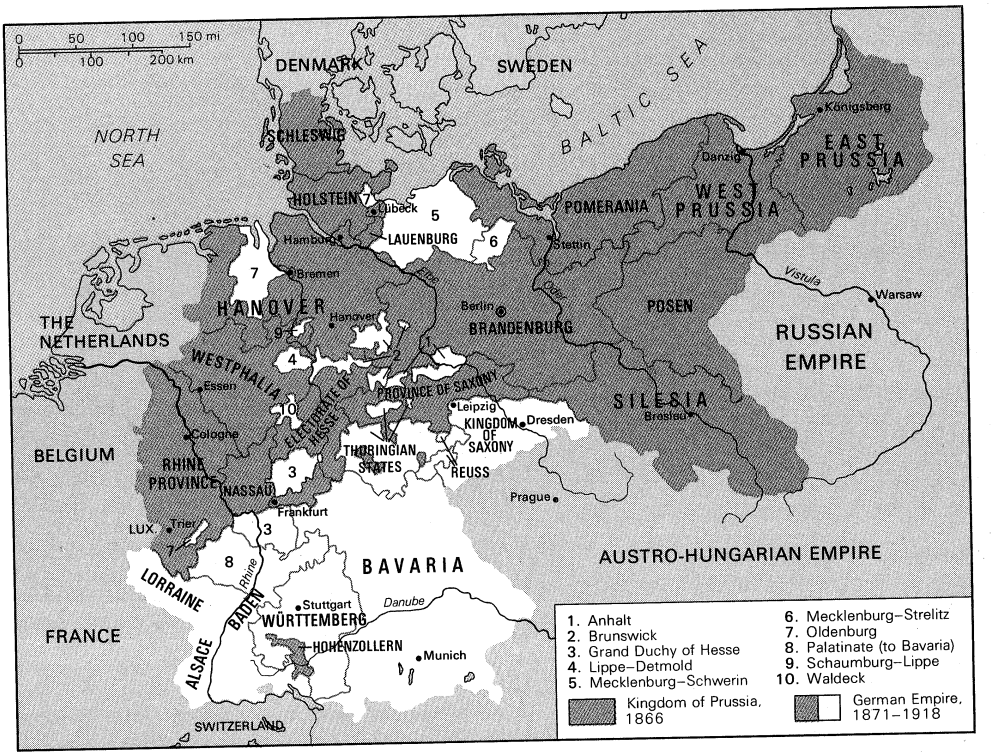
The second year of war saw the first serious efforts to mobilize and organize German resources; this was largely the work of an industrialist, Walter Rathenau, who convinced the government of the need for economic planning. Gradually, despite the success of Falkenhayn's strategy, it became clear that German interest had shifted to the west, mainly because control of the industrial resources of Belgium and of northeastern France became a major preoccupation of the German steel magnates. Annexation, or at least partial control, of Belgium began to be regarded as an essential condition for peace. Some Liberal elements continued to think that peace could be made without excessive demands, while a minority group of the Social Democrats became virtually pacifist.

In 1916 Falkenhayn's attempt to "bleed the French white" by the prolonged Battle of Verdun (February–June) proved almost as exhausting for the Germans. At the same time, an attempt to break British naval power by direct assault failed at the Battle of Jutland (May 31, 1916), and two months later William II dismissed Falkenhayn, appointing Paul von Hindenburg chief of staff, with Erich Ludendorff as his quartermaster general.

These appointments, reflecting a general desire for more vigorous policies, brought to an end any chances of a compromise peace and resulted in a major German political crisis. Bethmann had been negotiating with tsarist Russia for a peace without victory, but his moderation became unpopular, and in October 1916 the Reichstag passed a motion declaring that it had confidence in Bethmann so long as he possessed the confidence of the High Command. Thus Bethmann could oppose none of Hindenburg's and Ludendorff's policies. Ludendorff's insistence on the proclamation (November 1916) of an independent kingdom of Poland brought the Russian peace negotiations to an end. The decision (January 1917) to inaugurate unrestricted submarine warfare was made against Bethmann's judgment; the attempt to blockade Britain by this means was ultimately defeated by the British convoy system, while such warfare had the far graver consequence of bringing the United States into the war against Germany.

Unrestricted submarine warfare

There was a shortage of food during the winter of 1916–17, and war-weariness developed. The first Russian Revolution (March 1917) encouraged left-wing feeling, and soon afterward Matthias Erzberger, leader of the Centre, decided that it would be politically opportune to secure for it the position of most prominent anti-war



The German Empire (1871–1918).

party. In July he attacked Bethmann for advocating the policies of conquest that Ludendorff had, in fact, forced on him. Ludendorff, for quite opposite reasons, regarding Bethmann as too pacific, also wanted to oust the Chancellor and so welcomed Erzberger's attack. The politicians wanted to replace Bethmann by Bülow; but William II, remembering his humiliation by Bülow over the *Daily Telegraph* interview, refused to appoint him. Since the politicians could put forward no alternative candidate, Ludendorff then nominated an unknown official, Georg Michaelis, who would be a mere puppet.

Having failed to make a chancellor, the Reichstag was, however, allowed to dabble in policy making. On July 19 it passed a "peace resolution," which was little more than a string of innocuous phrases and lacked concrete proposals. Further apparent concessions were made to the Reichstag in October when, after dismissing Michaelis for incompetence, Ludendorff appointed as chancellor Georg, Graf von Hertling, who, as a Catholic, was particularly acceptable to the Centre Party. But Hertling, an old man, was allowed no influence on policy.

The German Empire was to have one last year of illusory success before its final collapse. Allied offensives on the western front were routed (1917) and Russian forces disintegrated, especially after the Bolshevik Revolution in November 1917. The Bolsheviks' belief that by offering a peace "without indemnities or annexations" they could induce the German workers to revolution proved unfounded; although a wave of strikes occurred in Germany early in 1918, they produced no political effect. The Social Democrats, who had declared themselves to be fighting against Russian tsardom, continued to fight after tsardom had disappeared. Thus, the Bolsheviks had to sign (March 1918) the Treaty of Brest-Litovsk, by which Russia lost 56,000,000 inhabitants, 79 percent of its iron, and 89 percent of its coal production. In May, Romania had to agree to a peace treaty that made Germany its economic master. Thus, for a few months, Germany had all of Europe east of the Rhine under its economic domination.

Treaty of Brest-Litovsk

The decisive battle remained, however, to be fought in the west. Ludendorff launched what he called "the emperor's battle" (much against the Emperor's wish) in March 1918, winning a battle against the British in April and against the French at the end of May. But these were not decisive, and in July the French struck back. The British broke through on August 8. The collapse of Bulgaria late in September and the imminent collapse of Austria-Hungary caused Ludendorff to decide to seek an immediate armistice. Approaches were made to the United States, whose president, Woodrow Wilson, had enumerated in January 1918 a list of Fourteen Points that he regarded as essential for a just and lasting peace. In order to make Germany's position more acceptable to the Allies, Ludendorff ordered that it should become a constitutional monarchy, and Prince Max of Baden, noted internationally for his liberalism, was made chancellor (October 3). The same day the political leaders were told that the war was lost. Ludendorff, who had not studied the Fourteen Points, wished to continue the war when he realized their implication, but he was overruled and resigned (October 26). Hindenberg remained at the head of the general staff.

The situation deteriorated in November, with the capitulation of Turkey and of Austria (November 3). The next day revolution broke out in Germany, and, on November 9, Prince Max ceded the chancellorship to the Socialist Friedrich Ebert. On the same day the abdication of William II was announced, and he fled to The Netherlands. The German delegates had already left Berlin and, with revolution at home as well as an untenable military position, were obliged to sign the Armistice on November 11.

Defeat of revolutionaries, 1918–19. A republic was hastily proclaimed on November 9 by a Social Democrat, Philipp Scheidemann, in order to forestall the proclamation of a soviet republic by Karl Liebknecht, leader of the extreme left-wing group known as Spartacists. The next day Ebert became chairman of the Council of People's Representatives, a body dominated by so-called Majority Social Democrats, who were opposed to revolution. Ebert made a bargain with Hindenburg, by which it was agreed that Ebert would resist revolution and Hindenburg would retain his command. Ebert then arranged for elections for a constituent assembly to take place on January 19. In the intervening period, the government survived various disorders and attacks, culminating in several days of street fighting in Berlin (January 6–15). Liebknecht and another prominent Spartacist, Rosa Luxemburg, were arrested and murdered. Thus the forces of social revolution were defeated, and the way was clear for the establishment of a democratic republic to preserve the economic order and military values of imperial Germany.

THE GERMAN REPUBLIC, 1919–33

The Weimar constitution. The new National Assembly met at Weimar on February 6, 1919. Ebert was elected president of the *Reich*, and Philipp Scheidemann as chancellor formed a ministry based on support from his own party, the Social Democrats, allied with the Centre and Democratic parties. The new constitution, promulgated in August, retained for the republic a federal basis, but the central government was much stronger than in imperial times, controlling all taxation and its laws overriding those of the separate *Länder*, or states. There were 17 *Länder* in all, ranging from Prussia, with a population (in 1925) of 38,000,000, to Schaumburg-Lippe, with 48,000. The only new *Land* was Thuringia, formed from the amalgamation of seven small principalities. The *Länder* were represented in the Reichsrat, but this chamber was subordinate to the Reichstag, to which alone the chancellor and his government were responsible. The Reichstag was to be chosen on a system of proportional representation, by all men and women over the age of 20.

As a counterweight to the Reichstag, the president as the chief executive was endowed with strong powers. He was to be elected by direct national vote for a term of seven years and was eligible for re-election. He was to make alliances and treaties, was the supreme commander of the armed forces, and could dissolve the Reichstag and submit any law enacted by it to a referendum. Finally, under the famous Article 48, he had the right to suspend civil liberties in case of emergency.

Powers of the president

The new German Republic was democratic but not Socialist. German industry continued to be marked by cartels and combines, control of which was increasingly concentrated in the hands of a very small group. The fact that there were no far-reaching plans for securing public control of industry or for breaking up the big landed estates had two serious consequences. The working class, although improving its political and economic status, remained resentful and increasingly supported the left-wing opposition, thus weakening both the Social Democratic Party and the republic. Meanwhile, economic power remained in the hands of those who either opposed the republic outright or, giving it nominal support, preferred more authoritarian forms of government.

The Treaty of Versailles. To many Germans, the break with the past was complete, and they imagined the program of self-determination and equality of rights originally set out in Pres. Woodrow Wilson's Fourteen Points to be binding on both sides. In the event, the Allies refused them the right to negotiate, and the terms presented on May 7, 1919, provoked bitter indignation throughout Germany.

Germany was to cede Alsace-Lorraine to France; Upper Silesia, most of Posen, and the so-called West Prussia to Poland; North Schleswig to Denmark; and three small frontier districts to Belgium. Danzig was to become an independent, free city; East Prussia was separated from the rest of the *Reich* by Polish Pomorze (modern Pomerania); and Memel was handed over to Lithuania. The union of Austria with the *Reich*, which was advocated in both countries, was expressly forbidden. All German colonies were to be handed over to the Allies. The left (German) bank of the Rhine and the right bank to a

depth of 30 miles (50 kilometres) were to be permanently demilitarized. The rich area of the Saar was to be governed by the League of Nations for 15 years and its coalfields administered by France. At the end of that time a plebiscite was to determine the Saar's future status. To ensure execution of the treaty's terms, Allied troops would occupy the left bank of the Rhine for a period of five to 15 years. The German Army was to be limited to 100,000 officers and men; the general staff was to be dissolved; vast quantities of war material were to be handed over and the manufacture of munitions rigidly curtailed. The navy was to be similarly reduced, and no military aircraft were allowed.

Reparations

A decision on financial reparations was deferred until 1921, but Germany was to make a provisional payment of 20,000,000,000 marks in gold, as well as deliveries in kind. Prewar commercial agreements with other countries were cancelled; German foreign financial holdings were confiscated, and the German merchant marine was reduced to less than one-tenth of its prewar size. As justification for these reparation claims, the Allies inserted into the treaty the famous war-guilt clause, Article 231, by which Germany and its allies were named the aggressors in the recent war and were held responsible for all loss and damage suffered by the Allies.

All the German political parties protested against these terms, and the belief that Germany had been tricked into signing the Armistice became widespread. But in June the Allies presented an ultimatum, and the German government faced the alternatives of either signing the treaty or suffering military invasion. Scheidemann, who opposed acceptance, resigned when his cabinet failed to agree. He was succeeded by Gustav Bauer, who formed an administration supported by Social Democrats and the Centre, but without the Democrats, most of whom joined the Nationalists (Deutschnationale Volkspartei) and the People's Party (Deutsche Volkspartei) in opposition. On June 23 the majority of the Reichstag, seeing no alternative, voted for acceptance of the terms, and the treaty was signed on June 28.

The Allies' insistence on terms so humiliating seriously weakened the republican regime. The legend that the German Army had never been defeated but was stabbed in the back by the Republicans, the Socialists, and the Jews was assiduously repeated by enemies of the republic and, in the mood of resentment created by the treaty, was increasingly accepted.

Years of crisis, 1920–23. The "Weimar coalition" of Social Democrats, Centre, and Democrats was re-established when the Democrats joined Bauer's government in October 1919 and was maintained under Hermann Müller, who succeeded Bauer as chancellor in March 1920. But, in the elections of June 1920, these parties lost heavily in favour of some more to the left and others more to the right. A Centrist, Konstantin Fehrenbach, became chancellor, leading a coalition of the Centre, the Democrats, and the right-wing People's Party.

There were some workers' risings in 1920 and 1921, but the danger to the republic from the left gradually became less. After the left wing of the Independent Social Democrats joined the Communists in December 1920, the remaining Independents drew closer to the Social Democrats, and a union of the two parties was achieved in September 1922. By far the greater danger to the republic came from the right. A coup d'etat, attempted in March 1920 by the conservative politician Wolfgang Kapp, was only defeated by a solid working class resistance and the organization by the trade unions of a general strike. Inside the Reichstag the Nationalists kept up an unrestrained attack upon the republic and its leaders, but even they were not sufficiently extreme to satisfy such groups as the German Racist Freedom Party in the north and Adolf Hitler's National Socialist German Workers Party in Munich, which combined anti-Semitism, anti-Communism, and inflammatory nationalism with open demands for the overthrow of the republic. Closely linked with these groups were various paramilitary organizations, drawing their membership largely from former servicemen and in close touch with army officers who provided them with arms. From this underworld of conspiracy, which was the breeding ground of the Nazi movement, were recruited gangs such as the notorious Organisation Konsul, responsible for several political assassinations. The stronghold of these counter-revolutionary forces was Bavaria, also the home of the Catholic Bavarian People's Party, which made no secret of its anti-republican and particularist views.

On April 27, 1921, the Allied Reparation Commission fixed the total to be paid by Germany at 132,000,000,000 gold marks. Regarding this sum as far in excess of what the country could pay, Fehrenbach and his government resigned, but, in face of an Allied ultimatum based on a threat of occupation of the Ruhr, the new chancellor, Karl Joseph Wirth (Centre), secured a reluctant vote from the Reichstag in favour of paying. It proved, however, impossible to pay the sums required on time, and the French, who had already engaged in frontier harassment, made a technical default by Germany in timber deliveries the pretext for occupying the Ruhr (January 1923).

One way of breaking the hostile ring with which the Germans felt themselves encircled was to ally with the other European outcast, the Soviet Union. Economic negotiations proved successful in 1921, and in April 1922 a treaty of friendship between Germany and Russia was signed at Rapallo. Reparation claims by both sides were abandoned, and the expansion of Russo-German trade was planned. The most important practical consequence of the treaty was the conclusion of secret agreements between the German and Russian armies, by which German units obtained experience with the Red Army in the use of weapons and aircraft forbidden to Germany by the Treaty of Versailles.

The Treaty of Rapallo

During these immediate postwar years, the value of the German mark steadily deteriorated, largely as a result of reparation payments and the savage penalties imposed on German trade. Faced with budgetary deficits, governments followed a practice already begun during the war, that of issuing more money to meet expenses. The result was a runaway inflation more severe than that experienced in any other part of Europe. During the year 1922, the value of the mark in terms of the U.S. dollar fell from 162 marks to more than 7,000 marks (the pre-1914 relation had been 4.20 marks to the dollar).

This serious situation was vastly worsened as a result of the French occupation of the Ruhr. The government of Wilhelm Cuno (Centre, People's Party, and Democrats), which succeeded that of Wirth in November 1922, ordered passive resistance to the French and Belgian attempts to get the mines and factories working and a ban on all reparation deliveries. The occupation forces replied with mass arrests and an economic blockade that cut off not only the Ruhr but the greater part of the occupied Rhineland from the rest of Germany. This was a devastating blow to the German economy, especially after the final loss in 1921, as a result of a plebiscite and a League of Nations' ruling, of the coal-mining and industrial areas of Upper Silesia. As a result, by July 1, 1923, the mark had fallen to 160,000 to the dollar; by November 20, 1923, it was down to 4,200,000,000,000 to the dollar. Barter replaced other commercial dealings, and food riots broke out. The heaviest losers were the middle classes and pensioners; but the drop in real wages also hit the workers hard. On the other hand, many businessmen made large profits, and everyone with debts to pay off gained immensely.

The extremist parties hastened to exploit this dire situation. Communists were active in Saxony and Thuringia and led a rising in Hamburg (October 1923). At the same time, the *Land* government in Bavaria openly defied the central government, while Hitler and the small National Socialist movement were urging the Munich authorities to stage a march on Berlin. In face of these troubles, Social Democratic pressure for a stronger policy led to the replacement of Cuno (August 1923) by Gustav Stresemann of the People's Party, who was supported by the Centre, the Democrats, and (until November) the Social Democrats. Stresemann courageously called off the cam-

paign of passive resistance in the Rhineland and lifted the ban on reparation deliveries. Declaring an emergency, he used the army to suppress the leftist governments in Saxony and Thuringia, put down the Hamburg rising, and took a strong line with the Bavarians. An attempted coup by Hitler (the Beer Hall Putsch) in Munich (November 8–9) failed to influence the authorities there, who suppressed the Nazis and hastened to make their peace with Berlin.

Attempts to stabilize the republic, 1923–30. The history of Germany in the mid-1920s is of a gradual return to normality, shown at home by the surmounting of the inflation problem and abroad by the establishment of friendly relations with the Allies, the withdrawal of the occupying forces from the Ruhr and the Rhineland, and the acceptance of Germany as a full member of the League of Nations. The election of Hindenburg as president of the republic on Ebert's death in February 1925 at first aroused concern among republicans and abroad, because of his monarchist and right-wing views; in fact, by his loyalty to the constitution during his first five years of office, he made an important contribution to the republic's stability.

Coalition governments

Throughout the years 1924 to 1928, Germany was governed by a succession of coalition cabinets, based on the support of the three bourgeois parties, the Centre, the Democrats, and the People's Party. Stresemann, defeated on a vote of confidence by an alliance of his enemies on the extreme left and right, had resigned in November 1923; his successor, Wilhelm Marx (Centre), was chancellor until January 1925. Marx was followed by Hans Luther, a former finance minister, but returned to a further term of office as chancellor from May 1926 to June 1928. The election of 1928 showed a swing to the left, and Hermann Müller, a Social Democrat, became chancellor, receiving additional support from the Democrats, the People's Party, and the Centre. During these years opposition, apart from that of extremists, came mainly from the Social Democrats before their accession to power and from the Nationalists. The former supported current foreign policy but criticized the continued concentration of economic power in the hands of a few industrialists and businessmen; the latter steadily criticized all attempts by German governments to comply with the terms of the Treaty of Versailles.

Before he fell from power as chancellor, Stresemann tackled the problem of inflation, introducing (November 1923), in strictly limited amounts, a new Rentenmark, for which cover was provided by mortgages on the entire industrial and agricultural resources of the country. Financial stability gradually returned; and in August 1924 the national bank (Reichsbank) was made independent of government control, and a new Reichsmark currency worth 1,000,000,000,000 original marks was issued. Foreign investment, mostly in the form of short-term loans, flowed back into Germany; industry was re-equipped, production boomed, and unemployment declined.

Relations with the Allies became easier, and, largely with the help of United States financiers, final decisions were reached on reparations. An interim arrangement, the Dawes Plan (1924), provided for a resumption of payments by Germany on a scale beginning at 1,000,000,000 gold marks a year, rising to 2,500,000,000 in 1928 and subsequent years. To help the German government at this stage, a foreign loan of 800,000,000 Reichsmarks was floated. The Dawes Plan had not tackled the question of Germany's total liability; this was studied in February 1929 by a committee, on which Germany was represented, presided over by the American Owen D. Young. As a result of its recommendations, it was finally agreed that the total amount payable by Germany in reparation would be 121,000,000,000 Reichsmarks, payable in 59 annual installments.

The Dawes and Young plans

Germany's acceptance of this liability was the occasion of particularly violent but unsuccessful criticism by the Nationalists, under their leader Alfred Hugenberg, and by the Nazi Party, which now, for the first time, made its mark in national as opposed to purely Bavarian politics.

The gradual removal, during these years, of foreign troops and controls and Germany's international rehabilitation were largely the work of Stresemann, foreign minister from November 1923 until his death in October 1929. His policy of "fulfillment" of Germany's obligations under the Treaty of Versailles was often attacked, but he always obtained other concessions for Germany and did a vast amount to restore his country to international prestige. He was particularly prominent in League of Nations disarmament discussions. In return for Germany's acceptance of the Dawes reparation plan, he secured French withdrawal from the Ruhr in 1924. Improved relations with France led to the conclusion of the Locarno Pact in December 1925, in one of the clauses of which Germany reaffirmed its renunciation of Alsace-Lorraine and undertook not to attempt by force any alteration of its frontiers with France and Belgium. These frontiers were also guaranteed by other signatory powers, including Britain and Italy. The signing of these treaties was followed by the Allies' evacuation of the first (Cologne) zone of the occupied Rhineland and by Germany's entry into the League of Nations (September 1926).

Finally, the Allied Military Control Commission withdrew from Germany in January 1927, and, as a result of agreements connected with the Young committee's decisions on reparations, the Allies promised final evacuation of the Rhineland by June 1930.

The end of the republic. German prosperity in the late 1920s, although real, was largely dependent on foreign credit. This was already being withdrawn early in 1929, and, with the crash of the New York Stock Exchange in October and the beginning of the world Depression, Germany was plunged into a slump more severe than that experienced by any other country. The Depression had immediate political repercussions, producing a notable increase in the support given to extremist parties, both of the right and of the left. Within two years the Nazis became the largest, and the Communists the third largest, of the parties. The Depression was the indispensable condition for the Nazis' rise to power.

The Depression

The immediate consequence of the slump was the breakup and resignation of Müller's coalition cabinet. The new chancellor, Heinrich Brüning, unable to get his budget through the Reichstag because of combined opposition from the Social Democrats, Communists, Nationalists, and Nazis, resorted to the emergency powers available under Article 48 of the constitution and put his economic program into effect by decree (July 1930). Such a development had been envisaged by the group of advisers to Hindenburg, prominent among whom was Gen. Kurt von Schleicher, head of the Ministry of Defense, who had suggested Brüning as chancellor. Defeated anew in the Reichstag, Brüning dissolved it and ordered fresh elections (September 1930). These were conducted in an atmosphere of public disorder, created largely by the Nazis and Communists, and resulted in a sensational rise in the vote for these two extremist parties.

Brüning decided to remain in office, nevertheless, and obtained sufficient support because the Social Democrats, alarmed at the advance of the extremists, now voted for him. His measures failed to check the slump, which was a worldwide phenomenon. A plan to create an Austro-German customs union (March 1931), which would have been a move in the breaking down of tariff barriers and also would have placated those in both Germany and Austria who wanted the countries combined, was wrecked by French and Italian opposition.

In July 1931, the big Darmstätter und Nationalbank failed, and, early in 1932, the number of unemployed rose to more than 6,000,000. In these circumstances, the prospect of a presidential election when Hindenburg's term of office expired was alarming. Attempts to extend the existing term were rejected, and, in an election held in March 1932, Hindenburg was opposed by Hitler and three other candidates. Hindenburg failed to obtain the absolute majority necessary, and, at a second election in April, he polled 19,359,642 votes against 13,417,460 cast for Hitler. In the same month, the increasing power of

the Nazis was shown when they became the largest single party in the Prussian Landtag.

Once Brüning had secured the re-election of Hindenberg, his usefulness in Schleicher's eyes had been exhausted. Hindenburg was persuaded to withdraw his confidence from Brüning (May 1932), who was replaced as chancellor by Franz von Papen, who, nominally a member of the Centre, was repudiated by his party and formed a nonparty cabinet in which Schleicher held office as minister of defense. Although unpopular in the country, Papen had presidential and army support and, in foreign policy, achieved a success for which Brüning had vainly striven; at a conference held in Lausanne, in Switzerland, in June 1932, reparations were virtually abolished in return for a payment of 3,000,000,000 Reichsmarks into a fund for European conservation.

Nazi rise to power

Elections held in July 1932 gave the Nazis 230 seats in the Reichstag. When it met in September, they outvoted Papen, and he dissolved the chamber, deciding to govern by emergency decree. In fresh elections held in November, the Nazis lost nearly 2,000,000 votes, but the prospect of deadlock continued. Schleicher decided to supersede Papen (December 1932), but, when he too failed to win support from moderate parties and did not pursue a course he had earlier advocated, of bringing the Nazis into the government, Hindenburg refused to allow him to dissolve the Reichstag. Meanwhile, Papen had made contact with Hitler and with the Nationalists and was able to offer the President the prospect of a coalition that could command a majority in the Reichstag.

In this bargain, Papen thought he had tied Hitler's hands by creating a coalition in which Nazi ministers were heavily outnumbered and held no key posts. Hitler was forced by the declining fortunes of his party to accept considerably less than he had demanded earlier in 1932. But he secured the chancellorship for himself, attaining this office (January 30, 1933) legally, as he had been determined to do, and not by revolution.

THE THIRD REICH, 1933–45

The Nazi revolution. Hitler's first step was to persuade the cabinet to agree to new elections, in order to achieve a majority in the Reichstag; he promised that, whatever the result, the composition of the coalition would remain unaltered. The elections were fixed for March 5, 1933, and the Nazis made full use of all the power they now possessed to launch an overwhelming campaign. The radio was extensively used for Nazi propaganda, and the meetings of other parties were broken up, and their newspapers suppressed. Hermann Göring, a *Reichsminister* without portfolio and Prussian minister of the interior, carried through a purge of the Prussian police force, bringing it effectively under Nazi control. The police were forbidden to interfere with the actions of the Nazi Sturmabteilung (SA, the Nazi militia), who were given the "freedom of the streets."

The Reichstag fire

On the night of February 27, the Reichstag building was burned down, and, on the pretext of a Communist plot to seize power, the *Reich* government assumed emergency powers. Despite this atmosphere of insecurity, the Nazis failed to gain an overall majority in the elections, both the Centre Party and the Social Democrats maintaining their strength. It was only with the support given by the 52 Nationalist members, in addition to that of the 288 Nazis, that Hitler achieved a bare majority in a Reichstag of 647 deputies.

Hitler's next step was to pass the Enabling Act, which would allow the government to issue decrees independently of Reichstag or president. For this he needed a two-thirds majority in the Reichstag. By arresting or excluding the 81 Communist deputies and by buying with promises and assurances the support of the Nationalist and Centre parties, he easily achieved this; the Social Democrats, sole opponents of the bill, were outvoted by 441 to 94. The Enabling Act remained the constitutional basis of Hitler's dictatorship in the Third Reich (*i.e.*, "empire"; the Second Reich was that created by Bismarck in 1871, while the first denoted the Holy Roman Empire of 962–1806).

Thus armed with overriding powers, Hitler proceeded to carry through a revolution with the authority of the state on his side. A series of decrees abolished the *Land* diets, transferring their sovereign powers to the central government. In May 1933, trade unions were suppressed, and their organizations merged in a German Labour Front. Soon afterward all political parties other than the National Socialist (Nazi) Party were suppressed. Cabinet opposition to these various measures crumpled before the wave of violence that swept the country. Official Nazi power was strengthend by the inclusion of Joseph Goebbels in the cabinet as minister of public enlightenment and propaganda, and Göring replaced Papen, who had been commissioner for Prussia. But in fact the cabinet had already ceased to matter, since all decisions were taken by the Nazi leaders on their own authority. The National Party was unable to retain its identity, and its leader, Hugenberg, resigned.

Hitler, for the sake of the country's economy, could not afford to quarrel with the industrialists and financiers and appointed as Hugenberg's successor at the Ministry of Economy the director of the largest insurance company in Germany. He was also anxious to enlist the support of conservative forces in resisting attempts by the more radical section of the Nazi movement to bring about a so-called second revolution. Hitler needed an accommodation with the army, still the most powerful independent institution of the country. He was anxious for its help in carrying through the rearmament of Germany and knew that he needed its support for his plan of succeeding to the presidency, which included the Supreme Command of the armed forces, when Hindenberg, now in his 87th year, should die. There was already conflict between the generals and the radical Nazis, whose leader, Ernst Röhm, chief of staff of the Nazi SA, was pressing for the SA to be incorporated into the army.

The tension and animosities came to a head in June 1934 when Hitler, visiting the dying Hindenburg, was met by an uncompromising demand from Gen. Werner von Blomberg, minister of defense, that, unless the government achieved a relaxation of the tension forthwith, Hindenberg would hand over power to the army. Hitler therefore reluctantly decided to act against Röhm and the SA; Röhm's most notable enemies, Göring and Heinrich Himmler (the leader of the SS, the party's special guard), planned the purge. On the weekend of June 30, 1934, Röhm and his chief lieutenants were seized and executed without trial. The opportunity was also taken to eliminate other enemies, including Schleicher. Thus the army's demands were satisfied, and, when Hindenburg died on August 2, it concurred in the merging of the offices of president and chancellor. Hitler assumed the title of *Führer und Reichskanzler* (literally "leader and state chancellor") and was confirmed in his new office by a plebiscite held on August 19.

The purge of June 1934

The totalitarian police state. The period 1934 to 1939 saw the steady elaboration of a totalitarian police state in Germany. The principal instrument of control was the unified police, security, and Schutzstaffel (SS) organization, run by Himmler. All educational and artistic institutions and media were forced into the pattern of Nazi regimentation, and, through schools and the compulsory Hitler Youth movement, determined efforts were made to indoctrinate the young. A concordat that the Vatican had signed with the German government in July 1933 did not protect German Catholics from constant harassment, and the refusal of German Protestants to accept a Nazi-sponsored German Christian movement brought them into an equally bitter conflict with the state.

The regime showed particular hostility toward the Jews. In April 1933 they were dismissed from government service and the universities and were debarred from entering the professions. The Nürnberg Laws of September 1935 forbade marriages between Jews and persons of German blood, and the Jews virtually lost all civil rights. Their persecution reached its climax in the pogrom of November 9–10, 1938, carried through by the SS. The greater part of all Jewish property was confiscated, and the surviving Jews were restricted to ghettos until World War

II, when they were systematically exterminated. Altogether, in German-occupied Europe, 6,000,000 out of a total of about 8,300,000 Jews were killed or died in concentration camps.

Hitler dealt with unemployment through a major program of public works, such as afforestation and road building; and, from 1935 onward, massive rearmament rapidly changed the problem to one of acute labour shortage. All other considerations were sacrificed to Hitler's demand for the rearmament of Germany at double the rate that the military and economic experts considered feasible. The four-year plan that Hitler proclaimed in September 1936 and that Göring was commissioned to carry out was regarded as unduly reckless by Hjalmar Schacht, minister of economic affairs and president of the Reichsbank; he resigned the former office in November 1937 and the latter in January 1939. Germany's expenditure on armaments rose from less than 2,000,000,000 Reichsmarks in 1933–34 to 16,000,000,000 in 1938–39.

In 1939 Hitler finally managed to bring under his control the two German institutions that had so far retained their independence—the army and the foreign service. By attacking circumstances of their private lives, he secured the removal of Blomberg from the office of commander in chief and of Gen. Werner von Fritsch from the Ministry of Defense. Hitler himself became commander in chief of the armed forces and replaced the Ministry of Defense by a separate high command that in fact acted as his personal staff. Sixteen senior generals were retired. Henceforth until the end of the war, Hitler's arbitrary power over Germany was complete.

Foreign expansion and defeat. *Hitler's early foreign policy.* Hitler's unswerving aim was to overthrow the 1919 peace settlement and establish a German hegemony in Europe. The principles of the League of Nations were quite alien to such an aim, and in October 1933 Germany withdrew from the League, ostensibly in disgust at the member nations' slowness in pursuing its disarmament policies. Although Hitler declared, after the reunion of the Saar with Germany as the result of a 90 percent plebiscite vote in favour of return, that Germany had no further cause of dispute with France, he avoided participation in schemes for a general European settlement. His announcement (March 1935) that Germany was reintroducing conscription in order to create an army of 36 divisions, a flagrant breach of the Treaty of Versailles, aroused nothing beyond protests from the other powers, and Hitler was encouraged to take bigger risks in the future.

Conscription reintroduced

During subsequent years Hitler played off the rivalries between other powers. A naval treaty made with the British (June 1935) annoyed the French, and he soon became the principal beneficiary of the quarrel between Italy and the Western powers over Ethiopia. The outbreak of the Spanish Civil War in 1936 enabled him to establish close working relations with Benito Mussolini, the Italian dictator. Meanwhile, while ratifying the treaty of mutual assistance with Soviet Russia, he had denounced the Locarno Pact made ten years earlier and remilitarized the Rhineland (1936). International acquiescence in this second open breach of the Versailles Treaty not only further increased his confidence but also weakened the complex system of alliances that France had built up in eastern Europe. When, in November 1937, Hitler's anti-Comintern (Communist International) Pact (1936) with Japan was joined by Italy, and German rearmament had made considerable progress, Hitler felt able to take the offensive in foreign policy. He was convinced that France and Britain would never fight; moreover, he had won the close cooperation of Italy and had driven a powerful wedge between the Soviet Union and the Western powers.

Annexation of Austria, Czechoslovakia, and Poland, 1938–39. Hitler's first objective was the annexation of Austria. An attempted Nazi coup there had been unsuccessful in 1934, but his alliance with Italy had now removed the major external support that those Austrians opposed to a union could have found. Kurt von Schuschnigg, the Austrian chancellor, was bullied by Hitler into acceding to far-reaching demands (February 12, 1938) but subsequently decided to conduct a plebiscite in Austria on the issues. Hitler acted quickly to avert this, German troops occupying Austria on March 12, 1938, the day before the plebiscite was to be held. Once again, the other powers merely protested, and Hitler turned rapidly to his second objective, Czechoslovakia. The demands of the Sudeten German minority in Czechoslovakia were skillfully used by Hitler to cause France and Britain to help him by bringing pressure on the Prague government. At the Munich Conference (September 29–30, 1938), the British prime minister, Neville Chamberlain, secured the cession of Sudetenland to Germany by the Czechoslovaks. But this was not all that Hitler wanted, and, in March 1939, he used the smoldering quarrel between Slovaks and Czechs as a pretext for occupying the whole of Bohemia and Moravia. In the same month he secured the return of Memel from Lithuania to Germany.

Hitler then turned to Poland, renewing demands for the return to Germany of the free city of Danzig and for the construction across Polish Pomerania of a German road and railway to link East Prussia with the rest of Germany. The demands were uncompromisingly rejected, and the British government, which had abandoned its policy of appeasement after the occupation of Bohemia-Moravia, guaranteed support for Poland in the event of aggression, and France followed suit. In the face of this diplomatic move, Hitler in May strengthened his alliance with Italy. Finally, he disrupted negotiations the British and French were trying to conduct with the Russians by making secret counterproposals that resulted in the signing of the German–Soviet Nonaggression Pact on the night of August 23–24. To the public pact of nonaggression was appended a secret treaty dividing the whole of eastern Europe into spheres of influence and partitioning Poland.

The Nazi–Soviet pact

Hitler had assumed that his Moscow pact would cause the British and French to withdraw their guarantees to Poland. When, on the contrary, Britain signed a pact of mutual assistance with Poland (August 25), Hitler tried to inaugurate negotiations. But the British would not bring pressure on the Poles as Chamberlain had done on the Czechoslovaks, and on September 1 the German Army invaded Poland. Hitler did not respond to the British and French ultimatum demanding withdrawal of the invading army, and on September 3 Great Britain and France declared war on Germany.

World War II, 1939–45. Hitler intended merely to wage a localized war against Poland and then make a peace settlement. But the speed and ease of his conquest tempted him to extend the war to the west. No fighting took place during the winter of 1939–40, but in April the Germans occupied Norway and Denmark, partly to safeguard their vital iron-ore supply route from northern Sweden through the Norwegian port of Narvik and partly to secure the Baltic Sea. The invasion of The Netherlands, Belgium, and France began on May 10, 1940. Breaking through the Ardennes sector of the Allies' front toward the English Channel coast, the German troops cut off the French and British fighting in Belgium. The Dutch and Belgian armies surrendered before the end of May, the British had to evacuate from Dunkirk, and in mid-June the French requested an armistice. Although he had had no plans to do so, Hitler then ordered preparations for an invasion of Britain. But the defeat of his air force in the Battle of Britain meant that the essential prerequisite, control of the air, was lacking, and in October 1940 the project was postponed indefinitely.

Encouraged by German successes, Italy had entered the war in June. Its attack on Greece in October 1940 complicated the situation for Hitler because it created a Balkan front of which the British might take advantage. Moreover, the Italians were totally unsuccessful, and Hitler had to help them in both North Africa and the Balkans. In the spring of 1941, with Hungary and Romania already satellites, Germany took over key posi-

tions in Bulgaria and invaded and occupied both Yugoslavia and Greece, including Crete.

In the view of Adm. Erich Raeder, commander in chief of the navy, and of other service leaders, these successes should have been followed by an all-out attack on the whole British position in the Middle East. But Hitler had already determined to attack the Soviet Union. The invasion began in June 1941, and at first the Germans drove deep into Soviet territory. But, at the end of the year, the horrors of Russian winter and an unexpected counteroffensive brought to a head the strained relations between Hitler and the army leaders. He assumed direct control of the field armies, and his success at any rate in holding the Soviet attacks during the winter increased his belief in his own military genius.

The Japanese attack on the United States' base at Pearl Harbor in December 1941 had made the war worldwide, and Germany declared war on the United States. Hitler failed to grasp the importance of sea power, and it was only in 1942 that Karl Dönitz, who was to succeed Raeder as naval commander in chief in January 1943, was able to persuade him to concentrate on submarine warfare. Allied shipping was hard pressed, and it was not until the end of 1943 that the Allies could be said to have won the Battle of the Atlantic. Meanwhile, in Russia a German drive to the Volga turned into a desperate contest for the city of Stalingrad (Volgograd), where Hitler's obstinate refusal to withdraw in time led to the encirclement and capitulation of the German armies in January 1943. At the same time, the British had renewed their Mediterranean offensive, and in May 1943 more than 250,000 German and Italian troops surrendered in Tunisia. This double defeat represented the turning point of the war.

Battle of Stalingrad

At the height of his success, Hitler was the master of the greater part of the European continent, with most states either directly occupied or in the position of subservient satellites. German exploitation of these territories was ruthless. Persons likely to provide leadership were, together with the Jews, exterminated, and vast numbers of people were deported to work in Germany. By the end of 1944, about 4,795,000 foreign workers had been recruited in this way, mainly Russians, Poles, and French. In certain occupied countries, the Germans encountered partisan guerilla warfare; almost everywhere there was some form of resistance movement.

In Germany itself the impact of war was not sharply felt until 1942, when casualties in Russia had first reached the scale of World War I. The effects of the Allies' naval blockade were reduced by the plundering of the occupied countries. Total mobilization was introduced early in 1942, and war production was maintained and even stepped up despite heavy air attacks on industry and communications. Göring, originally Hitler's second in command and named successor, had become completely eclipsed as a result of the failure of the air force, for which he was responsible, adequately to check or counter Allied air raids. His place was taken by Himmler, who extended the functions of the SS until it became virtually a state within a state, creating, with the *Waffen* (armed) SS divisions, a rival to the traditional army (*Wehrmacht*).

By the end of 1943 at the latest, Germany's defeat seemed certain to many of its military leaders. But Hitler would not admit it. During the year 1943, Mussolini was overthrown, Anglo-United States forces invaded Italy, and the Russians assumed the offensive. In June 1944 the British and Americans landed in Normandy. Meanwhile, Allied bombing of Germany became more and more extensive and did enormous damage. Hitler, who since mid-1941 had made his headquarters in a remote part of East Prussia, was now completely cut off from the life of the nation; he was scarcely ever seen in public and rarely spoke or broadcast.

Realizing the harm being done by Hitler's refusal to consider surrender, a group of patriots had for some time been plotting to assassinate him. The only institution capable of staging a successful coup d'etat was the army, and one of the principal centres of the plot had been the Abwehr, the army counterintelligence unit. Although the Abwehr was broken up by Himmler in 1943, a small group in the command headquarters of the reserve army became the new conspiratorial centre. On July 20, 1944, one of its members, Colonel Claus von Stauffenberg, placed a bomb carefully concealed in his briefcase under a table during a conference at Hitler's headquarters. But Hitler, although injured, was not among those killed. Attempts by the conspirators to seize power in Berlin and Paris were suppressed.

By late 1944, the Western Allies had reached the Rhine. The Russians swept through the Balkans and, by December 1944, were threatening East Prussia. Early the next year, they launched an attack from the Baltic to the Carpathian Mountains and broke into Germany. In March 1945 the British and Americans crossed the Rhine and entered from the west. At Hitler's command Germany was turned into a battlefield in a senseless campaign to delay the inevitable defeat. Despite talk of fleeing to an underground fortress in Bavaria, Hitler finally refused to leave Berlin. He appointed Admiral Dönitz his successor as head of state and Goebbels chancellor and, so far as is known, shot himself on the afternoon of April 30. Goebbels committed suicide the following day, and Himmler soon afterward. Most of the other Nazi leaders were captured and subsequently tried by the Allies as war criminals. Dönitz tried to negotiate with the Western powers, but the Allies insisted on unconditional surrender, and this was signed on May 7, 1945, to take effect at midnight on May 8–9.

Death of Hitler

GERMANY AFTER WORLD WAR II, 1945–49

With the unconditional surrender the German state ceased to exist, and responsibility for government was assumed by the four occupying powers, the United States, Great Britain, the Soviet Union, and France. Each power occupied a specific zone of Germany; Berlin, although situated in the Soviet zone, was to be occupied by all four powers and was to be governed by an inter-Allied authority. The Allies announced that they would determine the frontiers and status of Germany.

At the Potsdam Conference (July 17 to August 2, 1945), attended by the United States, British, and Soviet leaders, but not by the French, it was agreed that, although a program of decentralization should be carried out, Germany must be treated as a single economic unit. The British and Americans agreed to support at the settlement the Soviet annexation of the northern half of East Prussia; and, for the time being, the "former German territories" east of the Oder and Neisse rivers (Pomerania, Swinemünde, Silesia, and the southern half of East Prussia, as well as Danzig) should remain under Polish administration. These changes involved the loss of 23 percent of German territory as constituted by the Treaty of Versailles.

The Allies' plans for a common policy in Germany proved abortive because of their own divergent interests, and increasing deadlock characterized the meetings of the Allied Control Council. Between 1945 and 1947, attempts to heal the widening rift between Britain and the United States, on the one hand, and the Soviet Union, on the other, proved totally unavailing. Increasingly the British and United States authorities acted on their own initiative in dealing with problems in their own zones of occupation. In 1946 the British accepted an American proposal that their two zones should be merged for economic purposes, and the bizone came into existence on January 1, 1947. In both zones much administrative responsibility had already been handed over to the Germans, and state (*Länder*) governments had been set up. State parliaments (Landtage) were elected in the United States' zone (June 1946), in the British zone (April 1947), and in the French zone (May 1947). The three largest parties to emerge were the Christian Democratic Union, the Free Democratic Party, and the Social Democratic Party, the last named being closely bound up with the revived trade-union movement.

The Allies had undertaken the occupation of Germany with the purpose of rooting out the Nazi regime and de-

stroying the basis of German military power. Hence their original directives laid great stress on political and economic decentralization, disarmament, the arrest of war criminals, and the dismantling of German war industries. They rapidly found themselves forced to deal with quite a different set of problems: economic stagnation, the dislocation of communications, famine, a desperate housing shortage, an unstable currency, and a rampant black market. The problems were rendered much more acute by the influx into Germany of millions of refugees expelled from eastern Europe, partly as the result of decisions made at Potsdam. These problems, together with the sense of insecurity caused by dissension among the allies, added to German demoralization and apathy.

Postwar problems in Germany

The turning point in the postwar history of Germany was the year 1948. In January, the powers and composition of the German Economic Council for the Anglo-American bizone were changed to create the nucleus of a future German government. The council itself was expanded to 104 members, and a second chamber, a Länderrat, consisting of two representatives from each state, was set up. In June a plan of development that would include the French zone was agreed upon. It provided for a federal German government for all three zones and the summoning of a constituent assembly. Relations with the Allies were to be regulated by an occupation statute; an Allied military security board would enforce demilitarization; and an international Ruhr authority would control the coal and steel industries of the area. The Saar, which at the end of 1947 had become an autonomous territory economically attached to France, was excluded from this arrangement.

The Soviet government reacted strongly to these developments and began to interrupt communications between western Germany and Berlin. After a much-needed currency reform in the western zones (June 1948), the Soviet authorities blockaded the Allied garrisons and about 2,500,000 inhabitants of West Berlin, intending to drive out the Western powers. These powers counterblockaded the Soviet zone and organized an airlift to keep West Berlin supplied. The Soviet Union finally lifted the blockade in May 1949, but further disputes made it necessary for the airlift to be continued until September. Finally a tacit working agreement was reached that left Berlin divided into western and eastern sectors under separate administrations, but with the garrisons of the Western Allies still present.

Meanwhile, the plans for development of German government in the western zones had proceeded. A Parliamentary Council of 65 members, elected by the Landtage, met at Bonn on September 1, 1948, to draft a constitution. Their work went on while final attempts to achieve a unified policy for Germany among all four occupying powers again proved unsuccessful. In April 1949 the United States, British, and French foreign ministers published a new occupation statute governing their three respective zones. This guaranteed full powers of self-government to the new West German state, except for certain reserved subjects, such as the question of Berlin. If the political outlook for the new state remained difficult, its economic condition was encouraging. Currency reform and economic aid from the United States under the Marshall Plan had led to a remarkable recovery. In the second half of 1948, industrial production had risen from 45 percent to nearly 75 percent of the 1936 level, while steel production doubled during that year. Although the policy of free competition pursued by the German authorities led to steep price rises and consequent hardship for the poorer classes, in general the revival in German confidence was marked.

THE FEDERAL REPUBLIC OF GERMANY

The Basic Law

The Basic Law of the new state was passed by the Parliamentary Council on May 8, 1949, and four days later was approved by the Western Allied military governors with certain reservations, notably the exclusion of West Berlin, which had been proposed as the 12th *Land* of the federation. The 11 *Länder*, then, were Schleswig-Holstein, Hamburg, Bremen, Lower Saxony, North Rhine-Westphalia, Rhineland-Palatinate, Hesse, Bavaria, Baden, Württemberg-Baden, and Württemberg-Hohenzollern (the last three were merged in 1952 to form Baden-Württemberg, and in 1957, the Saar became the tenth *Land*). When the Occupation Statute came into force in September, the Allied military government was replaced by an Allied High Commission. In November the Petersberg Agreement made further concessions on dismantling industries and allowed the building of a merchant fleet. The Federal Republic became a direct member of the Ruhr Authority and the Council of Europe, and was able to establish consular and commercial relations abroad.

The constitution drawn up for the Federal Republic was not unlike that of Weimar, but it provided far greater democratic safeguards and established by law certain basic individual rights. The positions of president and chancellor were virtually reversed; in the new constitution the president was elected for a five-year term by a federal assembly (Bundesversammlung) made up of both legislative houses. The emergency powers available to the president under the Weimar constitution were now to be wielded by the chancellor and his cabinet. The chancellor, nominated by the president, had then to be elected by absolute majority in the Bundestag, or lower house. This house was elected by direct universal suffrage for a four-year term. The Bundesrat, or the upper house, represented the *Land* governments: no change could be made in the Basic Law without its consent.

The first federal election was held on August 14, 1949. Led by Konrad Adenauer, who had been chairman of the Parliamentary Council, the Christian Democrats won 139 seats out of 402 in the Bundestag. The Social Democrats won 131 seats. Of the other eight parties represented in the Bundestag, the most important were the Free Democratic, with 52 seats, followed by the conservative German Party and by the federalist Bavarian Party, with 17 seats each. The Communists obtained 15 seats. An alliance with the Free Democrats and the German Party enabled Adenauer to be elected chancellor by an absolute majority of one vote in the Bundestag and to form a coalition government. Theodor Heuss, of the Free Democratic Party, was elected president of the republic in September.

Chancellorship of Adenauer. *Foreign policy under Adenauer, 1949–63.* Adenauer, long a strong champion of European cooperation, regarded foreign policy of the greatest importance, and, when (March 1951) the Allied High Commission allowed the Federal Republic to establish a Foreign Office, Adenauer became foreign minister himself, as well as chancellor. Even when, in 1955, Heinrich von Brentano was made foreign minister, important decisions remained reserved to the Chancellor.

Adenauer's avowed aim was the reunification of Germany, but he regarded this as the Allies' responsibility, and continued differences between the Western and the Communist powers made its achievement unlikely. In this situation Adenauer supported the Western alliance under United States leadership and worked to establish a federal relationship between the countries of western Europe. The Social Democrats sharply opposed this order of priorities until 1956 and also opposed the idea of German rearmament, pressed on the Federal Republic by the United States. To prevent the re-emergence of a German national army, Sir Winston Churchill and René Pleven (of France) proposed the setting up of a European army, in which German units would be integrated under European command. The Social Democrats consistently opposed this project, until the elections of 1953 gave the government a two-thirds majority in both houses, which meant that, had they needed to amend the constitution on the rearmament issue, they could have done so. But the European Defense Community project was turned down by the French in August 1954, and, following agreements made in Paris in October 1954, the Federal Republic became, in May 1955, a sovereign state and a full member of the North Atlantic Treaty Organization (NATO), pledged to establish its own armed forces. Great Britain joined with

Membership in NATO

the six member countries of the European Coal and Steel Community (established in 1951) in a Western European Union, which was both to supervise German rearmament and to ensure the maintenance of sufficient British forces on the continent.

Adenauer still worked for closer European integration, and in 1955 Germany was prominent in the discussions that led in 1957 to the signing of the Rome Treaties establishing the European Atomic Energy Community and the European Economic Community (EEC). The Social Democrats voted for their ratification.

A meeting between Pres. Charles de Gaulle and Adenauer in September 1958 inaugurated a period of cooperation between the two men, which was crowned by a further increase in cordiality in 1962 and resulted in the signing of a Franco-German treaty of cooperation in 1963. The late 1950s also saw some coolness between the Federal Republic and Britain. Adenauer distrusted the approaches that the British prime minister, Harold Macmillan, was making to Moscow. Britain, on its side, was anxious about the number of former Nazis holding government and judicial office in Germany. Some cordiality was re-established in the early 1960s.

Outside Europe, Germany tried hard, especially among the emergent nations of Africa and Asia, to establish itself as a noncolonial power able to aid their economic development. Israel was a special case, receiving more than 3,000,000,000 Deutsche Marks in material reparation for Germany's treatment of the Jews. Throughout the 1950s Germany was on excellent terms with the United States, especially after the 1956 involvement of France and Britain in the Suez. A peak of cordiality was reached in 1963 when President John F. Kennedy visited West Germany and West Berlin.

Adenauer went to Moscow in 1955 and agreed to establish diplomatic relations with the Soviet Union, in return for the repatriation of nearly 10,000 German prisoners of war still held in the Soviet Union. A constant problem was the two countries' divergent views about East Germany (the Democratic Republic). Moscow considered that there were two German states, but the Federal Republic would not recognize the legal existence of the East German government. Although a trade agreement was signed with the Soviet Union in 1958, a crisis developed over Berlin later in that year, when the Soviets demanded the withdrawal of Western troops and wanted to declare Berlin a free city. A series of international meetings failed to solve the conflict, and in 1961 the Soviets authorized the building of the Berlin Wall, separating the eastern from the western sectors.

Domestic affairs under Adenauer, 1949–63. The most urgent internal problem for Adenauer's first administration was the resettlement of refugees. At the end of the war, about 8,000,000 Germans had been expelled from Polish territories or from the Sudeten lands; there were others from the Balkans, and each year tens of thousands of additional refugees were arriving from East Germany. The presence of these refugees put a social burden on the Federal Republic, but their assimilation proved surprisingly easy. Many of the refugees, indeed, were skilled, enterprising, and readily adaptable, and their labour proved an important factor in the "economic miracle" of West Germany's speedy recovery.

German refugees

The first legislative period (1949–53) was marked by the struggle over the Law of Codetermination demanded by the Trade Union Federation. As a result, "labour directors" nominated by the unions were made compulsory in the coal and steel industries. But, in spite of their strongly held views about the distribution of the nation's wealth, in the matter of wage demands the German unions were extremely moderate throughout the postwar decade, thereby facilitating economic growth. German industry was rapidly re-equipping itself, and the business community was thought to have a powerful influence on politics behind-the-scenes.

In the elections of September 1953, the Christian Democrats won 243 seats out of 487 in the Bundestag. Economic progress, Adenauer's personal prestige, and the desire to remain close to the protection of the West were the chief reasons for this success. The Social Democrats won only 151 seats. The Free Democrats, who as junior partners in Adenauer's coalition could neither attack its record nor claim much credit for its achievements, dropped to 48 seats. While the German Party obtained 15 seats, the All-German Bloc, a specifically refugee party, won 27. Adenauer formed a second coalition government with the Free Democrats, the German Party, and the All-German Bloc. Heuss was re-elected president in July 1954.

Conflicts soon developed. The trade unions and the Social Democrats launched a campaign against proposals for rearmament and conscription, although they were not able to prevent the passage in the Bundestag in July 1954 of a law authorizing the formation of a volunteer army. The Conscription Bill was passed two years later, and the first recruits were called up in 1957.

During this term of office, Adenauer lost the effective support of some of the smaller parties in his coalition. Most of the All-German Bloc went into opposition in 1955, and the next year, after a bitter quarrel over an electoral law, the Free Democrats also left the government. But any further erosion of Adenauer's power was averted by external events. The brutal suppression of the Hungarian rising by the Soviet Union in 1956 caused a swing back to the support of a government firmly in alliance with the West, while showing up as impractical various plans for reunification with East Germany that some of the Social Democrats and Free Democrats had been entertaining. Government popularity was also enhanced by the social security system in 1957; in future, pensions were to be tied not only to the cost of living but to the growth in national income. Thus, the government was able to point to impressive achievements when launching its campaign for the elections of 1957; it also claimed much of the credit for Germany's phenomenal economic progress under Ludwig Erhard, the minister of economics. As a result, the Christian Democrats increased their proportion of seats in the Bundestag to 270 out of 497. Adenauer formed his third coalition in alliance only with the German Party, which had won 17 seats in the election.

Again, rearmament proved a bone of contention, the Social Democrats and the trade unions this time protesting against the possible equipment of the German army with nuclear weapons. They also called for the withdrawal of foreign (that is, those of the North Atlantic Treaty Organization) troops from German soil. But the trade unions refused to call a strike over the issue; the constitution allowed no referendum on it; and in 1958 the federal armed forces received dual-purpose missiles capable of carrying atomic warheads. By the end of 1960, there were more than 290,000 men in the German forces.

The year 1959 saw the first attempt to challenge Adenauer's power from within the Christian Democratic Union. The crisis was provoked when, as Heuss's second term of office was drawing to an end, Adenauer proposed to have Erhard nominated to succeed him as president, thereby eliminating him from any contest for the chancellorship. Erhard refused to stand, and, in a surprise move, the party induced Adenauer himself to accept nomination for the presidency. But when he saw that the party would go on and make Erhard chancellor, Adenauer reversed his decision. Though, in the course of the crisis, Adenauer had cast grave aspersions on Erhard's political experience and reliability, he was able easily to crush the revolt in the party. The candidate finally elected president, in July 1959, was Heinrich Lübke, a Christian Democrat.

Adenauer's attempt to exclude Erhard

A notable change in the policy of the Social Democrats began in November 1959, with their adoption of the Bad Godesberg Program. This policy document showed them to have completely abandoned Marxism, while even socialization was scarcely mentioned. Roughly a year later, at Hannover, the party conference confirmed Willy Brandt, the mayor of West Berlin, as its candidate for the chancellorship. Brandt reversed the party's policy on the only major issue that now separated it from the government by accepting both conscription and the possibility

that German troops might be armed with nuclear weapons. The party now differed little in policy from the Christian Democrats and fought the electoral campaign of 1961 chiefly by contrasting Brandt with the octogenarian Adenauer.

The elections of 1961 returned only three parties to the Bundestag: the Christian Democrats with 242 seats, the Social Democrats with 190, and the Free Democrats with 67. Having lost their absolute majority, the Christian Democrats allied themselves with the Free Democrats. Although that party's leader, Erich Mende, wanted a new chancellor, Adenauer was in fact re-elected, after the Free Democrats had elicited an understanding that he would retire before the parliamentary term expired.

Chancellorships of Erhard and his successors. *Erhard.* Adenauer resigned in October 1963, and Erhard was elected chancellor, forming a coalition Cabinet very similar to Adenauer's but including Mende as vice chancellor and minister for all-German affairs. Lübke was re-elected president in July 1964.

Germany's relations with the West were complicated by the problem of reconciling "Atlanticism," European unity, and the Franco-German treaty—particularly as France under Pres. Charles de Gaulle became more estranged from NATO and insisted on the exclusion of Britain from the Common Market. Further difficulties arose over Germany's status in any plan for a "multilateral nuclear force" within NATO; France wished to restrict control of its nuclear deterrent to itself, while Britain did not want to allow Germany authority in decisions about the use of nuclear weapons. The whole plan, however, was abandoned in December 1965 when Erhard had talks in Washington with Pres. Lyndon B. Johnson. German–French relations remained at a stalemate, and relations with Britain, temporarily improved by a state visit from Queen Elizabeth II in May 1965, were clouded by disputes about liability for the financial upkeep of British troops stationed in Germany.

Although the results of the election of September 1965, when the Christian Democrats won an unexpected 245 seats against 202 won by the Social Democrats (the Free Democrats won 49), were regarded as a personal victory for Erhard, his government was running into difficulties. For the first time since the currency reform of 1948, the growth rate of the German economy slackened, and in 1967 industrial production actually fell. Widespread fear of economic recession, coupled with the lack of clear political leadership, led to two developments. In the country at large, the National Democratic Party—a neo-Nazi right-wing group—won some support, especially in Bavaria and Hesse, and in October 1966 the Free Democrats, worried about their own future as junior partners in a government in difficulty, withdrew from Erhard's Cabinet.

Economic difficulties

Kiesinger. The withdrawal of the coalition partner left the Cabinet without a parliamentary majority and the Christian Democrats without a leader who could command adequate parliamentary support. In November Kurt Georg Kiesinger, previously minister-president of Baden-Württemberg, became the candidate for the chancellorship and, after complicated negotiations, effected an unprecedented political alliance with the Social Democrats. Kiesinger was elected chancellor, and Brandt became vice chancellor and foreign minister.

Kiesinger sought a détente with the East and an improvement of relations with France. But, although ambassadors were exchanged with the Soviet Union in 1967, the Soviet occupation of Czechoslovakia in 1968 caused renewed tension. Friction with France not only continued but was increased by Brandt's warm endorsement of Britain's application to join the EEC, which was vetoed in December 1967 by President de Gaulle. Fresh conflict arose over the currency crisis of November 1968, when the Germans not only resisted pressure from France and other countries to revalue the Deutsche Mark but recommended the devaluation of the French franc, a step de Gaulle refused to take.

Despite the fact that the Kiesinger government presided over a return to economic growth and full employment, it encountered vociferous opposition from both right and left. In June 1969 the problem of trials of "war criminals" had to be faced again, since the regulations extended in 1965 would soon run out. A time limit for the bringing of charges of genocide was abandoned, and the limit for murder was extended to 30 years from the end of the war.

Early in 1969, President Lübke announced his resignation, and the decision to hold the new presidential election in West Berlin provoked reprisals from the East German authorities. The elections were held, nevertheless, and the new president was Gustav Heinemann, a Social Democrat.

In elections for the Bundestag held in September 1969, the Christian Democrats won 242 seats, the Free Democrats only 30, and the Social Democrats 224. With these results, Brandt made his bid for the chancellorship.

Brandt. On October 21, 1969, Willy Brandt became West Germany's first Social Democratic chancellor. His coalition government was made possible by the support given by the Free Democrats, who, despite their small showing, were in the position of offering an alliance to give power to either of the major parties. Later, in 1970, the fortune of the Free Democrats declined in local elections, and a split in the party weakened the coalition, reducing Brandt's majority in the Bundestag.

As a left-wing government, the Social Democrats had planned a program of domestic reform. In practice, however, much of their time was occupied in combatting inflation. Their first action on achieving power was to devalue the Deutsche Mark by 9.29 percent; in May 1971 they had to float the Mark against international currencies.

Brandt's most important work, and that which caused the most debate, was his *Ostpolitik,* his attempt to come to terms with Germany's Communist neighbours, notably the Soviet Union, Poland, the German Democratic Republic, and Czechoslovakia. In August 1970 he signed a treaty with the Soviet Union confirming the existing frontiers and pledging the renunciation of force to achieve political ends. It was to be ratified by the Bundestag upon the successful conclusion of talks on Berlin by the four Allied occupying powers—the Soviet Union, the United States, Great Britain, and France. They reached agreement in August 1971, and talks with the East Germans about improving facilities for communications between West Germany and West Berlin, which formed a part of the agreement, began. Brandt had had two meetings in 1970 with the East German premier, Willi Stoph, although little had been achieved; but now talks began in earnest. In October 1972 the two Germanies reached a transportation agreement, and that November they initialled a pact that made possible formal relations and cooperation; as a result, families were reunited, visits allowed, prisoners released, and relations formalized between the two Germanies. Meanwhile, in December 1970, West Germany had signed a treaty with Poland ratifying existing borders, and in December 1973 a similar treaty with Czechoslovakia was signed, thus completing the foundations of the *Ostpolitik.*

Brandt's *Ostpolitik*

Although Brandt's *Ostpolitik* was a major issue in the elections of November 1972, his coalition was re-elected with a majority of 46 seats in the 496-seat Bundestag. Brandt resigned as chancellor in May 1974 after accepting responsibility for negligence in a scandal involving one of his personal aides who was accused of spying.

Resignation of Brandt

Schmidt. It was taken for granted that Helmut Schmidt would succeed Brandt as chancellor, and he did so on May 16. In July Walter Scheel became the fourth president of the republic, succeeding Gustav Heinemann on his retirement. Both Schmidt and Scheel were returned to office in the federal elections of 1976.

The change of leadership in both West Germany and France in 1974 brought to power two men who had become friends as their countries' finance ministers, Schmidt and Valéry Giscard d'Estaing. This friendship formed a new basis for German–French relations. Schmidt's foreign policy emphasized European unity, and his domestic policy stressed economic stability.

THE GERMAN DEMOCRATIC REPUBLIC

For the East German zone a Soviet Military Administration was formed in June 1945, with the dual aim of providing a pro-Soviet regime in Germany and of collecting war reparations. The blocking of all bank accounts reduced everyone to an equal need to find work; on July 26 all private banks were closed, and all companies and individuals were ordered to surrender all currency, bullion, deeds, and valuables. Factories were dismantled, and rolling stock and even the rails themselves were transferred to the Soviet Union. All estates of more than 100 hectares (250 acres) were expropriated without compensation.

Thus, about one-third of the country's agricultural land, previously owned by about 3,000 landlords, passed into the hands of about 544,000 farm workers, smallholders, and refugees.

The Soviet Zone was divided into five *Länder:* Saxony, Brandenburg, Mecklenburg (including part of Pomerania), Saxony-Anhalt, and Thuringia. The political freedom of all "anti-Fascist" parties was proclaimed (June 1945), and, besides the Communist Party of Germany (KPD), three other parties emerged: Social Democrats (SPD), Christian Democrats (CDU), and Liberal Democrats (LDPD). In April 1946 the Communists and Social Democrats merged to form the Socialist Unity Party of Germany (Sozialistische Einheitspartei Deutschlands, or SED); Wilhelm Pieck and Otto Grotewohl were elected joint chairmen, and Walter Ulbricht became secretary general (later first secretary). Later in 1946, the Socialist Unity Party won the largest number of votes in municipal elections and in elections for the Landtage, or provincial councils.

Political parties

In 1948, when the United States and British authorities set up the German Economic Council in their bizone, the Soviet Union created a similar 25-member Economic Commission. At a People's Congress held by the Socialist Unity Party (March 1948), a Volksrat, or People's Council, of 400 members was set up and, later in the year, began to draft a constitution for the German Democratic Republic. The constitution was adopted on May 30, 1949, by another People's Congress of 1,525 elected members. A new Volksrat of 400 was appointed and transformed into a Volkskammer (People's Chamber) in October. A Länderkammer (Chamber of States) was also set up, and the two houses elected Pieck president of the republic; but the upper house disappeared on July 23, 1952, when the five *Länder* were subdivided to form 15 *Bezirke,* or districts. A government was formed, with Grotewohl as premier and Ulbricht as first deputy premier. The Soviet Military Authority formally transferred power to the government and was replaced by a Soviet Control Commission.

In July 1950 a treaty was signed with Poland, recognizing the Oder–Neisse line as the permanent frontier between the two countries.

Gradually Soviet claims became less harsh, and the GDR moved toward autonomy. In May 1950 Joseph Stalin announced a 50 percent reduction in reparations still held to be due. In October that year, the republic was admitted to the Council for Mutual Economic Assistance (Comecon), composed of the Soviet Union and its European allies.

In 1953 the Control Commission was replaced by a single high commissioner, as part of a new policy of concessions to German public opinion. Already, however, the Germans had been exasperated by persecution of the churches, by a serious food shortage, and by collectivization in 1952 of the smaller farms not involved in the agricultural expropriations of 1945.

When the production requirements of workers were raised, a wave of strikes culminated in an uprising on June 17, 1953, and Soviet troops were called in to restore order. Collection of reparations by the Soviet Union ceased in 1954, and on March 25 of that year it was announced that the GDR was a sovereign state. Soon afterward it became a founder-member of the Warsaw Treaty Organization, the unified military command established by the Soviet Union and its eastern European allies.

When Pieck died in September 1960, the presidency was abolished, and a Council of State was formed, with Ulbricht as chairman. He remained the first secretary of the Socialist Unity Party.

As early as 1952, the GDR took steps to isolate its territory from West Germany; a police-guarded cordon of land, three miles wide, was created along the whole frontier.

This action left Berlin as the easiest aperture through which persons dissatisfied with the East German regime could leave the country. Between 1945 and 1961, about 3,500,000 people left East Germany for the West. Though it is estimated that 500,000 of these eventually returned, the loss of 3,000,000 people, many of whom were highly trained technical specialists, severely damaged the East German economy.

To deal with this crisis, the Volkskammer in August 1961 adopted a decree for sealing off West Berlin by building a barbed-wire barrier that was later replaced by a concrete wall.

The Berlin Wall

This closing of the "Berlin gap" marked the beginning of an economic revival in East Germany. Rebuilding of East Berlin, Leipzig, Dresden, Magdeburg, Halle, and other cities was rapid and spectacular. In external affairs, the GDR's position also improved, and it made headway in its struggle for international recognition outside the Communist world. A bilateral treaty of alliance with the Soviet Union was signed in 1964, similar treaties with four other Soviet-allied countries in 1967. On October 15, 1971, the German Democratic Republic submitted to the United Nations a memorandum claiming that its membership in that institution would help to solve the outstanding problems between the two German states. At that time the German Democratic Republic maintained normal diplomatic relations with 30 states, as well as consular or commercial official relations with 32 other states.

The East German claim to recognition by West Germany had long been rejected by West German statesmen, but, from 1966 onward, positive attempts at improved relations were made, especially after Willy Brandt became chancellor of the Federal Republic in 1969. In 1970 Willi Stoph, who had succeeded Grotewohl as premier in 1964, had two meetings with Brandt. In 1971 Walter Ulbricht resigned as party chairman and was replaced by Erich Honecker, who also was named chairman of the Council of State in 1976. The question of mutual recognition between East and West Germany was solved with the treaty signed by the two countries on December 21, 1972. The treaty paved the way for them to be admitted to the United Nations in September 1973, and in June 1974 they exchanged diplomatic representatives. In September of that year the U.S. was the last member of the NATO alliance to grant formal recognition to East Germany. (Ed.)

BIBLIOGRAPHY

Medieval Germany to 1250: GEOFFREY BARRACLOUGH, *The Origins of Modern Germany*, 2nd ed. rev. (1947, reprinted 1966), probably the best history of medieval Germany available in English, its perspective enhanced by the author's study of the Middle Ages as an integral part of German history as a whole; and (ed. and trans.), *Mediaeval Germany, 911–1250*, 2 vol. (1938, reprinted 1967), a collection of essays by notable German scholars that presents a kaleidoscopic picture of political and constitutional developments of the medieval German state in addition to providing a concise history of the period; WILLIAM STUBBS, *Germany in the Early Middle Ages, 476–1250* (1908, reprinted 1969), an excellent analysis of the early medieval period, though dated in some respects; JAMES WESTFALL THOMPSON, *Feudal Germany*, 2 vol. (1928, reprinted 1962), the most thorough examination of the German system of feudalism as well as a competent general history; FRANZ H. BAUML, *Medieval Civilization in Germany, 800–1273* (1969), a somewhat simplified yet richly illustrated investigation of German medieval arts, architecture, culture, and social life; KARL BOSL, *Herrscher und Beherrschte im deutschen Reich des 10.–12. Jahrhunderts* (1963), a short statement of the power relationships in medieval German political and ecclesiastical life.

From 1250 to 1493: WILLY ANDREAS, *Deutschland vor der Reformation*, 6th rev. ed. (1959), a broad survey of German

society on the eve of the Reformation; GEOFFREY BARRACLOUGH (*op. cit.*), valuable for political trends and constitutional issues; LEO JUST (ed.), *Deutsche Geschichte bis zum Ausgang des Mittelalters* (1957), a cooperative work by various specialists in the period; HEINZ ANGERMEIER, *Königtum und Landfriede im deutschen Spätmittelalter* (1966), a critical analysis of the various means employed to preserve the public peace; OTTO BRUNNER, *Land und Herrschaft*, 4th ed. (1959), an original study of the growth of territorial lordship; FRANCIS L. CARSTEN, *Princes and Parliaments in Germany, from the Fifteenth to the Eighteenth Century* (1959), an analysis of the relationship between the territorial princes and the assemblies of estates; FRIEDRICH LUTGE, *Deutsche Sozial- und Wirtschaftsgeschichte* (1952), a wide-ranging work on social and economic conditions in late medieval Germany; MICHAEL SEIDLMAYER, *Currents of Mediaeval Thought* (1959), a thoughtful elucidation of general intellectual trends.

From 1493 to c. 1760: HAJO HOLBORN, *A History of Modern Germany*, vol. 1, *The Reformation* (1964), the best general history of Germany for the period from 1500 to 1648, focussing on political and intellectual developments of the Reformation and Counter-Reformation and the severe conflicts they provoked; JOHANNES JANSSEN, *Geschichte des deutschen Volkes seit dem Ausgang des Mittelalters*, 4 vol. (1883–86; Eng. trans., *History of the German People at the Close of the Middle Ages*, 17 vol., 1896–1925), by far the most exhaustive and all-inclusive examination of Germany from the later 15th century to the outbreak of the Thirty Years' War (1618)—still provides an almost unsurpassed wealth of detailed information about the period; RICARDA OCTAVIA HUCH, *Das Zeitalter der Glaubensspaltung* (1937), a German classic on the era of the Reformation to the end of the Thirty Years' War (1648), which contains much valuable material on social and intellectual life during this period; JOSEPH LORTZ, *Die Reformation in Deutschland*, 5th ed., 2 vol. (1965; Eng. trans., *The Reformation in Germany*, 2 vol., 1968), one of the most thorough studies of the German Reformation during the first half of the 16th century, the first volume focussing on Luther's life and historical impact and the second dealing with the resulting Protestant struggles against a conservative ecclesiastical and political establishment; MANFRED BENSING and SIEGFRIED HOYER, *Der deutsche Bauernkrieg, 1524–1526* (1965), an excellent examination of the German Peasants' War of the early 16th century, should be read with care, as the author's analysis is strictly Marxist, drawing heavily on Friedrich Engels' classic work on the subject; KARL BRANDI, *Kaiser Karl V*, 7th ed. (1964; Eng. trans., *The Emperor Charles V*, 1939, reprinted 1963), probably the best biography of the first man to rule a truly world-wide empire—succinctly documents his immeasurable role in the course of European history; LEOPOLD VON RANKE, *Zur deutschen Geschichte vom Religionsfrieden bis zum dreissigjährigen Krieg . . .* (1868), a classic German treatise on the period from 1555 to the beginning of the Thirty Years' War (1618), which still retains much of its validity today, though it is limited by its focus on political history; C.V. WEDGEWOOD, *The Thirty Years War* (1938, reprinted 1961), the standard work on this conflict that was fought almost exclusively in Germany, emphasis on political history; WILHELM TREUE, *Deutsche Geschichte von 1648 bis 1740* (1956), a short history of Germany for the century following the Thirty Years' War, provides an excellent general account of the developing Austro-Prussian and Franco-German rivalries but lacks detailed analysis due to its brevity; W.H. BRUFORD, *Germany in the Eighteenth Century* (1935, reprinted 1965), a well-rounded portrait of Germany in the century before the French Revolution emphasizing cultural and literary developments.

From c. 1760 to 1871: For a good account of Central Europe during the period of the Enlightenment, see W.H. BRUFORD (*op. cit.*). FRANCIS L. CARSTEN (*op. cit.*), deals with the form of government, while HANS ROSENBERG, *Bureaucracy, Aristocracy, and Autocracy* (1958), describes the nature of authority. The transformation of civic ideals is the theme of FRIEDRICH MEINECKE, *Weltbürgertum und Nationalstaat* (1962; Eng. trans., *Cosmopolitanism and the National State*, 1970). GERHARD RITTER, *Friedrich der Grosse*, 3rd ed. (1954; Eng. trans., 1968), analyzes the achievements of the famous king. The prevailing attitudes in the Holy Roman Empire toward the fall of the old order are examined by G.P. GOOCH, *Germany and the French Revolution* (1920, reprinted 1966). REINHOLD ARIS, *History of Political Thought in Germany from 1789 to 1815* (1936, reprinted 1965); and EUGENE N. ANDERSON, *Nationalism and the Cultural Crisis in Prussia, 1806–1815* (1939, reprinted 1966), depict the intellectual effects of the French hegemony. For political developments, see H.A.L. FISHER, *Studies in Napoleonic Statesmanship: Germany* (1903, reprinted 1968); and GUY S. FORD, *Stein and the Era of Reform in Prussia, 1807–1815* (1922). ENNO KRAEHE, *Metternich's German Policy* (1963), describes the War of Liberation, while HAROLD G. NICOLSON, *The Congress of Vienna* (1946, reprinted 1970), deals with the reconstruction of Europe. For the period of the German Confederation, see the vigorous but opinionated account by HEINRICH G. VON TREITSCHKE, *Deutsche Geschichte im neunzehnten Jahrhundert*, 5 vol. (1890–96; Eng. trans., *History of Germany in the Nineteenth Century*, 7 vol., 1915–19). THEODORE S. HAMEROW, *Restoration, Revolution, Reaction* (1958), stresses the connection between politics and economics. JOHN H. CLAPHAM, *The Economic Development of France and Germany, 1815–1914*, 4th ed. (1936); and WILLIAM O. HENDERSON, *The Zollverein* (1939), examine the transformation of the economy of Central Europe. VEIT VALENTIN, *Geschichte der deutschen Revolution von 1848–49* (1930–31; Eng. trans., *1848: Chapters of German History*, 1940), emphasizes the political events of the Revolution; WERNER E. MOSSE, *The European Powers and the German Question, 1848–71* (1958), looks at the diplomatic aspects. The achievement of national unification is described from the Prussian point of view by HEINRICH VON SYBEL, *Die Begründung des deutschen Reiches durch Wilhelm I*, 7 vol. (1889–94; Eng. trans., *The Founding of the German Empire by William I*, 7 vol., 1890–98, reprinted 1968); while HEINRICH FRIEDJUNG, *Der Kampf um die Vorherrschaft in Deutschland, 1859 bis 1866*, 10th ed. (1916–17; Eng. trans., *The Struggle for Supremacy in Germany, 1859–1866*, 1935, reprinted 1966), looks at it from the Austrian side. EUGENE N. ANDERSON, *The Social and Political Conflict in Prussia, 1858–1864* (1954), deals with the constitutional struggle in the Hohenzollern kingdom. OTTO PFLANZE, *Bismarck and the Development of Germany: The Period of Unification, 1815–1871* (1963); and ERICH EYCK, *Bismarck and the German Empire* (1950), portray the architect of the new Germany. The origins of the Franco-Prussian War are traced by HERMANN ONCKEN, *Die Rheinpolitik Kaiser Napoleons III* (1926; Eng. trans. from vol. 1, *Napoleon III and the Rhine*, 1928, reprinted 1967).

From 1871 to the present: HAJO HOLBORN, *A History of Modern Germany*, vol. 3, *1840–1945* (1969), perhaps the best one-volume work on recent German history, which, while focussing on political developments, provides valuable insights into the cultural and intellectural life that has made modern German history such an enigma; KOPPEL S. PINSON, *Modern Germany*, 2nd ed. (1966), another standard work on German history from the French Revolution to the present, provides a concise yet comprehensive analysis that includes much neglected material on the post-World-War II period; ARTHUR ROSENBERG, *Die Entstehung der Deutschen Republik, 1871–1918* (1928; Eng. trans., *Imperial Germany, The Birth of the German Republic, 1871–1918*, 1928, reprinted 1966), one of the best analyses of the German Second Empire (the author's socialist views are readily apparent); ERICH EYCK, *Bismarck and the German Empire*, 2nd ed. (1960), perhaps the best one-volume study of imperial Germany's greatest statesman and his decisive impact on the European balance of power; S. WILLIAM HALPERIN, *Germany Tried Democracy* (1946, reprinted 1965), still the most satisfactory one-volume history of the Weimar Republic (1918–33), most of whose arguments remain valid today despite the availability of new primary evidence since the book's publication more than 25 years ago; ARTHUR ROSENBERG, *Geschichte der Deutschen Republik* (1935; Eng. trans., *A History of the German Republic*, 1936, reprinted 1965), a valuable companion volume to Halperin, which views the Weimar period from a socialist perspective; WILLIAM L. SHIRER, *The Rise and Fall of the Third Reich* (1960), one of the most detailed one-volume accounts of the Hitler era beginning with the dictator's political career after World War I, but focusses on political and military events; ALAN BULLOCK, *Hitler: A Study in Tyranny*, rev. ed. (1962), an excellent comprehensive biography of Adolf Hitler indispensable for an understanding of the Third Reich; ALFRED GROSSER, *L'Allemagne de notre temps* (1970; Eng. trans., *Germany in Our Time*, 1971), traces Germany's history from the collapse of the Third Reich in 1945 to the present focussing mainly on West Germany—almost exclusively political history; MICHAEL BALFOUR, *West Germany* (1968), places the postwar Federal Republic in historical perspective while concentrating on the conservative Adenauer era during which West Germany developed close ties to the Western Allies, recovered economically, and began its rearmament program; JOHN DORNBERG, *The Other Germany* (1968), one of the few serious studies of the German Democratic Republic available in English but unfortunately concentrates on the post-1961 period, providing only a very limited account of the origin and development of Germany's socialist state.

(H.D.S./C.C.B./T.S.H.)

Germfree Life

Germfree life refers to plants and animals living in the absence of living micro-organisms. Concepts equivalent to germfree, including axenic and sterile, are included in the word gnotobiotic (the *g* is silent), which is derived from a Greek word meaning "known biota." Gnotobiology comprises the study of germfree plants and animals, as well as living things in which specific micro-organisms, added by experimental methods, are known to be present. When one or more known species of micro-organisms are added experimentally to a germfree plant or animal, the host, of course, is no longer germfree; both the host and the introduced species are gnotobiotic, however, since all added species are known to the investigator.

Definition of gnotobiology

Neither germfree nor conventional animals should be thought of as abnormal; each reflects a "normal state." Diseases may occur, for example, in either germfree or conventional animals. In most comparisons, the major differences between the animal types are the micro-organisms associated with conventional animals and the absence of all detectable living micro-organisms in germfree animals. Precise comparisons between germfree and conventional animals cannot be made unless both are isolated from the environment and fed the same sterile diet.

Gnotobiotic research seeks to explore the effects of micro-organisms in aging and in physiological diseases, the specific causative agent in infectious diseases, and the role of bacteria in protozoan and viral infections. Germfree research currently is directed toward studying the reactions of germfree animals after they have been inoculated with specific known micro-organisms. The vigorous activity now in progress in germfree research is reflected in an increasing number of publications on germfree life and gnotobiology, in the thousands of germfree animals raised each year, and in the laboratories that are contributing knowledge to interactions between micro-organisms and their hosts (which range from plants, flies, and silkworms to mice, rats, guinea pigs, rabbits, pigs, dogs, cats, chicks, turkeys, Japanese quail, monkeys, goats, sheep, and horses). Research, especially with invertebrate animals, may advance the knowledge of basic principles concerning host–microbe interactions.

GENERAL BACKGROUND

Germfree plants

Before methods for obtaining germfree animals had been developed, germfree plants were being used to resolve a controversy concerning the fixation of nitrogen into chemical compounds by leguminous plants. In 1858 it was found that nitrogen fixation occurred in plants grown on nonsterile soil but not in those grown on sterile soil. The fixation process was thought to occur in organic corpuscles (or nodes) found on the roots of plants grown in nonsterile soil. These nodes were shown to contain bacteria when the experiment was repeated in 1888.

After he presented a student's paper on germfree peas to the French Academy of Science in 1885, Pasteur commented that animals should not exist without bacteria to help digestion; in opposition, others argued that life would be better in the absence of bacteria. The first attempts to grow germfree animals, however, were not performed until 1895, with guinea pigs at the Hygiene Institute of Berlin; experiments were continued with chicks for more than a decade with no success. The first successful germfree vertebrate experiments (with chicks) were begun around 1912; shortly thereafter, germfree goats were kept alive for two months.

Progress in modern germfree research may be attributed to a few important developments. Methods developed at the University of Notre Dame in the United States have advanced the systematic exploration of germfree life so that it has become a useful tool in biological research. Work directed at the Kungliga Karolinska Institute in Stockholm, Sweden, pioneered the development between 1960 and 1970 of several centres for germfree research in Europe. Notable work also is carried out in France, Belgium, the Netherlands, and England. From beginnings at Kyoto University, work in government, university, and private laboratories in Japan includes germfree research centres at Nagoya Tokyo, Chiba, and Gifu universities. The Tokyo Sericulture Experiment Station is developing a germfree silkworm industry.

IMPORTANCE OF GERMFREE LIFE RESEARCH

Human health. The addition of one or two specific micro-organisms to germfree animals can clarify cause and effect relationships that are important in human disease. Sometimes, germfree animals become infected and die after they have been inoculated with bacteria commonly found in the intestines of conventional animals. Such bacteria include *Escherichia coli* and *Bacillus subtilis*. In contrast, inoculation of pathogenic (*i.e.*, disease-causing) organisms such as *Entamoeba histolytica* (which causes amebic dysentery) into guinea pigs, or *Shigella flexneri* (which causes enteritis) into rats and mice, kills conventional animals and has almost no effect on germfree ones. In some diseases caused by bacteria and viruses (*e.g.*, rabies), specific bacteria normally found in the intestinal tracts of conventional animals collaborate with pathogens to produce the usual clinical symptoms of disease and death; *E. coli* often collaborates in such cases.

Gnotobiotic research involving the simultaneous introduction of more than two micro-organisms may ultimately allow man to control his microflora (*i.e.*, the types of micro-organisms in his body). Experiments in which some germfree animals are inoculated with a pathogen and others are inoculated with both a pathogen and another micro-organism reveal that the apparent strength of the pathogen changes in the presence of different microfloral environments. Such results provide direct evidence for an increasingly popular concept that the microflora (in the alimentary tract of a conventional animal) are important in defending an animal from many potentially pathogenic micro-organisms frequently found in various parts of the alimentary tract.

Tooth decay

The techniques of gnotobiology have been used to clarify the causes of certain common human health problems. Gnotobiotic studies have shown, for example, that tooth decay is caused by certain infectious bacteria. Only a few of many microbial species added to germ-free rats have produced decay. It appears that tooth decay in man is an infectious disease caused by a few bacteria.

Germfree animals are used in toxicology, pollution control, and vaccine tests. The effects of an external force (*e.g.*, radiation or a noxious gas) are easy to distinguish because there is no interference from infection. Sterile surgery is another application of gnotobiotic procedures. Surgical entry into a patient can be made by a single incision through a sterile isolator and the skin which have been glued together. Both patient and surgeon are outside the isolator, while the sterile instruments and the area of the patient that is to be operated on cannot be contaminated by micro-organisms from the field of operation.

Patients with underdeveloped or suppressed immunological defenses against bacteria can be placed in complete biological isolation using gnotobiotic techniques. Babies suspected of lacking the ability to synthesize gammaglobulins (blood proteins that include antibodies) have been delivered into germfree isolators and maintained there until laboratory tests have shown that they could synthesize gammaglobulins. Some hospitals for cancer patients can maintain patients in an environment containing greatly reduced numbers of bacteria, particularly pathogens. Such facilities for humans are comparable to research facilities in which animal colonies are maintained free of specific pathogens. Precautions that precede heart transplants may include elaborate gnotobiotic rooms and procedures to prevent an immune-suppressed patient from coming into contact with pathogenic micro-organisms. Increasing success in antibiotic treatment of biologically isolated patients may culminate in gnotobiotic hospitals.

Biological research. Germfree animals are used to determine the effects of specific micro-organisms that are

added to a germfree host. Knowledge of the characteristics of germfree animals, therefore is important. The effects of the microbial population, which lives in the human alimentary tract and on the skin (every human harbours several hundred billion micro-organisms), can be inferred if experimental animals without micro-organisms are compared with those to which one (or more) species of micro-organisms has been added. The total numbers of each added microbial species in any part of an animal can be counted with standard microbiological techniques. Such qualitative and quantitative estimates permit the assessment of the biological component of the host environment with an exactness equivalent to that regarding physical and chemical components of the environment.

Except in ruminants (*e.g.*, cows), micro-organisms have been taken for granted in the ecology of healthy animals. In nature, however, ruminants cannot live without micro-organisms which digest cellulose and synthesize essential amino acids and vitamins. One challenge to germfree research, therefore, is to determine which of the specific nutrients produced by alimentary tract microflora are needed for the survival of a host. Many germfree species cannot survive if their diets are not fortified with vitamins that are synthesized in conventional animals by their microflora.

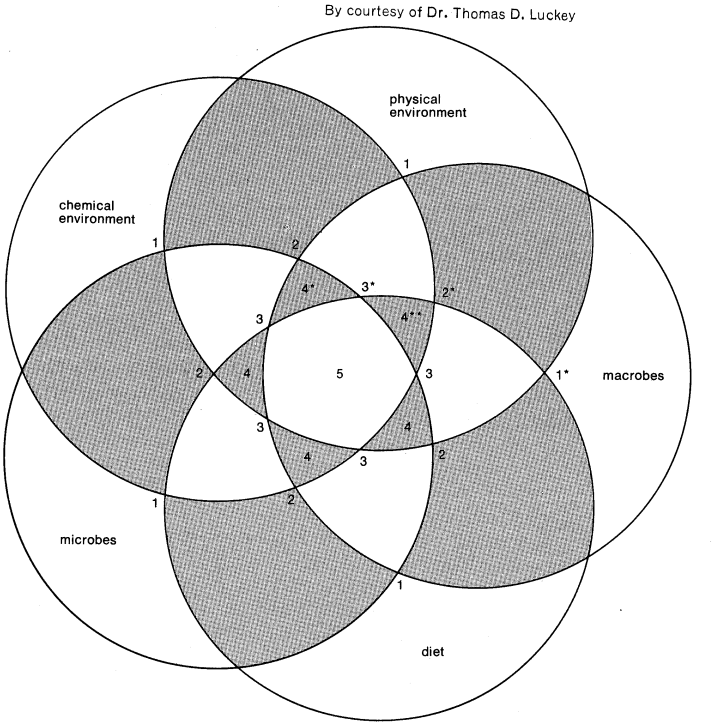
Other subtle interactions between host and microflora are studied with gnotobiotic techniques; for example, a balanced microflora is a defense mechanism that helps to prevent the accumulation of pathogenic micro-organisms within the alimentary tract of a host. Prolonged isolation of an animal may result in changes in its normal intestinal flora; these may lead, in turn, to complications generally termed microbial shock. Isolation of humans for transplantations, for immunological deficiencies at birth, for cancer treatments, or for prolonged space flights raises as yet unanswered questions about the functions, maintenance, and reconstitution of the normal alimentary microflora of humans.

Germfree research provides information about biologically pure animals. Life now is possible in the absence of the ubiquitous micro-organism, which was an important environmental component during vertebrate evolution. Germfree research provides a comparative basis for evaluating conventional life, which is intimately associated with micro-organisms. Aging processes, physiological activities, diseases, and specific nutritive requirements can be studied with gnotobiotic techniques in order to determine important aspects of host-microbe relationships. Experiments performed on germfree animal embryos, for example, provide insight into the development of organs, biochemical systems, physiological mechanisms, and immunological responses.

Outer space research. The germfree animal is a useful tool in the continuing search for life on the moon and other planets and in protecting life on Earth from possible extraterrestrial life forms. Space rockets, sterilized to avoid contaminating other planets with micro-organisms from Earth, can be used to bring materials from other planets back to Earth. Moon dust and moon rock samples, for example, have been examined physically, chemically, and biologically using gnotobiotic procedures. Moon dust has been fed to (and inoculated into) germfree animals to test its toxicity and to determine if it contains any moon micro-organisms that could grow or cause infection.

Industry. Industries exist that produce germfree animals, isolators, and other equipment. Germfree animals provide an ideal source for the culture of epithelial tissue (*e.g.*, skin) because skin removed from conventional animals is always contaminated with micro-organisms. Gnotobiotic techniques also provide stock animals that can be used to start disease-free colonies. After laboratory animals are used to start a new colony, the new colony can be maintained free of specific pathogens (and the animals then are known as specific pathogen-free animals) and are free from all common pathogenic micro-organisms. With animals such as these, better disease control and greater uniformity of animals used in research and drug assays can be maintained. Clean room procedures and isolation equipment also are being developed and improved continuously.

By courtesy of Dr. Thomas D. Luckey



Possible Interactions of major components of ecology (see text).

GENERAL FEATURES OF GERMFREE LIFE

Germfree life is an extension of the pure culture concept in microbiology, in which one species can be isolated from all other species. Manipulation and description of the biological component of the environment of an animal are the objectives of gnotobiology. Gnotobiology provides a way to understand the ecology of interactions between host and micro-organisms.

Interaction of ecological components

Germfree research utilizes scientific methods, in which a complex subject is separated into its component parts and then studied in any desired combination. The component parts of a living system may be treated as units after the host (also called the macrobe) is separated from its normal complement of micro-organisms (also called the microbes). Interaction of major ecological components (*e.g.*, physical, chemical, and biological aspects of the environment) may be studied in a variety of ways (see drawing). A mammalian embryo in the uterus of its mother, for example, is in state 1*; that is, normally, no micro-organisms are present, and the environment of the embryo is the same as that of the mother. State 2* can be represented by a chick embryo. In this case, temperature regulation (a component of the physical environment) originates outside the egg; chemicals, including air and water vapour, must pass through egg membranes before entering the embryo. A newborn mammal is in state 2* until the amnion (the sac that encloses the fetus) is broken; then it passes to state 3*, and both chemical and physical environmental components act directly on it. State 3* usually lasts only a short time, because the first contacts of a newborn animal with the outside world contaminate it with many types of micro-organisms that immediately establish themselves on the skin, in the nose, and in the alimentary tract of the animal. The newborn animal then passes to state 4* and enters state 5 when it eats.

An animal may be born or hatched into a germfree isolator, and since there are no micro-organisms present, the animal will remain in state 3* until it is fed. Germfree piglets kept in this state for one or two days have been used as antigen-free animals in immunochemical studies. When it is fed, such an animal enters state 4**,

which is the usual state of the germfree animal. If a germfree animal is inoculated with one or more pure cultures of micro-organisms, both animal and micro-organisms enter state 5. Micro-organisms, for example, may be studied as they exist in various states, including the effect of the host upon the microbe in state 5.

Research may eventually show that most of the effects now attributed to microflora can be duplicated by the administration of certain compounds; *e.g.*, antigen (a substance that stimulates formation of antibodies) injection is known to cause changes in lymph node size, cell numbers, and antibody production. In this sense the germfree animal becomes an experimental base, or control animal; animals containing micro-organisms are in a more complex experimental condition.

METHODS

Isolation. Physical methods of isolating a desired organism from a group of micro-organisms (*i.e.*, the microbiological techniques developed for pure culture of micro-organisms) usually are not used with vertebrates. Although antibiotics and germicides are used routinely by some investigators to decontaminate animals, most germfree animals are transferred from microbe-free environments in nature into a sterile isolator by aseptic (*i.e.*, sterile) transfer. If, for example, a seed or an egg contains no internal microbial contaminants, germfree plants and egg-laying animals may be obtained from seeds or eggs whose surfaces have been sterilized. Germfree chicks, turkeys, and Japanese quail can be obtained by passing such surface sterilized eggs through a temperature controlled, germicidal trap and into a sterile isolator in which they are incubated and hatched. The eggs must be obtained from flocks free of micro-organisms that can invade the egg before it leaves the oviduct of a hen. The fetus of mammals, normally free of micro-organisms, can be removed from the mother and transferred to a sterile isolator. The young animal must be fed by hand in order to avoid contamination from its mother. When a hand-reared germfree mammal is allowed to reproduce, germfree colonies of that species are obtained. Germfree rats or mice can now be purchased commercially. In this case, the animals are shipped under sterile conditions and then transferred (also under sterile conditions) to an experimental isolator.

Eggs and seeds

Isolator. The isolator, a barrier that prohibits the passage of living micro-organisms, may be small or large, and usually is constructed of glass and steel, or plastic. The physical barrier to contamination may be a flow of sterile air from inside the isolator to the outside; *e.g.*, an air outlet in the isolator. The isolator has a sterile entry for placement of the animal and its food; a source of air with exhaust; and, usually, armlength rubber gloves. Special attachments may include a liquid dip tank, a transfer cage, a bacterial filter for liquids, or a second isolator. Transfers are made aseptically after the entry is sterilized with a dilute peracetic acid fog. The most widely used isolator is a soft plastic sack-like container that is available in many shapes and sizes. A simplified shipping cage for small germfree animals is a large plastic, screw-cap bottle with screening and a glass wool filter at the cap end. Isolators are usually maintained with a slight positive internal pressure. Steel isolators are used if a negative pressure is needed to prevent the escape of pathogenic micro-organisms or viruses inside the isolator. Plastic isolators are designed to surround the equipment needed for an experiment; for a given experiment, a standard commercial isolator can be modified or a special isolator constructed. Convenience and freedom from contamination are important characteristics of isolators.

Sterilization. Sterilization of metal isolators and most utensils is accomplished with steam under pressure. Germicidal vapour sterilization is used for plastic isolators that cannot withstand the heat of steam sterilization. Air filters are usually sterilized with dry heat. Air is sterilized by filtering it through a bed of fine glass wool; electrostatic forces trap particles too small to be stopped by the glass wool. Eggs and seeds are surface sterilized with compounds such as mercuric chloride, peracetic acid, or formalin. Food and water can be sterilized by steam, irradiation, filtration, or chemical agents.

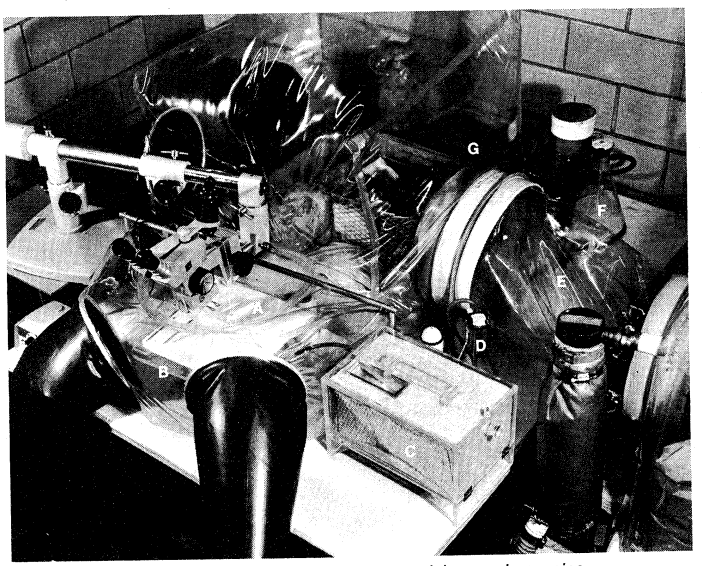
Nutrition. Lack of knowledge about the exact nutritional requirements of specific animals has been the major obstacle to the successful development of germfree animal colonies. Special nourishing formulas, based upon analyses of colostrum and milk, have been used to hand feed rats, mice, guinea pigs, and rabbits removed surgically from their mothers. Once a germfree species has been reared and has reproduced, new strains can be raised by using foster germfree mothers with milk to suckle young animals removed from their mothers under sterile conditions. Germfree hamsters, however, cannot yet be reared either by hand feeding or by cross suckling to another species; apparently they require an unknown nutrient or environmental factor, which has not yet been discovered.

Cross suckling

The diets of germfree animals are generally comparable to those of conventional animals. Extra vitamins are added to the diets of germfree animals to compensate for losses during sterilization procedures and for vitamins originally supplied to the animal by microbial synthesis in the alimentary tract. Fibrous materials are added to diets for germfree rodents and rabbits to help reduce distention of the cecum, a pouch in the large intestine, in these species. Antigen-free diets fed to germfree animals provide a base for immunological studies; such chemically defined diets, which have maintained germfree mice through three generations, provide confirmatory evidence that all the nutrients required by mice are known.

Determination of absence of microbes. Microbiological tests for the presence of living micro-organisms in the germfree animal and its isolator are essential in germfree research. The criterion of germfreeness (*i.e.*, lack of contaminating living micro-organisms) is established by direct examination and exhaustive laboratory tests. In addition, two weeks of data from an organism may be required to establish that a plant or an animal is germfree. Dead micro-organisms or microbial products are permissible in a germfree experiment. Standard microbiological methods are used to detect viruses, rickettsiae, bacteria, fungi, protozoans, worms, and arthropods that could live in intimate association with a germfree host. The experimental base of the germfree concept varies according to the adequacy of methods used to detect viable (*i.e.*, capable of becoming active) organisms intimately associated with the host; for example, the virus-

By courtesy of Patricia M. Bealmear, University of Notre Dame, South Bend, Indiana



Aseptic surgical unit for thymectomizing newborn mice. (A) Operating board. (B) Plexiglas stand. (C) Plexiglas insulated cooling chamber. (D) Suction flask. (E) Transfer port. (F) Safety trap for pump. (G) Vacuum pump.

free status of germfree animals can be questioned since radiation is known to activate viruses in all germfree mice and in some strains of germfree rats.

Inoculation of germ-free animals

Oral inoculation of germfree animals with one or more pure cultures of micro-organisms is used to study host-microbe interrelationships. Although the gnotobiotic animals are no longer germ free when this is done, the methods used to maintain them are comparable to those used for germfree animals. A drop of a uniform suspension of a pure culture of micro-organisms (called an inoculum) usually is placed in the mouth of each germfree animal in an isolator. The drop usually contains between one thousand and one million micro-organisms. If these micro-organisms become established in the alimentary tract (or some other area) of the inoculated animal, it is called a "gnotophoric" animal; *i.e.*, it carries one or more known species of microbes and no other species can be found by microbiological tests.

GENERAL CHARACTERISTICS OF GERMFREE ANIMALS

Appearance. Germfree chickens, turkeys, Japanese quail, guppies, flies, dogs, cats, calves, burros, sheep, pigs, goats, and monkeys are normal in appearance. On the other hand, germfree rodents and rabbits often appear to have an overfull abdomen; this is caused by cecal distension. When reared under conditions in which they remain clean, most germfree animals can be differentiated from conventional animals only by the lack of a characteristic odour in the former.

Classes of organs and tissues

Macroscopic characteristics. Laboratory comparison of gross morphological characteristics of germfree and healthy, non-infected conventional animals reveals five classes of organs and tissues, depending on the type of contact the organ or tissue has with the microflora or its products. First, organs and tissues remote from any physical or chemical activity of external microbes show no gross differences under germfree and conventional conditions. Brain, bone, cartilage, skeletal muscle, most organs, eyes, tongue, nose, lungs, and peripheral vascular system (alimentary tract excluded) have similar gross morphology under germfree and conventional conditions. Organs that act together with defense mechanisms comprise a second group. Liver and heart function, the intestinal vascular system, and perhaps the adrenal glands may be less prominent in germfree animals. The liver detoxifies compounds absorbed from the intestine. Since less blood flows to the liver and there is less inflammation of the intestine in germfree animals, cardiac output also decreases. Tissues that are targets for invasion by micro-organisms, as well as tissues that maintain function and integrity for reasons other than microbial defense are different in subtle ways in germfree animals. Skin, teeth, and mucosa of the respiratory tract, sinuses, pharynx, and gastrointestinal tract, for example, are more complex when they exist in animals containing micro-organisms than they are in germfree animals. These tissues have less blood circulation in germfree animals than in conventional animals. The intestinal wall of a germfree animal is somewhat lighter in weight and colour than that of a conventional animal. The mouth, urogenital tract, and eyelids have yet to be studied extensively. Organs and tissues of the specialized defense systems of the host comprise the fourth class. The lymphatic system (lymph nodes, thymus, spleen, white blood cells) generally is underdeveloped in the germfree animal that has not been stimulated by antigens to form antibodies. In general, the defense systems are very poorly developed in germfree animals. Special features that result from the state of germfreeness, (*i.e.*, the cecal wall of rodents) is the last class of morphological distinctions between germfree and conventional animals.

Although cecal distension is not common in most germfree species, the cecum may constitute up to 25 percent of the body weight of germfree rodents. This physically inert mass causes the animal to be lethargic, reluctant to breed, and susceptible to premature death. Since the weight of the cecal wall does not change, cecal distension is caused primarily by a flabby intestinal wall, in which muscle tone is decreased. Although the addition of fibrous materials to the diet reduces cecal distension in rats and mice, germfree guinea pigs and rabbits are not helped much by the dietary supplement. The compounds that cause the decrease in muscle tone may be decomposed by certain intestinal bacteria; therefore, such compounds, present in greater concentration in germfree animals than in conventional ones, may cause the condition.

Microscopic characteristics. Microscopic examination of the tissues of a germfree animal, particularly the lower intestine and lungs, reveals a lack of inflammation; the clarity of germfree tissues contrasts with the mildly inflammatory digestive and respiratory tissues of conventional animals. Germfree tissues, therefore, are particularly valuable for studying the first lesions produced by a respiratory or an alimentary toxin. Germfree guinea pigs, for example, have been valuable in studying lung lesions caused by air pollution. Microscopic examination of the small intestine shows fewer damaged absorptive surfaces in germfree animals; in addition, the villi (projections that line the small intestine) are more free and uniform, and there is less fibrous tissue in the intestinal wall. Peyer's patches, small lymph nodes associated with the intestinal wall, are reduced in numbers and size in germfree animals.

Functional characteristics. *Defense mechanisms.* A variety of immunological studies can be carried out on germfree animals for several reasons: the lymphatic system is poorly developed, and there is negligible antibody production. The blood of germfree animals has only one third as many white cells as that of conventional animals and few gammaglobulins (blood proteins that include antibodies). No gammaglobulins are present in germfree pigs until they are given antigen to stimulate antibody production. Plasma cells and other defense centres, absent in lymphatic and thymic tissues of germfree guinea pigs, are present in decreased numbers in other species. Since the defense mechanisms can be activated, however, it is known that they are not defective; rather, they act only when they are needed.

Nutrition and growth. Growth rates of most species of germfree animals are equal to, or faster than, those of conventional animals. Nutritional studies have shown that germfree animals use food more efficiently than do conventional animals. Apparently the germfree animal, which uses no energy to defend itself from invasion by undesirable micro-organisms, has more energy available for growth. Intestinal micro-organisms produce some vitamins that are utilized by the conventional animal. Since they lack such microbial synthesis, germfree rats require folic acid, biotin, and vitamin K in their diets. On the other hand, germfree rats may require smaller quantities of amino acids than do conventional rats, suggesting that the intestinal microflora may compete with the host for certain nutrients. Iron salts are poorly absorbed by the intestines of germfree rodents; calcium is more readily absorbed. Germfree mice and rats have more calculi ("stones") in the kidney and bladder than do conventional mice. Dietary antibiotics stimulate growth to a lesser extent in germfree chicks and turkeys than they do in conventional animals; most important in this case is the indirect action of the antibiotics on the intestinal micro-organisms. Germfree animals have more digestive enzymes in the lower digestive tract and excreta than do conventional animals; apparently intestinal bacteria degrade these products in conventional animals.

Reproduction and tumour growth. Germfree mice, rats, rabbits, guinea pigs, chicks, Japanese quail and dogs have reproduced. The average number of young mice at birth and at weaning is high in germfree colonies. With the exception of infections that end in death, both germfree and conventional animals have similar symptoms and pathology in several stress reactions (*e.g.*, radiation sickness, hemorrhagic shock, anal block). Cancer follows a similar route when chemical carcinogens are applied to germfree and conventional animals; tumours sometimes grow faster in germfree animals. Spontaneous tumours are rare in germfree animals; transmissible tumours

(*e.g.*, Bittner mammary tumour) do not occur spontaneously in germfree mice. Temperature, respiratory exchange, and cardiac output may be lower in germfree rats than in conventional ones. Although longevity has not yet been studied systematically, the lifespan of germfree mice appears to be about 25 percent longer than that of conventional mice. Male mice have been shown to survive longer than females.

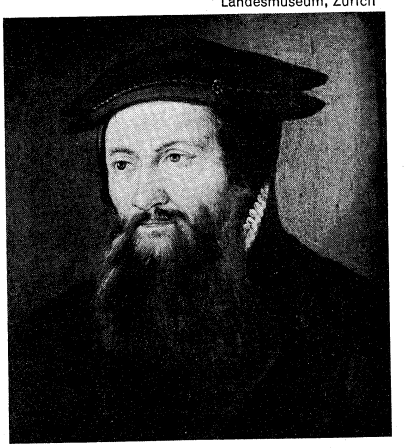
BIBLIOGRAPHY. M.E. COATES *et al.* (eds.), *The Germ-Free Animal in Research* (1968), useful for the novice in gnotobiology, the chapter on nutrition is especially good; T.D. LUCKEY, *Germfree Life and Gnotobiology* (1963), a thorough coverage of broad concepts, history, techniques, animal characteristics, and selected experiments in the field, including an extensive bibliography and chronology of important events in gnotobiological research, "Effects of Microbes on Germfree Animals," in W.W. UMBREIT (ed.), *Advances in Applied Microbiology*, vol. 7, pp. 169–223 (1965), a survey of the effect of inoculating microbes into germfree animals using a microbiological classification, and "Gnotobiology and Aerospace Systems," in M. MIYAKAWA and T.D. LUCKEY (eds.), *Advances in Germfree Research and Gnotobiology*, pp. 317–352 (1968), a table and discussion of the effect of microbes when added to animals using an animal classification; E.A. MIRAND and N. BACK (eds.), *Germfree Biology* (1969), a good presentation of immunology, patient care in isolation, and other topics approached from the viewpoint of germfree research; M. MIYAKAWA and B.S. WOSTMANN (eds.), *Technology in Germfree and Gnotobiotic Life Research* (1969), a book presenting technical details useful for beginning workers; A. and V.B. SILVERSTEIN, *Germfree Life* (1970), a simplified introduction to the topic, well illustrated and easily read by the nonscientists; B.A. TEAH, *Bibliography of Germfree Research, 1885–1963* (1964), good coverage of scientific papers are provided in this and in annual supplements to the 1885–1963 issue.

(T.D.L.)

Gesner, Conrad

A Swiss physician and naturalist, honoured as one of the founders of modern zoology and of bibliography and as a pioneer in both botany and mountaineering, Gesner is best known for his bibliography *Bibliotheca universalis*, which provides a checklist of European authors and their works in Latin, Greek, and Hebrew up to the Renaissance; for his *Historiae animalium*, which summarizes European knowledge and belief in his time of the Earth's animal life; and for his collection of notes and wood engravings on plant life.

By courtesy of the Schweizerisches Landesmuseum, Zurich



Gesner, portrait by an unknown artist, second half of the 16th century. In the Schweizerisches Landesmuseum, Zürich.

Education

Gesner was born in Zürich on March 26, 1516. Noting his learning ability at an early age, his father, an impecunious furrier, placed him for schooling in the household of a great-uncle, who augmented his income by growing and collecting medicinal herbs. There young Conrad acquired a basic knowledge of plants and their medicinal uses that led to a lifelong interest in natural history.

Gesner's progress at school, largely concerned with gaining facility in reading the classic works of Latin and Greek authors, so impressed his teachers that a number of them sponsored his continued education. One acted as his foster father after his own father had been killed in 1531 during one of the many religious conflicts of the times; another fed and sheltered him for three years; and a third saw him through upper school at Strassburg. Together they promoted a scholarship for him to study at Bourges and Paris. Even when Gesner committed what his sponsors considered the fatal mistake at the age of 19 of marrying a young lady who had no dowry, his sponsors did not forsake him but rather found a teaching position for him in Zürich and then managed to persuade the authorities to grant him a leave of absence with pay so that he could undertake formal study of medicine in the city of Basel.

The first fruits of such faith was a Greek–Latin dictionary Gesner published in 1537, having prepared it in his spare time at Basel. At the age of 21, he was appointed professor of Greek at the Lausanne Academy. Three years of teaching brought him enough money for another year of studying medicine, and in 1541 he received his doctoral degree. Gesner spent the rest of his life practicing medicine in Zürich, serving also as a lecturer in Aristotelian physics at the Collegium Carolinum and, after 1554, as city physician.

During the years that followed, he continued to read prodigiously while at the same time, despite his many professional duties and recurring poor health, making field trips, starting a museum, organizing medical instruction, and publishing the 70 or so books that he either wrote or edited.

In an early work, a medical tract on the virtues of milk, *Libellus de lacte et operibus lactariis* (1545), he included a letter to a friend in which he extolled mountains as one of the greatest wonders of nature. This reference and a later account of his scaling of Mt. Pilatus (1555) provide one of the first records of mountain climbing.

In 1545 Gesner published his *Bibliotheca universalis*, the first bibliography of its kind, listing about 1,800 authors alphabetically with the titles of their works, annotations, evaluations, and comments on the nature and merit of each entry.

This monumental reference was followed in 1548 by the encyclopaedic work *Pandectarum sive Partitionum universalium Conradi Gesneri . . . libri xxi*, in which Gesner attempted to survey the recorded knowledge of the world under 21 headings. The first 19 books were published in 1548; the last, devoted to theological thought, was published in 1549, while the 20th, on medicine, was never completed.

Contributions to zoology and botany

Gesner's next monumental achievement was a compendium of recorded knowledge concerning animal life, the *Historiae animalium*, in which he sought to distinguish observed facts from myths and popular errors. The first volume (1551), a generously illustrated work of 1,100 folio pages, dealt with viviparous quadrupeds (four-footed animals that bear living young). Later volumes on oviparous quadrupeds (those that hatch the young from eggs), birds, and fishes and other aquatic animals followed in 1554, 1555, and 1556; the partially completed fifth volume, on serpents, was published posthumously in 1587.

Gesner never completed a similarly comprehensive survey of plant life, but his notes and about 1,500 wood engravings including the important flowers and seeds were used by other authors for two centuries after his death. In his own lifetime, he was best known for his botanical work.

Gesner also published *Mithridates: De differentis linguis* (1555), an account of about 130 then-known languages, and an edition (1556) of the works of the 3rd-century Roman miscellaneous writer Claudius Aelian.

Gesner died on December 13, 1565, from plague contracted during an epidemic in Zürich. By the standards of his time Gesner as a scientist showed good judgment and industry. His use of woodcuts was significant in fixing the accuracy of his data and made possible the eventual emergence of a scientific zoology and botany. His writ-

ings about his mountain excursions further helped to emphasize the importance of the empirical study of nature. Few in his time did so much to comprehend and extend the range of man's knowledge of his natural surroundings.

BIBLIOGRAPHY. Major biographies are: WILLY LEY, *Konrad Gesner* (1929), a summary in German of his life and contributions in zoology, botany, paleontology, and medicine, with a catalog of his publications and a bibliography of source materials; and *Dawn of Zoology* (1968), the most extensive discussion in English on Gesner's place in the history of zoology, with reproductions of his wood engravings of animals and birds; and JOHANNES HANHART, *Conrad Gessner* (1824), the first definitive and still valuable biography (in German), based on materials in the Bibliotheca Carolina and the Collectio Simmleriana in Zürich, the Kirchen-archive von Zürich and Basel, and the Collectio Vadiana in St. Gallen. See also JOSIAS SIMLER, *Vita Conradi Gesneri* (1566), a biography in Latin by the executor of his scientific papers, and a translation of this work into German: DAVID RICHTERN, *Des Welt-berühmten Medici . . .* (1711). Particular aspects of this work are discussed in the following: DIETHELM FRETZ, *Konrad Gessner als Gartner* (1948), based on Zürich records, details his early life in the home of his great-uncle, who was interested in herbal medicine, with a survey of his contributions to botany; J. CHRISTIAN BAY, *Conrad Gesner (1516–65) the Father of Bibliography: An Appreciation* (1916), discusses how Gesner's *Bibliotheca universalis* brought to public attention the existing sources of knowledge, thereby stimulating learning in the Renaissance; SIR WILLIAM JARDINE (ed.), "Memoir of Gesner," *The Naturalist's Library*, vol. 20, pp. 1–58 (1866), cites his contributions to natural history; and J.M. THORINGTON, *On Conrad Gesner* (1937), assesses his contributions to mountaineering, with an English translation of his 1543 letter on mountain climbing, and his account of scaling Mt. Pilatus in the Swiss Alps in 1555.

(G.A.P.)

Geysers and Fumaroles

Geysers and fumaroles are both manifestations of hot springs, which disperse groundwater from the upper parts of the Earth's crust after it has been heated by magma (molten silicate material) and its vapours from below. The thermal energy and part of the material in the hot springs are derived from volcanic activity. As magmas begin to solidify to form crystalline rocks, their gases become concentrated under ever-increasing pressure in the residual, uncrystallized liquid. In the last stages of crystallization, this liquid consists largely of hot water and contains gases and many soluble substances in solution. When the pressure becomes sufficiently high, the liquid is forced into cracks in the surrounding solid rock. If a crack extends upward and opens at the surface, then a fumarole, or a steam vent, is formed. If no surface opening exists, the liquid percolates upward, mixes with groundwater, and eventually emerges as a hot spring, possibly showing geyser action. The distinctions among geysers, fumaroles, and kindred phenomena are as follows:

Geysers (Icelandic *geysir*, from *geysa*, "to rush forth," or *gjósa*, "to gush") are spouting hot springs that throw forth intermittent jets of water and steam. Many hot springs that are not geysers may exhibit periodicity: some show periods of agitation or of violent boiling, and some alternately discharge water and gas. A sharp distinction between geysers and other hot springs cannot always be drawn.

Definitions and interrelationships

Fumaroles are steam vents (Latin *fumus*, "smoke"); the "smoke" is water vapour, the dominant constituent, but acid gases, such as carbon dioxide and hydrogen sulfide, characteristically occur.

Solfataras (Italian *zolfo*, "sulfur") are fumaroles that yield hot vapour and sulfurous gases. The temperature of fumaroles varies widely and in any one fumarole is subject to great variations over time. In the famous Valley of Ten Thousand Smokes, which was formed by the tremendous eruption of the volcano Katmai, Alaska, in 1912, temperatures up to 645° C (1,193° F) were measured. In the dying volcano La Solfatara (or Forum Vulcani) at Pozzuoli, Italy, known from antiquity, the temperature of the steam oscillates between 100° and 165° C (212° and 329° F).

With this group are included the paint pots, porridge pots (Figure 1), and mud volcanoes, all of which are hot springs of limited water supply and acid reaction. Their immediately adjacent surface rocks are undergoing active chemical attack from magmatic steam and acid gases to provide the porridge, whereas the iron content of the rocks provides the paint.

By courtesy of T.F.W. Barth



Figure 1: Porridge pots in the Solfatara Crater, Pozzuoli, Italy.

Mofettes (German *Muff*, "musty smell") are vents, issuing carbon dioxide, some methane and other hydrocarbons, nitrogen, and oxygen at low temperature; they mark the last stage of volcanic activity.

The intimate connection between fumaroles and simple hot springs is obvious in areas in which there exists a strong contrast between dry and wet seasons. During the dry season, hot springs are transformed into fumaroles, which in turn, during the wet seasons, become hot springs again. Growler Spring, in Yellowstone National Park, is a good example: its temperature is at a boiling point appropriate to its elevation (93° C or 199° F), but after long-continued dry weather it becomes a fumarole with a temperature of 103° C (217° F). Such seasonal variations indicate that hot-spring waters are derived from groundwater that is heated by magmatic gases and that hot springs are essentially "drowned fumaroles." Recent studies of water chemistry support this view and indicate that all thermal waters are essentially of meteoric (atmospheric) origin. This includes high-temperature fumarolic steam that contains such magmatic matter as sulfates, borates, and arsenates that are not present in the rock through which the water has passed.

Treated in this article are the physical and chemical characteristics of hot-spring waters and their deposits and the nature of geysers and geyser activity. For additional information on magmas and the water in the ground that is heated, see VOLCANOES; IGNEOUS ROCKS; and GROUNDWATER. See also SPRINGS AND WELLS and, for information on the transfer of heat within the earth, EARTH, HEAT FLOW IN.

PHYSICAL AND CHEMICAL CHARACTERISTICS OF HOT SPRINGS

Hot springs are fed by ordinary groundwater heated by magmatic steam. The development of various types of

hot springs, such as fumaroles, acid or alkaline springs, geysers, and tepid pools, is therefore dependent on the groundwater conditions at the locality. Such factors as topography, porosity and other rock properties, and local meteorologic conditions, particularly rainfall, are all important. The development of hot springs requires a source of local volcanic heat.

Water composition. The waters of hot springs fall roughly into three groups: (1) waters carrying calcium carbonate in solution, (2) siliceous acid waters, and (3) siliceous alkaline waters. The acid springs are of low volume and contain hydrochloric or sulfuric acid.

Calcium carbonate (or bicarbonate) lends the water special properties that are easily observed, particularly the ability to precipitate large quantities of travertine, or hard tufa (sedimentary rocks formed by precipitation from springwaters), consisting predominantly of the mineral calcite but also of aragonite, on the surface. Huge terraces of exquisite beauty may develop from this process. In Yellowstone National Park the Mammoth Hot Springs (Figure 2) provide an outstanding example. A unique scenic aspect is the deposit form of cones and basins; the high terraces were deposited along a steep slope by small rushing streams of springwater and now resemble the successive billows of a frozen cataract. In order for this type of springwater to develop, the source water must pass through limestone or other calcite-bearing rocks. In volcanic regions that are underlain by limestone or dolomite, as in Italy and Greece, extensive deposits of travertine are common.

By courtesy of the National Park Service; photograph, M. Woodridge Williams

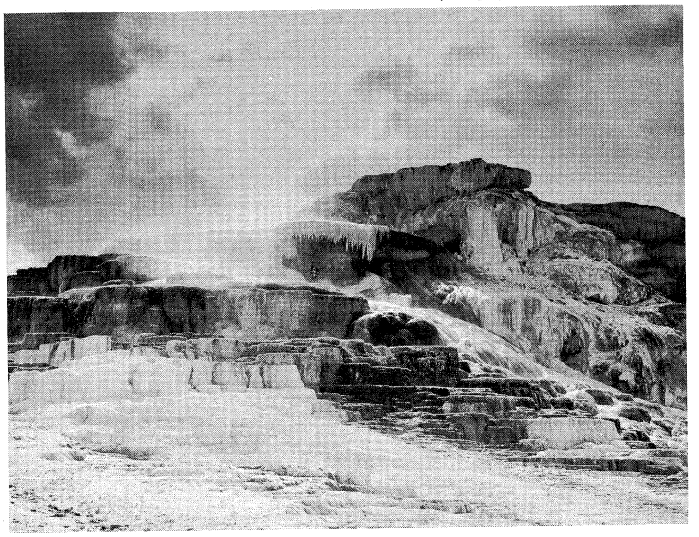


Figure 2: Mammoth Hot Springs terraces, Yellowstone National Park, Wyoming.

Siliceous water types and their evolution

Siliceous types of springwater form an evolutionary series from acid to alkaline and throw much light on the chemical reactions taking place underground. There is no doubt that steam prevails among the gases given off by a magma. But, in addition, acid gases, such as carbon dioxide and hydrogen sulfide, characteristically occur. If oxygen is available, then sulfur dioxide forms, and the waters carrying the emanations will be highly acid. On the pH scale, a measure of the hydrogen-ion content, which ranges from 1 (extremely acid) to 14 (extremely alkaline), a pH of 7 is neutral, but these waters typically exhibit a pH of 4.

In the early evolutionary stages, hydrogen sulfide in acid solution will effectively precipitate silver, mercury, lead, bismuth, copper, cadmium, arsenic, antimony, and tin supplied by the magma. The acid solutions attack the rock minerals and mix with large quantities of groundwater; thus, they become neutral or alkaline, the pH value increasing from 4 to perhaps 10. The resulting alkaline waters then remove silica in colloidal solution (a very fine grained suspension). Thus, the general relation produced is one of increasing pH values with increasing distance from the mother magma to the place of eruption of a particular spring.

Many elements are precipitated as the solutions become alkaline. Iron and aluminum are the commonest of these elements; they are present only in springwater more acid than that with a pH of about 5. Calcium and magnesium are present in waters with a pH of about 7. Sodium and potassium are not affected by acidity, but sodium is always present in great excess over potassium. Except where there is a surplus of calcium carbonate, the issuing waters thus contain silica in great relative excess over all other rock constituents except sodium. The relative contents of the spring gases also evolve. Hydrogen and hydrogen sulfide in acid waters gradually diminish; they are still less abundant in neutral waters and are completely absent in alkaline waters.

Incrustations and deposits. High-temperature deposits from vapours around acid fumaroles and hot-gas vents are dominantly sal ammoniac. Sulfate-bearing incrustations and efflorescences (mainly of ammonia, magnesium, iron, and aluminum) occur in deposits of somewhat lower temperature, and large amounts of native sulfur and gypsum are deposited around acid hot springs at the boiling temperature. Minor amounts of other sulfates and of pyrite, limonite, and opal also occur.

Siliceous sinter deposits

Siliceous sinter (a variety of opal) is the deposit typically formed by alkaline hot springs. In some areas it comprises extensive deposits of domes, terraces, or flats. The most distinguished example was the great "White Terrace" beside Lake Rotomahana, in New Zealand, which was destroyed in a volcanic eruption in 1886. Alkaline springs are almost always deep and are often distinguished by the brilliant, bluish colouring of the water, which contrasts beautifully with their white sinter deposits. The sinter is white or grayish, scaly to massive, and sometimes porous, filamentous, or cauliflower-like. Chemical analyses show 80–90 percent silica and about 10 percent water; the remainder is mostly aluminum, iron, and small amounts of magnesium, calcium, and sodium.

Rather unusual and related to travertine deposits are small globules of mercury and cinnabar (a mercury sulfide mineral) deposited as coatings on calcareous tufa of some boiling springs in Idaho and California.

Another special occurrence is a salt crust on Lake Magadi, Kenya, which is a natural evaporation pan, fed by four copious hot springs (aggregate flow is about 750 gallons per second). This crust yields 80,000 tons of soda ash and 40,000 tons of salt annually, but the inflow of sodium salts far exceeds the rate of exploitation. Mineral-spring brines are worked at many places in the world.

Hot springs and fumaroles in Larderello, Italy, also have been the subject of considerable attention. Boric acid, ammonium sulfate, and carbon dioxide are recovered.

Discharge and thermal energy. Large amounts of hot groundwater emerge in alkaline hot-spring areas, whereas acid hot springs, solfataras, and fumaroles characteristically occur in drier areas and have a much smaller discharge. Indeed, observations indicate that acid waters make up 1 or 2 percent of the total hot-water discharge in Iceland and in Yellowstone National Park, although the acid-hot-spring areas are more extensive than those of the alkaline springs.

Because of the great porosity of rocks in Iceland, much hot water runs off as overland flow and subsurface discharge to be lost to the sea (Figure 3). Drilling has raised the yield considerably; the surface discharge in low-temperature areas of Iceland amounts to 300 gallons per second, but the additional discharge from drill holes is 400 gallons per second. The surface discharge in Yellowstone National Park is 800 gallons per second, and the total discharge in all of Japan is 4,400 gallons per second.

A conspicuous example of the copious water supply of alkaline springs can be seen at Deildartunguhver, 30 miles north of Reykjavík, Iceland. Cascades of boiling water belch forth from several rents and fissures in a steep wall of hardened boulder clay of red, blue, and yellow colours; they meet to form a steaming river, the

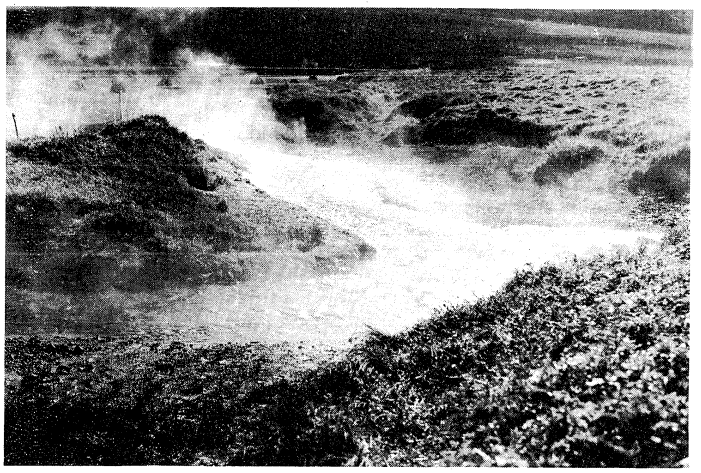


Figure 3: Deildartunguhver River, north of Reykjavík, Iceland.
By courtesy of T.F.W. Barth

largest "boiling river" in the world, yielding approximately 65 gallons per second. The temperature at the place of issue is approximately 99° C (210° F); at the confluence with Reykjadalsá, about 130 yards downstream, the temperature is 82° C (180° F).

Utilization of thermal energy

Although a tremendous amount of heat is released at the surface by the activity of hot springs, the utilization of this energy has been relatively slight.

Thermal energy is used only for home heating, greenhouses, and the drying of such substances as seaweed and diatomaceous earth. Leaky underground chambers and other technical difficulties have restricted the use of the mechanical energy (pressure) of fumaroles and geysers. Only at Larderello, southwest of Florence, have larger energy projects been developed. In 1904 the first generating station operated on natural pressurized steam was installed. Subsequently, a great number of additional borings were made, and the energy output increased from 3,000 kilowatt-hours in 1916 to 2,694,000,000 kilowatt-hours in 1968. This was approximately 2.6 percent of the total energy consumption of Italy in 1968.

Some electricity is generated from wells in the geyser region of Sonoma County, California, and the utilization of hot springs also has been accomplished in New Zealand.

GEYSERS AND GEYSER ACTIVITY

There are three major areas of geyser action in the world: Yellowstone National Park contains at least 200 geysers, or about 10 percent of the total number of hot springs in the park; Iceland has about 30 active geysers, or less than 1 percent of the total number of hot springs; and New Zealand has fewer geysers than Iceland but is famous for having had the greatest one in the world, the Waimangu Geyser, which played daily to heights above 300 metres (1,000 feet) between 1902 and 1905.

True geyser action, one of the rarest phenomena in nature, is one of the most beautiful and conspicuous to behold. Geysir, in Iceland, is the most famous geyser in the world. It was known to science centuries before Waimangu, in New Zealand, had been heard of or before the performances of Old Faithful, in Yellowstone National Park, had been seen by Western men. The behaviour of Geysir had attracted tourists as well as learned men, and the notes and descriptions of this geyser from the 17th century to the present are legion.

Eruption patterns. The magnitudes of geyser eruption range from insignificant (throwing the water a few inches high) to those of famous fountains that playfully attain heights of many hundred feet. Each geyser has its individual performance pattern; some are regular and rhythmical, whereas some are unpredictable. This is because no two geyser edifices and water paths are alike, and the precise causes of eruption cannot always be followed in detail.

Some geysers have negligible discharge, but the larger geysers disperse an amount of water that is many times greater than the volume of their accessible tubes during eruption. One or more storage basins of considerable size are therefore necessary to account for the volume of the visible eruptions. This fact was ascertained for many geysers of Yellowstone National Park and for the Geysir, in Iceland, which in a single 24-hour period (including numerous eruptions) may eject an amount of water equal to at least 100 days' normal or customary discharge.

Discharge and temperature of geysers

Most geysers erupt close to the boiling point, 100° C (212° F) at sea level, and some also will boil between eruptions; others will cool considerably. In the upper part of a geyser vent, the temperature often increases with depth, whereas in the lower part it is constant or decreases. This is a general pattern for all hot springs. Many geysers exhibit very irregular temperature—depth curves, but one important point should be emphasized: boiling temperatures are attained only in the uppermost parts of a geyser system; at greater depths the temperature recedes from the boiling point for that depth.

In Yellowstone National Park, the changing behaviour of geysers is well-known. Contrary to common belief, most geysers do not erupt with regularity, nor are they constant in the intensity of water jets or the heights attained. Even Old Faithful (Figure 4)—which in all tourist brochures during the 1920s and '30s was said to erupt at intervals of precisely one hour and five minutes —when under continuous observation was actually shown to erupt at intervals varying between 30 and 90 minutes. The highest measured eruption, originating in Beehive Geyser, was 67 metres (219 feet). It is possible that the great Excelsior Geyser, active in the 1890s, may have attained greater heights.

Crater Hills Geyser differs in pattern from most geysers: an eruption lasts for about 30 minutes, followed by a quiet interval of only two minutes. In most geysers the eruption period is shorter than the quiet period. There is no discharge, even during an eruption (all water falls back into the basin), and this is another unique characteristic. The eruption starts when the water level in the basin is low, whereas most geysers erupt only when the basin is full. And, finally, the maximum temperature during eruptions in August 1947 was 73° C (163° F); most geysers erupt when their water temperature is close to the boiling point.

Geyser eruptions have been compared to volcanic eruptions. The dimensions are usually different, but, when a

UPI Compix

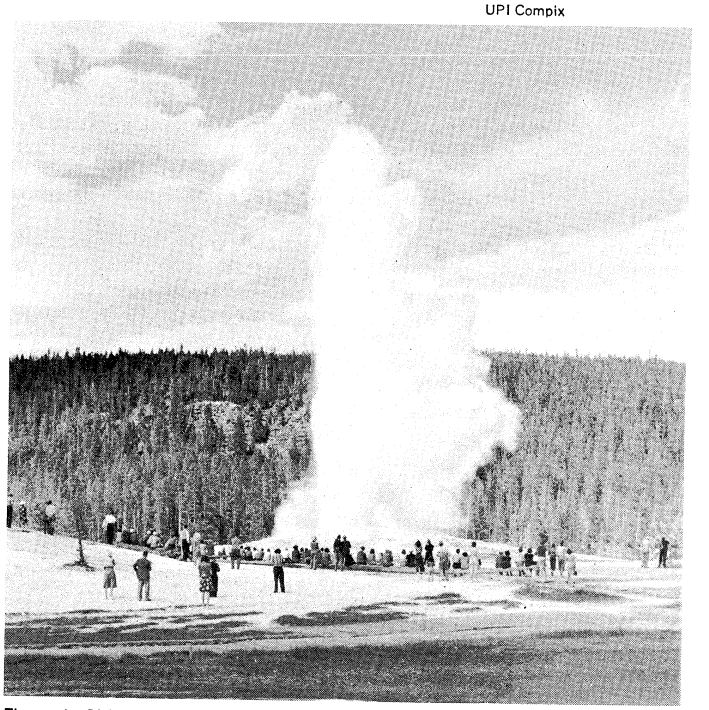


Figure 4: Old Faithful geyser, Yellowstone National Park, Wyoming.

geyser throws steam and water 300 metres (1,000 feet) into the air, it approaches the activity of a small volcano. The peculiar volcano Siretoko-Iō-Zan, in Japan, acts in some ways as if it were a tremendous irregular geyser. After a dormancy of 46 years, it sustained a continuous eruption for several months in 1936, during which it successively developed the following stages: (1) liquid-sulfur eruptions, (2) explosive emissions of hot water and steam, and (3) intermittent gushing of boiling water and steam.

Seawater geysers and *nagawakis*

A geyser spouting seawater, situated at the southwestern tip of Iceland, probably first erupted in 1906 and was said to be an active geyser at flood tide, quiet at ebb. In 1927 it was a true geyser with a regular period of 15 minutes, and the splashes were between three and six metres (ten to 20 feet) high. But in July 1928 the spring underwent a marked change. The period became irregular and, on the average, somewhat longer; after 20 to 25 minutes the water jets became lower. Eventually, the geyser action ceased, and in 1937 the spring became an incessantly boiling pot.

A real geyser, manifestly spouting seawater, is certainly one of the rarest phenomena of the natural world. The former geyser at Atami, Japan, situated 1,000 metres (3,300 feet) from the sea at an altitude of 18 metres (60 feet), is the best example. In many respects it was a peculiar spring. It dispersed diluted seawater, and in 1906 it threw small water jets 0.5 metre (1.5 feet) high, at regular intervals of about five hours. The rhythmic activity was broken by irregularly occurring eruptions, called *nagawaki*, that would go on for 12 hours, discharging large amounts of water and vapour. *Nagawakis* were known to occur at irregular intervals of from one to six months.

Age and history of geysers. Within the three principal geyser regions of the world, clusters of geysers form local areas, and these undoubtedly correspond to situations particularly favourable for geyser action. Although the life of an individual geyser may be ephemeral, geyser areas have persisted for centuries. Hveragerdhi, just east of Reykjavík, Iceland, has been known as a famous geyser area from the earliest historical records (Figure 5),

By courtesy of the Carnegie Institution, Washington, D.C

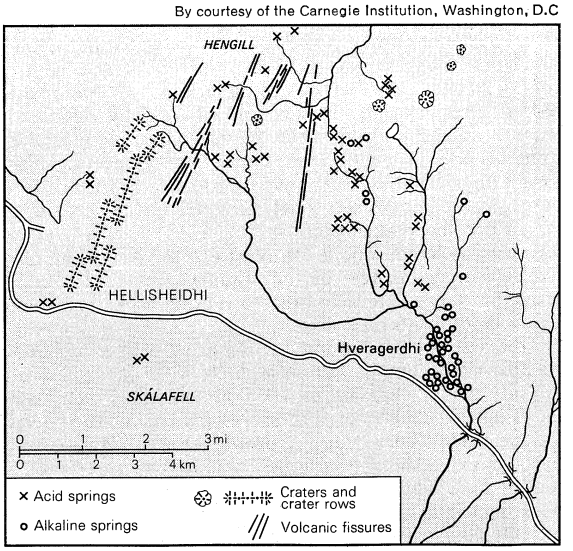


Figure 5: Distribution of hot springs in the Hveragerdhi area, Iceland.

but individual geysers have suffered radical changes; some have ceased spouting and have become ordinary hot springs, and new geysers have replaced the old. When the birth of a geyser was observed here in 1896, it was accompanied by much noise and the rush of qreat quantities of steam, mixed with dirt, clay, and rock fragments, which were shot 180 to 210 metres (600 to 700 feet) into the air. The force of the geyser decreased very soon, and after two days the jets were only ten feet high. After ten days the geyser activity ceased.

The most famous geyser area in Iceland, first mentioned in 1294, is an extensive plain of silica sinter at Haukadal, about 30 miles east of Reykjavík. About 50 individual hot springs occur here, ten of which are active or dead geysers, among them the Geysir. Steam escapes by seepage over the whole area, and the name Geysir was first applied in 1647; all spouting springs are now called geysers for this reason. The probability ratio of a hot spring being a geyser is low elsewhere in Iceland (less than 1 percent) but is much higher (about 25 percent) within the several local geyser areas.

History of the Icelandic Geysir

The Geysir has been known as a spouting geyser for 350 years, but it may be older. Its orifice is probably several thousand years old. It is surrounded by a regular dome-shaped eminence of silica sinter topped by a flat basin or crater 14 metres (45 feet) wide with a central vent extending about vertically downward. During an eruption, columns of water are shot 45 to 60 metres (150 to 200 feet) into the air. Geysir managed to survive many earthquakes and volcanic eruptions that eliminated several of its competitors and brought radical changes to the entire Haukadal area. Toward the end of the last century, however, Geysir became less active and eventually fell dormant until 1935. At that time, a gap was artificially cut in the lip of the basin, forcing the water level down about three feet. Consequently, Geysir was rejuvenated and began to spout as willingly as before. The gap was later filled, and the water resumed its old, high level; in spite of this, Geysir has continued to spout and still spouts today at irregular intervals. These eruptions last for two or three minutes, after which the spring is quiet for about ten minutes.

The renowned geysers Uxahver and Ystihver, in northern Iceland, also were "improved" by slightly lowering the water table in the vents. At Larderello, Italy, fumaroles and hot springs are numerous over an area of about 250 square kilometres (about 100 square miles). Industrial development that has been in progress for almost a century has completely altered the natural surface and has given rise to artificial pools, steam wells, and so-called geysers. It is regrettable that the modern commercial exploitation in many hot-spring areas severely encroaches on natural conditions.

Theories of geyser activity. Many authorities have endeavoured to determine the special features causing geyser action, but none of the solutions set forth has been altogether convincing. The theory designed specifically to account for the activity of the Geysir is best known, although Littli Geysir, Strokkur, and many other Icelandic geysers are not reconcilable with it. The theory dates from 1846 and is based on the idea that a narrow column of water would, when rapidly heated at the bottom, form a continuously boiling spring. Slower heating from below and the presence of a sufficiently wide column to permit considerable cooling at the surface would result in a temperature distribution that is typical of a geyser before eruption. The temperature, therefore, is thought to gradually increase because of the entering of hot steam from below until boiling temperatures are attained at a depth of 12 to 15 metres (40 to 50 feet); the vapour thus produced reduces the pressure in the geyser's vent and an eruption follows (Figure 6).

Most investigators who have given special thought to the subject of mechanisms have come to the conclusion that this theory, even when limited in its application to Geysir, is unworkable. A closed water circuit is assumed for the geyser models, but in most natural geysers (including Geysir) only a negligible fraction of the erupted water is returned to the basin. The termination of the eruption must also be considered. In addition to postulating that an influx of cool water causes termination of the eruption, convectional cooling in the original theory was thought to prevent boiling from beginning in the central vent, and therefore boiling would have to begin in a side chamber. More recent theorists have stressed the fact that underground chambers serve two main purposes in geyser action: some must contain and, during eruption, contribute much water; others must act as steam chambers wherein the expansion and contractions of steam may cause the fluctuation in surface level noted in

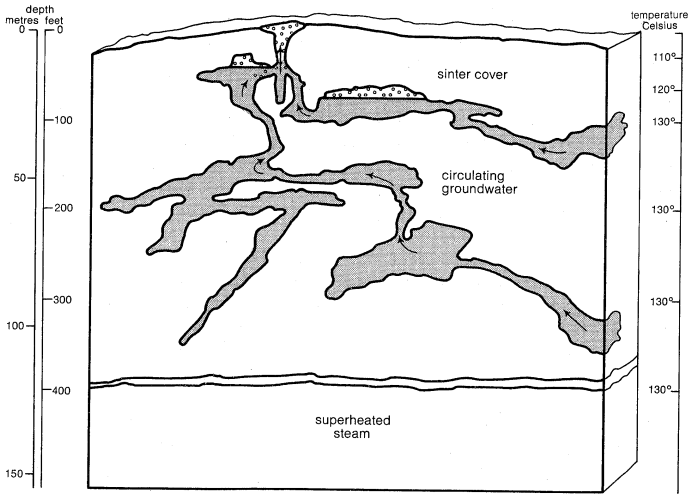


Figure 6: Cross section of geyser substructure.
By courtesy of the Carnegie Institution, Washington, D.C.

geysers. Other workers contend that hot-spring waters must of necessity occlude a certain amount of gas (such as nitrogen, argon, and methane), which is carried up with the water to reach a "critical level," at which small bubbles of gas form simultaneously throughout the liquid (like the bubbles that form when a bottle of soda water is uncorked). This is shown in Figure 6, and the location of this level will determine whether geyser action or perpetual spouting ensues.

BIBLIOGRAPHY. Geysers, fumaroles, and hot springs are fed by magmatic heat and are therefore spatially connected with the active volcanic belts on the continents and on the sea floors. The world distribution of these springs with data on the salient properties and mode of occurrence has been compiled in an impressive publication by G.A. WARING, "Thermal Springs of the United States and Other Countries of the World," *Prof. Pap. U.S. Geol. Surv. 492*, rev. by R.R. BLANKENSHIP and R. BENTALL (1965). A summary report on thermal waters of the world was edited by G. KACURA, *Mineral and Thermal Waters of the World*, 2 vol. (1968), the proceedings of the 2nd Symposium of the 23rd International Geological Congress; and E.T. ALLEN and A.L. DAY published a most illuminating study in *Hot Springs of the Yellowstone National Park* (1935). A similar monograph exists on the springs of Iceland: T.F.W. BARTH, *Volcanic Geology, Hot Springs and Geysers of Iceland* (1950); and an interesting account of the Valley of Ten Thousand Smokes was provided by E.G. ZIES, "The Fumarolic Incrustations in the Valley of Ten Thousand Smokes," *Tech. Pap. Natn. Geog. Soc.*, Katmai series, vol. 1, no. 3 (1924).

(T.F.W.B.)

Ghana

Ghana is a West African republic situated on the coast of the Gulf of Guinea. It has an area of 92,100 square miles (238,539 square kilometres), and, in the early 1970s, a population of 8,500,000. Although relatively small in area and population, Ghana is among the leading countries of Africa, partly because of its considerable natural wealth, and partly because it was the first black African country south of the Sahara to achieve independence from colonial rule.

Ghana, which achieved independence on March 6, 1957, consists of the former British colony of the Gold Coast, and the part of Togoland that was formerly a United Nations Trust Territory under British administration. It is bordered on the west by the Ivory Coast, on the northwest and north by Upper Volta, on the east by Togo, and on the south by the Atlantic Ocean.

The economy is dominated by the cocoa crop, of which the country is the world's largest exporter; other important exports are diamonds, gold, manganese, bauxite, and timber. After a slow start, industrialization in the decade of the 1970s was beginning to loom as a feature in the economy. Ample electricity is available from a large hydroelectric project on the Volta River at Akosombo. The capital is Accra (see city article ACCRA). An associated physical feature may be found under VOLTA RIVER; and historical aspects may be found under WEST AFRICA, HISTORY OF.

PHYSICAL AND NATURAL FEATURES

Relief. Relief throughout Ghana is generally low, with altitudes nowhere exceeding 3,000 feet. The southwestern, northwestern, and extreme northern parts of the country consist of a dissected peneplain—a land surface worn down by erosion to a nearly flat plain, later cut by erosion into hills and valleys or into flat uplands separated by valleys; it is made of Precambrian rocks (from 570,000,000 to 4,600,000,000 years old). Most of the remainder of the country consists of Paleozoic deposits (*i.e.*, from 225,000,000 to 570,000,000 years old), which are thought to rest on older rocks. The Paleozoic sediments are composed mostly of beds of shales (laminated sediments consisting mostly of particles of clay) and sandstones in which strata of limestone occur in places. They occupy a large area, known as the Voltaian Basin, in the north central part of the country where the altitude rarely exceeds 500 feet. Along the north and south, and to some extent along the west, the uplifted edges of the basin give rise to narrow plateaus between 1,000 and 2,000 feet high, bordered by impressive scarps. The most outstanding of these are the Kwahu Scarp in the south and the Gambaga Scarp in the north.

The Voltaian Basin

Surrounding the basin on all of its sides, except in the east, is the dissected Precambrian peneplain, which rises to elevations of 500 to 1,000 feet above sea level and contains several distinct ranges as high as 2,000 feet.

Along the eastern edge of the Voltaian Basin, and extending from the Togo border to the sea immediately west of Accra, is a narrow zone of folded Precambrian rocks running northeast to southwest, forming the scenically attractive Akwapim-Togo Ranges, which vary in height from 1,000 to 3,000 feet. Here are found the highest points in Ghana, including Mt. Afadjato (2,905 feet), Mt. Djebobo (2,873 feet), and Mt. Torogbani (2,862 feet), all of them situated east of the Volta River near the Togo border. These ranges are part of the Togo-Atakora Mountains, which extend northward into Togo and Dahomey.

The southeast corner of the country, between the Akwapim-Togo Ranges and the sea, consists of the gently rolling Accra Plains, which are underlain by some of the oldest Precambrian rocks known—mostly gneisses (coarse-grained rocks in which bands containing granular minerals alternate with bands containing micaceous minerals); in places they rise above the surface to form inselbergs (prominent steep-sided hills left after erosion), which dot the plains. Only in the wide, lagoon-fringed delta of the Volta, about 50 miles east of Accra, and in the extreme southwest of the country, along the Axim coast, are there extensive areas of young rocks less than 136,000,000 years old.

In the east, the predominant rocks are less than 65,000,000 years old, though there is a patch of Cretaceous sediments (from 65,000,000 to 136,000,000 years old) near the Ghana-Togo border. In the western part of the country, near the Ivory Coast frontier, west of Axim, the rocks date to the Cretaceous Period. The intervening coastal zone between eastern and western extremes contains scattered patches of Devonian sediments (from 345,000,000 to 395,000,000 years old). In combination with the older and more resistant rocks of the Precambrian peneplain, these form a low but picturesque coastline of sandy bays and rocky promontories.

The drainage system is dominated by the Volta River Basin, which includes the artificially created Lake Volta. Most of the other rivers, such as the Pra, the Ankobra, the Tano, and a number of smaller ones, flow directly south into the sea from the watershed formed by the Kwahu Plateau, which separates them from the Volta drainage system. South of Kumasi, in the south central part of the country, is Ghana's only true natural lake—Bosumtwi—lying in a deep basin without any outlet to the sea. It is believed to be of volcanic origin, although theories of a possible meteorite origin have also been ad-

The rivers

vanced. Along the coast are numerous lagoons, most of them formed at the mouths of small streams.

Over much of the surface of Ghana, the rocks are weathered, and great spreads of laterite (red, leached, iron-bearing soil) and lesser spreads of bauxite and manganese are found on the flat tops of hills and mountains. Although the movements of the earth's crust that produced the basic geological structure of the country have now virtually ceased, periodic earthquakes are still experienced, especially in the immediate vicinity of Accra along the eastern foot of the Akwapim-Togo Ranges.

Climate, soils, and vegetation. Ghana's climate, like that of the rest of the Guinea coast, is mainly determined by the interplay between the tropical continental air mass, known as the harmattan, which consists of hot, dry, dust-laden air moving from the northeast across the Sahara, and the tropical maritime air mass known as the monsoon, which consists of moist and relatively cool air moving from the southwest across the southern Atlantic. The zone where these two air masses converge is characterized by seasonal line squall rainfall. The convergence zone itself oscillates north and south, following the seasonal movement of the sun overhead; it reaches its most northerly position in the central Sahara, about latitude 21° N, in August, and its most southerly position about 7° N, a few miles north of the Ghana coastline, in January. Rains occur when the dominant air mass is monsoonal (*i.e.*, characterized by rain-bearing winds), while drought prevails when the harmattan dominates.

Rainfall and temperature

In the savanna (grassy parkland) country north of the Kwahu Plateau, the year consists of two seasons—a dry season from November to March, with hot days and cool nights under clear skies, and a wet season that reaches its peak in August and September. The mean annual rainfall ranges between 45 and 50 inches, but there is a marked moisture deficit because of the long and intensely dry season that follows. In the southern forest country, where the annual mean rainfall from north to south has an approximate range of 50 to 86 inches, there are two rainy seasons—one from April to July and a lesser one from September to November—separated by two relatively dry periods that occur during the harmattan season, from December to February, and in August, which is a cool, misty month along the coast. In the Accra Plains, anomalously low annual mean rainfall figures vary between 40 inches to less than 30 inches, and the rainfall variability and the vegetation both bear close resemblance to conditions in the northern savanna zone.

In contrast to the rainfall, temperatures show much more regional uniformity. The annual mean temperature is from 78° to 84° F (26° to 29° C), and the daily range only some 10° to 15° F (6° to 8° C) along the coast, and some 13° to 30° F (7° to 17° C) in the north. Relative humidities range from nearly 100 percent in the south to 65 percent in the north, although during the harmattan season figures as low as 12 percent have been recorded in the north and around Accra. Enervating conditions produced locally by the combination of high temperatures and high humidities are considerably moderated by altitude in the higher parts and by regular land and sea breezes along the coast. In general, the hottest months are February and March, just before the rains, and the lowest temperatures occur in January or—along the coast—in August.

Although soils and biotic factors (*i.e.*, those pertaining to living organisms, including man) are important, vegetation is primarily determined by rainfall. There are three principal types of vegetation from south to north occurring in the coastal savanna, in the forest zone, and in the northern savanna zone.

The coastal savanna in the southeastern plains around Accra consists of a mixture of scrub and tall grass (mostly Guinea grass), with giant anthills, often ten to 14 feet high, forming a prominent feature of the landscape and providing an anchorage for thicket clumps that often include *Elaeophorbia* (a fleshy-leaved plant containing caustic latex) and other drought and fire-resistant species.

In the forest zone (the southern third of the country and the area along the Akwapim-Togo Ranges, where the mean annual rainfall exceeds 45 inches and is well distributed throughout the year without a pronounced dry season), the predominant vegetation is evergreen and tropical semideciduous forest. Here are tall trees of varying heights, forming a closed canopy at the top, above which tower a few forest giants, such as the silk cotton tree, the wawa tree (African whitewood, a hardwood), and the African mahogany. The evergreen forest is in the extreme southwest, where the rainfall exceeds 65 inches a year, while there is a semideciduous forest further north.

The forest zone

The dense forest zone covered an area of 30,000 square miles until farming activities and timber exploitation reduced it to about 10,000 square miles, including about 6,000 square miles of forest reserves.

The third vegetation type, the northern savanna, is found in the northern two-thirds of the country, where the low annual rainfall, between 45 and 30 inches, occurs in a single season and is followed by a period of intense drought. Here, the vegetation consists mostly of tall Guinea grass, together with a scattering of low trees, such as the shea butter tree, and various species of acacia. Along the northern border the Guinea savanna gives way to a more open type of grassland that has developed largely as a result of prolonged human interference with the natural vegetation pattern.

Types of soil

There is a close relationship between the major soil types and the broad geological, climatic, and vegetational features. Topography also plays an important local role. Throughout the country, weathering, leaching, and the formation of hard pans (hard impervious layers, composed chiefly of clay cemented by relatively insoluble materials), by capillary movement (the movement of water containing mineral salts to the surface) and evaporation, are common processes that vary in importance according to the characteristics of each locality. Leaching is more pronounced in the wet south, while the formation of hard pans and laterite is most widespread in the drier north. In general, most soils are formed in place from parent rock material, which has been subjected to prolonged denudation and consequently has limited fertility.

In the forest zone the soils are mostly lateritic, ranging in colour from reddish brown to orange brown. They are subdivided into relatively fertile and less acidic ochrosols (red, brown, and yellow-brown, relatively well-drained soils) in areas of moderate rainfall and into more highly acidic and less fertile oxysols in the extreme southwest, where the annual rainfall exceeds 65 inches. Ochrosols also occur over considerable areas within the coastal and northern savanna zones. As in the forest zone, they provide the best soils for agriculture.

The varied geology of the coastal savanna zone is reflected in an abundance of soil types, including tropical black earths, tropical gray earths, acid vleisols, and sodium vleisols. Except for the tropical black earths, known locally as Akuse clays, most of these soils are of little importance agriculturally. The Akuse clays fill a broad zone across the coastal savanna plains; although heavy and intractable, they respond well to cropping under irrigation and mechanical cultivation.

Because of their intrinsic poverty in nutrients, most of the soils are heavily dependent upon the humus supplied by the vegetation cover. There is, thus, a delicate balance between vegetation and soil fertility, which may be upset by uncontrolled burning, or overuse.

Animal life. Although it has been depleted by too much hunting and the spread of human settlement, Ghana is still comparatively rich in animal life. The large mammals include lions, leopards, hyenas, antelope, elephants, buffalo, wild hogs, chimpanzees, and many kinds of monkeys, including the black and white *Colobus* monkeys. Among the snakes are pythons, cobras, horned and puff adders, and green mambas. Crocodiles and the diminishing remnants of manatees (large water animals characterized by two flippers and a spoon-shaped tail) and otters are found in the rivers and lagoons. Hippopotamuses are found in the Volta River. There are many varieties of lizards and tortoises and giant snails.

Among the numerous birds are parrots, hornbills, kingfishers, eagles, kites, herons, cuckoos, nightjars, sunbirds, doves, pigeons, egrets, swallows, vultures, snakebirds, and plantain eaters.

Fish and other aquatic life

The adjacent ocean, the rivers, and the inland lakes are rich in fish and other forms of life. Herring in large shoals arrive seasonally in the coastal waters; other fish include mackerel, soles, skates, mullet, bonitos, flying fish, lungfish, elephant fish, sea bream, and sharks. Edible turtles, barracuda, and stingrays are fairly common, mussels and several varieties of crabs, lobsters, and prawns also are found.

Insect life is particularly abundant. There are beetles, fireflies, ants, termites, butterflies, crickets, and bugs. Among the most dangerous insects are mosquitoes, tsetse flies, and simuliids (biting flies), which are responsible for transmitting the endemic diseases of malaria and yellow fever, trypanosomiasis (sleeping sickness), and onchocerciasis (a parasitic disease), respectively.

The Mole Game Reserve in the western part of the Northern Region is 1,500 square miles in extent; other reserves are planned further south, notably on some of the islands in Lake Volta.

Traditional regions. Ghana falls naturally into three major regions, the boundaries of which are not always clearly defined.

The coastal zone. Apart from the urban centres, which are found here more frequently than in any other region, the coastal zone is traditionally a region of fishermen and small-scale food farmers. The region is occupied by a series of small kingdoms the inhabitants of which were the first to be exposed to European contact—from the 15th century onward, perhaps even earlier. From east to west, the principal ethnic groups are the Ewes, the Adangmes (Adangbes), the Gas, the Effutus, the Fantis, the Ahantas, and the Nzimas. While it is difficult to speak of any cultural cohesion embracing the whole region, the seaboard has made the region an important centre for commerce, causing population concentration exemplified by such urban centres as Accra, Cape Coast, Sekondi-Takoradi, and the new port and industrial town of Tema, a few miles east of Accra.

The forest zone. Further inland, occupying about a third of the country, is the forest region with its relatively large and prosperous traditional states and rich agricultural lands. West of the Volta these states consist mostly of Akan peoples; to the east the Ewes predominate. The forest environment and the economic activities and modes of life engendered by it, especially since the introduction of cocoa farming in 1879, have served to give the region a common stamp. Apart from the Ewes, the major ethnic groups are the Akwapims and Kwahus in the east, the Akims in the south, the Ashantis and Brongs in the centre and north, and the Wassaws and Sefwis in the west. While all of these peoples have a relatively long history of settlement and political activity to their credit, those with the most impressive record are the Ashantis, who, from the 17th to late 19th centuries, built up a large political empire centred on Kumasi, the present Ashanti capital, that included a large number of subject and satellite states spread throughout the region and in both the coastal and northern savanna zones.

The Ashantis

Practically all the timber, cocoa, and exploited mineral wealth of the country, as well as a number of minor cash crops grown for export, and a large proportion of the foodstuffs consumed in Ghana are derived from this one region. Population density is relatively high, amounting to an average of 50 to 200 persons per square mile, particularly in the cocoa-growing areas. Except for Kumasi, there are few really large urban centres.

The northern savanna zone. This region covers approximately two-thirds of the country but is economically the most backward and neglected. The largest ethnic groupings are the Dagombas, Gonjas, and Mossi (Moshis). The region has a harsh environment because of its low rainfall. The period of drought is known as the "hungry season." The southern part immediately adjoining the forest zone, forming part of the disease-ridden "middle belt" of West Africa that combines the worst features of the forest and the savanna environments, is especially unattractive for settlement. In the past it was also subject to extensive slave raiding, from both north and south. Its distance from the sea and consequent insulation from active European contact over a long period retarded the development of the northern region.

Among the advantages of the northern region, especially in the most northerly part, which is relatively free from the tsetse fly so deadly to cattle, is an extensive savanna vegetation that offers excellent prospects for livestock breeding. Its relatively light soils and the rainfall regime favour the cultivation of both yams and cereals. Although agriculture is mostly of the traditional subsistence type, irrigation and mechanized cultivation have opened up new prospects. The recently created Lake Volta extends far into the heart of the region and offers prospects for comparatively cheap communication with the south, as well as a reservoir of water for agricultural and other uses.

Population distribution is uneven, but average densities are generally low—*i.e.*, under 50 persons per square mile. In the Mamprusi area in the extreme northeast, however, where instead of compact villages the settlements consist of compound houses set amid their fields, densities of more than 200 persons to the square mile are found. Urban life is not well developed.

The landscape under human settlement. *Rural settlement.* Ghana's population is predominantly rural. In 1970 only 28 percent of the population lived in settlements of 5,000 or more, the remaining 72 percent being spread over a large number of small rural settlements, mostly with populations of under 1,000. Almost everywhere agriculture is extensive, rather than intensive, and the rural settlements form isolated nuclei surrounded by land that is either under crops or is undergoing regeneration.

The pattern of settlement

The only true examples of permanent or continuous cropping are to be found in the extreme northeast, where settlements consist of isolated compound houses, each surrounded by its own farm. Elsewhere, agriculture is based on rotational farming, a system in which land is cropped for two or three years and then abandoned for from four to seven years, in order to regenerate. When cocoa or other tree crops are grown, however, cultivation is usually permanent.

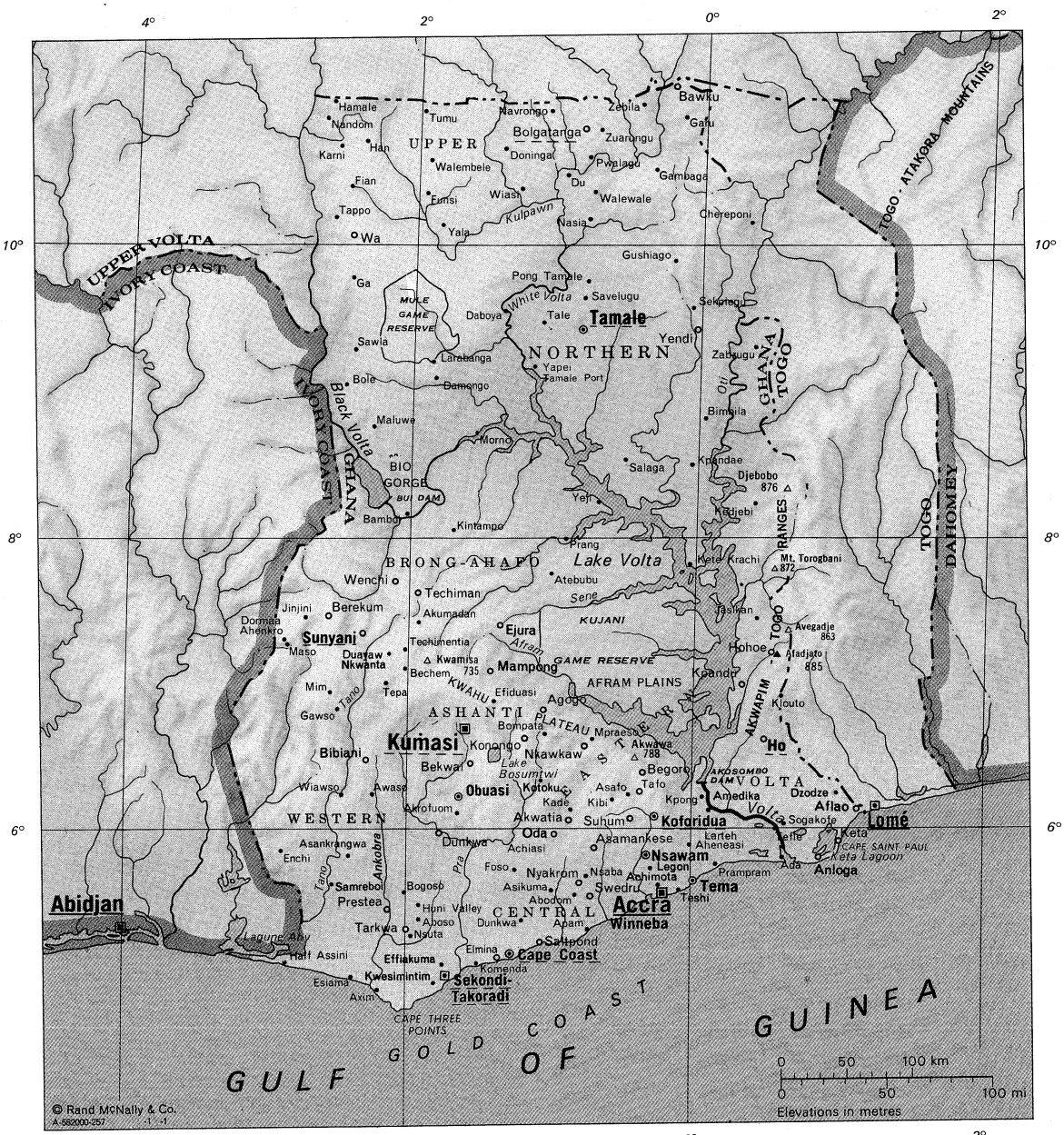
Urban settlement. Although many of the existing urban centres are expanding rapidly in size and population, they remain small by world standards. The Accra-Tema municipality, largest in the country, has a population of over 735,000, followed by Kumasi with about 350,000, and Sekondi-Takoradi with over 220,000 inhabitants.

PEOPLE AND POPULATION

Groups historically associated with the contemporary country. *Ethnic and linguistic groups.* Ethnically, the people of Ghana may be said to belong to the one broad group within the Negro family, but there is a large variety of tribal, or subethnic, units. On the basis of language, it is possible to distinguish at least 75 different tribes. Many of these, however, are very small, and only ten of them consist of more than 1 percent of the total population. The largest of these are the Akan, Mole-Dagbani, Ewe, and Ga-Adangme (Ga-Adangbe). Fortunately for Ghana, despite its tribal variety, no serious tribal divisions and animosities have asserted themselves since the country became independent in 1957, although tribal consciousness still persists in many quarters. At all levels in government and in public life, a conscious effort is made to play down tribal differences, a policy that has been helped in no small measure by the adoption of English as the country's official language.

The largest tribal groups

Religious groups. The main religions in Ghana are the indigenous religions of the various tribes, Christianity, and Islām. Although the indigenous religions are widespread and deep-rooted, they are completely lacking in any systematic body of doctrines, and are only vaguely conceptualized. Though they are based, in general, on belief in the existence of a supreme being, a number of lesser deities associated with various natural phenomena



GHANA

MAP INDEX

Political subdivisions

Accra..... 5·40n 0·10w
Ashanti..... 6·45n 1·30w
Brong-Ahafo..... 7·45n 1·30w
Central..... 5·30n 1·00w
Eastern..... 6·30n 0·30w
Northern..... 9·30n 1·00w
Upper..... 10·30n 1·30w
Volta..... 7·00n 0·30e
Western..... 5·30n 2·30w

The name of a political subdivision if not shown on the map is the same as that of its capital city.

Cities and towns

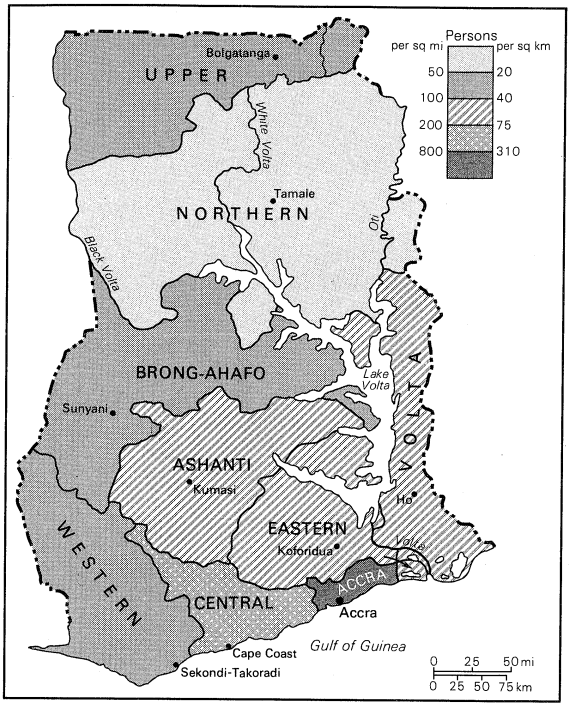
Abodom..... 5·32n 0·49w
Aboso..... 5·22n 1·56w
Accra..... 5·33n 0·13w
Achiasi..... 5·52n 1·00w
Achimota..... 5·37n 0·14w
Ada..... 5·47n 0·24e
Aflao..... 6·07n 1·11e
Agogo..... 6·47n 1·04w
Akrofuom..... 6·07n 1·39w
Akumadan..... 7·24n 1·57w
Akwatia..... 6·04n 0·49w
Amedika..... 6·06n 0·08e
Anloga..... 5·48n 0·54e
Apam..... 5·17n 0·44w
Asafo..... 6·11n 0·28w
Asamankese..... 5·52n 0·42w
Asankrangwa..... 5·48n 2·26w
Asikuma..... 5·35n 1·00w
Atebubu..... 7·45n 0·59w
Awaso..... 6·14n 2·16w
Axim..... 4·52n 2·14w
Bamboi..... 8·10n 2·02w
Bawku..... 11·05n 0·14w
Bechem..... 7·05n 2·02w
Begoro..... 6·23n 0·23w
Bekwai..... 6·27n 1·35w
Berekum..... 7·27n 2·37w
Bibiani..... 6·28n 2·20w
Bimbila..... 8·51n 0·04e
Bogoso..... 5·34n 2·01w
Bole..... 9·02n 2·29w
Bolgatanga..... 10·46n 0·52w
Bompata..... 6·38n 1·04w
Cape Coast..... 5·05n 1·15w
Chereponi..... 10·09n 0·17e
Daboya..... 9·32n 1·23w
Damongo..... 9·05n 1·49w
Doninga..... 10·37n 1·26w
Dormaa Ahenkro.. 7·17n 2·53w
Du..... 10·30n 0·59w
Duayaw Nkwanta. 7·10n 2·06w
Dunkwa..... 5·58n 1·46w
Dzodze..... 6·14n 1·00e
Effiakuma..... 5·06n 1·39w
Efiduasi..... 6·51n 1·24w
Ejura..... 7·23n 1·22w
Elmina..... 5·05n 1·21w
Enchi..... 5·49n 2·49w
Esiama..... 4·56n 2·21w
Fian..... 10·23n 2·29w
Foso..... 5·42n 1·17w
Funsi..... 10·17n 1·58w
Ga..... 9·47n 2·30w
Gambaga..... 10·32n 0·26w
Garu..... 10·51n 0·11w
Gawso..... 6·48n 2·31w
Gushiago..... 9·55n 0·12w
Half Assini..... 5·03n 2·53w
Hamale..... 10·59n 2·44w
Han..... 10·41n 2·27w
Ho..... 6·35n 0·30e
Hohoe..... 7·09n 0·28e
Huni Valley..... 5·28n 1·55w
Jasikan..... 7·24n 0·28e
Jinjini..... 7·26n 2·39w
Kade..... 6·05n 0·50w
Karni..... 10·40n 2·37w
Kedjebi..... 8·12n 0·25e
Keta..... 5·55n 1·00e
Kete Krachi..... 7·46n 0·03w
Kibi..... 6·10n 0·33w
Kintampo..... 8·03n 1·43w
Koforidua..... 6·03n 0·17w
Komenda..... 5·03n 1·29w
Kononga..... 6·37n 1·11w
Kotoku..... 6·12n 1·07w
Kpandae..... 8·28n 0·01w
Kpandu..... 7·00n 0·18e
Kpong..... 6·09n 0·04e
Kumasi..... 6·41n 1·35w
Kwesimintim..... 4·54n 1·47w
Larabanga..... 9·13n 1·51w
Larteh Aheneasi... 5·56n 0·04e
Legon..... 5·39n 0·11w
Maluwe..... 8·40n 2·17w
Mampong..... 7·04n 1·24w
Maso..... 7·14n 2·53w
Mim..... 6·54n 2·34w
Morno..... 8·41n 1·31w
Mpraeso..... 6·35n 0·44w
Nandom..... 10·51n 2·45w
Nasia..... 10·09n 0·48w
Navrongo..... 10·54n 1·06w
Nkawkaw..... 6·33n 0·47w
Nsaba..... 5·39n 0·45w
Nsawam..... 5·50n 0·20w
Nsuta..... 5·17n 1·58w
Nyakrom..... 5·37n 0·48w
Obuasi..... 6·14n 1·39w
Oda..... 5·55n 0·59w
Pong Tamale..... 9·41n 0·49w
Prampram..... 5·42n 0·07e
Prang..... 7·59n 0·53w
Prestea..... 5·27n 2·08w
Pwalagu..... 10·35n 0·50w
Salaga..... 8·33n 0·31w
Salt Pond..... 5·12n 1·04w
Samreboi..... 5·36n 2·34w
Savelugu..... 9·37n 0·49w
Sawla..... 9·17n 2·25w
Sekondi-Takoradi. 4·59n 1·43w
Sekpiegu..... 9·33n 0·02w
Sogakofe..... 6·00n 0·36e
Suhum..... 6·05n 0·27w
Sunyani..... 7·20n 2·20w
Swedru..... 5·32n 0·43w
Tafo..... 6·31n 0·22w
Takoradi, see Sekondi-Takoradi
Tale..... 9·26n 1·07w

MAP INDEX (continued)

Tamale..... 9·25n 0·50w
Tamale Port see Yapei
Tappo.....10·12n 2·38w
Tarkwa..... 5·19n 1·59w
Techiman..... 7·35n 1·56w
Techimentia..... 7·11n 2·02w
Tefle..... 5·59n 0·35e
Tema..... 5·38n 0·01e
Tepa..... 7·00n 2·10w
Teshi..... 5·35n 0·05w
Tumu.....10·52n 1·59w
Wa.....10·04n 2·29w
Walembele.....10·30n 1·58w
Walewale.....10·21n 0·48w
Wenchi..... 7·42n 2·07w
Wiasi.....10·21n 1·20w
Wiawso..... 6·12n 2·29w
Winneba..... 5·25n 0·36w
Yala.....10·07n 1·52w
Yapei (Tamale Port)..... 9·10n 1·10w
Yeji..... 8·13n 0·39w
Yendi..... 9·26n 0·10w
Zabzugu..... 9·17n 0·22e
Zebila.....10·56n 0·29w
Zuarungu.....10·47n 0·48w

Physical features and points of interest

Afadjato, *mountain*..... 7·05n 0·35e
Afram, *river*..... 7·00n 0·52e
Afram Plains..... 6·50n 0·10w
Akosombo Dam... 6·16n 0·03e
Akwapim-Togo Ranges, *mountain range*..... 7·25n 0·40e
Akwawa, *hill*..... 6·27n 0·25w
Ankobra, *river*.... 4·53n 2·17w
Bio Gorge..... 8·20n 2·20w
Black Volta, *river*.. 8·41n 1·33w
Bosumtwi, Lake... 6·30n 1·25w
Bui Dam..... 8·22n 2·10w
Djebobo, *mountain*..... 8·21n 0·38e
Gold Coast..... 5·20n 0·45w
Guinea, Gulf of.... 4·40n 1·30w
Keta Lagoon..... 5·54n 0·56e
Kujani Game Reserve, *wildlife refuge*.... 7·10n 0·50w
Kulpawn, *river*....10·21n 1·05w
Kwahu Plateau... 6·30n 0·30w
Kwamisa, *mountain*..... 7·08n 1·53w
Mole Game Reserve, *wildlife refuge*.... 9·30n 2·00w
Oti..... 8·40n 0·13e
Pra, *river*..... 5·01n 1·37w
Saint Paul, Cape.. 5·49n 0·57e
Sene, *river*..... 7·30n 0·33w
Tano, *river*..... 5·07n 2·56w
Three Points, Cape..... 4·45n 2·06w
Togo-Atakora Mountains.....10·45n 1·30e
Torogbani, Mount, *mountain*..... 7·45n 0·37e
Volta, Lake..... 7·30n 0·15e
Volta, *river*..... 5·46n 0·41e
White Volta, *river*.. 9·10n 1·15w



Population density of Ghana.

are also recognized. Considerable prominence is given to dead ancestors, who are considered to be ever present, capable of influencing the course of events for the living, and capable of serving as intermediaries between the living and the divinities.

Over the years, Christianity has steadily gained ground at the expense of the indigenous religions. Christian influence is most dominant in the southern part of the country, while Islām is strongest in the extreme north and in the larger urban centres, which contain some immigrant populations from Muslim regions of West Africa. Since the 1950s a large number of spiritual churches claiming adherence to Christianity have appeared, but the main divisions of the Christian Church remain: Roman Catholics, who constitute more than 14 percent of the adult population; Methodists (10 percent); Presbyterians (just under 10 percent); Anglicans (less than 3 percent), and other smaller denominations (less than 7 percent). Islām is adhered to by about 12 percent of the adult population and 38 percent adheres to the traditional religions.

Demography. On the basis of the 1970 census, the population is generally taken to be about 8,500,000, and to be increasing at the very high rate of between 2.4 and 3 percent a year. Detailed analyses of the 1970 census are not yet available, so that the following analysis is based on 1960 figures.

With an average population density of 100 persons to the square mile, Ghana is still far from overcrowded, although there is concern for the future because of the relationship between available resources and the rapid population growth.

In 1960, when the population was about 6,700,000, there was a slight excess of males over females, due to an immigrant population of more than 800,000. Despite governmental measures to control immigration, Ghana still attracts immigrants, especially from Upper Volta, Togo, and Nigeria.

Life expectancy

A striking characteristic is the youthfulness of the population as a whole. In 1960 about 12 percent of the entire population were 45 years old or above, and 44 percent were under 15 years old. Thanks to medical facilities and advances in medicine, life expectancy has increased in recent years; the average life expectancy is now 48 years, which in 1921 was perhaps about 28 years, and the figure is generally higher than this in the urban centres of the south. Although infant mortality rates are high, the ratio of survival among children is increasing.

THE NATIONAL ECONOMY

A comparative view. Despite the economic difficulties it is now facing, Ghana is still one of the richest and most developed countries in tropical Africa. In 1968 the per capita income was $238, the fourth highest in tropical Africa. There has, however, been little substantial change in the pattern of the economy since independence was attained in 1957. Exports are still dominated by primary raw materials, both mineral and agricultural, while the bulk of imports consist of manufactured and capital goods. In recent years, however, attempts have been made to broaden the basis of agricultural exports, and local manufacturing industry for the domestic market is gradually coming into prominence. The completion of the Volta hydroelectric project in the 1960s also added a significant new factor to economic life.

Natural resources. *Mineral resources.* While Ghana possesses a wide range of minerals, only a few, namely gold, diamonds, manganese and bauxite—in that order of importance—are exploited. These are found mostly in the southern part of the country. The gold industry, with an unbroken history going back 500 years, is the oldest; the others are of more recent origin—the working of manganese dating from 1915, diamonds from 1919, and bauxite from 1942. Recently, reasonable promising reserves of iron ore and limestone have been located, and

Ghana, Area and Population

	area		population	
	sq mi	sq km	1960 census	1970 census
Capital District				
Accra	995	2,577	492,000	849,000
Regions				
Ashanti	9,417	24,390	1,109,000	1,477,000
Brong-Ahafo	15,273	39,557	588,000	763,000
Central	3,815	9,881	751,000	893,000
Eastern	7,698	19,938	1,094,000	1,263,000
Northern	27,175	70,383	532,000	729,000
Upper	10,548	27,319	757,000	857,000
Volta	7,943	20,572	777,000	947,000
Western	9,236	23,921	626,000	768,000
Total Ghana	92,100	238,539*	6,727,000†	8,546,000†

*Converted area figures do not add to total given because of rounding. †Figures do not add to total given because of rounding.
Source: Official government figures.

in 1970 oil was discovered by offshore prospecting between Saltpond and Cape Coast. Although this discovery was initially classified as noncommercial, the steep world oil price increases of 1973–74 caused the Ghanaian government to reclassify it as commercial in 1974 and to undertake development. In that year there were discoveries of substantial amounts of natural gas offshore to the south and west of Cape Three Points. Although neither of these developments was expected to improve Ghana's energy situation or balance of payments substantially, they did lead to a revival of exploration activity in the mid-1970s. Salt is obtained from the sea and lagoons.

The mining industry

During the 1960s there was a general decline in the mining industry, and the possibility of further discoveries or exploitation of gold, diamonds, and manganese was thought to be unpromising. During the 1970s, however, increased world market prices for gold and manganese, new discoveries in the Prestea goldfields, and the formation of a National Manganese Corporation in 1975 to carry out a five-year plan to rehabilitate the manganese mines at Nsuta all combined to improve prospects for these metals. Bauxite reserves amount to more than 300,000,000 tons, though the output, all of which is exported, is only about 400,000 tons a year. A giant aluminum smelter, completed at Tema in 1966, relies entirely on imported alumina, although the use of bauxite found nearby, at Kibi, for conversion into alumina to feed the smelter is under development. There are also extensive supplies of building stone, gravel, and sand. High-quality sand in the Tarkwa mining area provides the basis for a small but important glass industry. Cement factories have been developed at Tema and Takoradi.

Biological resources. Biological resources, unlike mineral, are not finite. Soil and climatic conditions favour a wide range of crops. By far the most important of these is cocoa, of which Ghana is the leading world producer. Maize (corn) is gradually becoming an important cash crop and in 1975–76 was exported to Mali, Somalia, Upper Volta, and Togo. Other agricultural products that are exported are kola nuts, copra, lime juice, palm oil and kernels, rubber, and bananas.

Timber and the crops of the forest zone constitute additional important biological resources. Yams and such cereals as rice and millet are produced primarily in the northern savannah zone; cattle are also raised there. Ghana's offshore waters are rich in fish, and the creation of Lake Volta added another important source of fish for the domestic market. The various types of fish caught include cape hake, grunt, sea bream, tilapia, herring, mackerel, barracuda, and tuna. Most of the catch is sun-dried or smoked and is consumed locally. Government and private research agencies are carrying out a project designed to increase the cash income of rural fishing communities and also to modernize processing, distribution, and marketing of the catch. Game is another source of protein.

Hydroelectric resources. Apart from offshore oil, Ghana's only sources of power lie in the hydroelectric potential of its rivers, many of which have the requisite regimes and rates of flow to permit exploitation. The

The Akosombo Dam

Akosombo Dam on the Volta River has a total generating capacity of 768,000 kilowatts, which is to be augmented by the construction of an additional dam a few miles downstream at Kpong. Plans exist for the construction of another dam at Bui on the Black Volta, a short distance above the head of Lake Volta, for the generation of a further 190,000 kilowatts. Electricity from Akosombo, nowever, meets all of the domestic and industrial requirements, leaving a surplus for sale to Togo and Dahomey. Half the power from Akosombo is consumed by the aluminum smelter at Tema; the rest is distributed throughout southern Ghana by a grid.

Sources of national income. National income is derived primarily from agricultural and mineral output and only to a limited extent from manufacturing and services. Most of the agricultural and mineral products are for export.

Agriculture, forestry, and fishing. Apart from providing the bulk of national income, agriculture, forestry, and fishing constitute the preponderant occupations, involving about 60 percent of the population. The annual output of cocoa is around 400,000 tons, making Ghana the world's leading producer; cocoa provides about half of the country's total revenue from exports. Cocoa is of such vital importance that the world price paid for the crop directly determines Ghana's economic fortunes. Predictions that Ghana's cocoa industry would be second to that of the Ivory Coast by the mid-1980s led the government to undertake programs intended to assure its continued leadership; these included cultivation of virgin areas in Brong-Ahafo and Western regions, rehabilitation of farms withdrawn from production, and research into disease control at the Cocoa Research Institute.

Cocoa exports

Timber is also an important foreign exchange earner (9 to 17 percent of export value in the 1970s). Ghana's timber marketing is controlled by the Ghana Timber Marketing Board. From the beginning of 1973 all foreign-owned timber export firms were required to incorporate locally, so that the purchasing of timber is conducted only through the Ghana Timber Marketing Board.

The importance of the Ghanaian domestic market should not be underestimated; the value of food produced for local consumption is very considerable, as is that of fish, both marine and freshwater.

Mining and quarrying. As mentioned above, the four principal minerals extracted are gold, diamonds, manganese, and bauxite; these account for a considerable part of the country's export earnings.

Manufacturing. A policy of industrialization has resulted in the establishment of a wide range of manufacturing industries, producing food products, beverages, tobacco, textiles, clothes, footwear, timber and wood products, chemicals and pharmaceuticals, and metals, including steel and steel products. These are produced mostly for local consumption. Among the announced program directives of the five-year plan begun in 1975, however, was the maintenance of a reasonable balance on external trade, and certain major industrial projects under way in the mid-1970s were directed to the export market in either the short or long term. These included an expansion of the Aboso glass factory, to be completed by 1978; reactivation in 1976 of a large prefabricated concrete panel factory in Accra; and construction of an asphalt plant at the Tema oil refinery utilizing residual products of the plant's operations.

Management of the economy. *Private and public sectors.* The economy is a mixture of private and public enterprise. Before independence the government's role was confined mainly to the provision of such basic utilities as water, electricity, railways, roads, and postal services. Agriculture, commerce, banking, and such industry as existed were almost wholly in private hands, with foreign companies and interests controlling the greater share in all of them except agriculture.

Shortly after independence, the government set out to extend its control over the economy by setting up a large number of state-owned enterprises in both agriculture and industry. At the same time, in order to make up for the local shortage of capital and entrepreneurial skills, various measures were adopted to attract foreign investors operating either wholly on their own or in partnership with the government. These policies did not achieve the desired results because of poor planning and corrupt administration, and by the time the administration of Pres. Kwame Nkrumah (*q.v.*) was overthrown, in 1966, the crippling weight of the heavy overseas borrowing upon which the government had relied for the realization of its ambitious economic programs had succeeded not only in dissipating practically the whole of the country's overseas reserves but also in saddling the economy with external and internal debts totalling some $1,000,000,000.

One of the first concerns of governments since that time has been the rehabilitation of the economy. They have sought to deal with the adverse balance of payments, to arrest inflation, to secure a rescheduling of overseas debts, to increase agricultural productivity, and to establish industrial development on a rational basis,

as well as to save scarce foreign exchange by encouraging exports of locally manufactured goods.

Between 1966 and 1972 there was a marked contraction in governmental involvement in economic matters. The government continued to provide basic utilities, however, and remained the largest single employer of labour. After the military coup of 1972, policy makers returned to the concept of a centralized economy. The considerable debt owed to four British companies was repudiated, imports were cut, industrial projects abandoned after the fall of Nkrumah were resuscitated, and a policy of increased nationalization and state control was begun.

Taxation. A large proportion of government revenue is derived from taxation, levied in a variety of forms, including a duty on cocoa, import duty, customs and excise duties, sales tax, income tax, and a number of other taxes. Tax concessions are available to businessmen. A surcharge on many imports, including the processed and semi-processed raw material on which many local industries depend, was introduced.

Tourism. Revenue from tourism is gradually on the increase, with most of the tourists coming from the United States, United Kingdom, West Germany, Canada, Nigeria, Upper Volta, and Ivory Coast. The government has established two bodies—the Ghana Tourist Control Board and Ghana Tourist Development Company—for the regulation, financing, and development of the tourist industry. Hotels are located at Accra, Tema, Takoradi, and Kumasi, and there is a hotel at Akosombo overlooking Lake Volta. Construction of a resort complex, including a large hotel, at Ada, a coastal town about 60 miles (100 kilometres) east of Accra, began.

Trade unions. The trade-union movement played a role in the struggle for self-government, and after independence the government, no doubt recognizing the importance of the movement as a political force, sought to make it a more direct instrument of policy. All trade unions in the country were brought under the authority of the Trade Union Congress, which was virtually an integral part of the government; this development curtailed the freedom of workers to bargain with employers in general and with the government in particular. After the fall of the Nkrumah government, the monopoly of the Trade Union Congress was abolished and other unions were able to function. In 1972, however, restrictions were again placed on labour activity.

TRANSPORTATION

The principal means of transport, in order of importance, are motor vehicles, railways, and aircraft. Animals are scarcely used except in the extreme north, where horses and donkeys are sometimes employed. Ghanaians are inveterate travellers, and all available forms of transport are well patronized for social as well as business purposes.

Roads and railways. The density of roads and railways is much greater in the southern part of the country than in the north, and even in the south the cocoa-growing areas and the coastal zone tend to be favoured at the expense of other parts. It has been estimated that a combined road and rail network at least 30,000 miles in length is needed to ensure that, on the average, no place will be more than three miles from a road or railway. The country has around two-thirds of that mileage.

Rail transport

Rail transport was introduced in the early 20th century, but only a small part of the country is served by it. The rail system forms a triangle joining Sekondi-Takoradi, Kumasi, and Accra; in addition, a central line runs from Huni Valley, on the Takoradi–Kumasi line, to Kade in the centre of the triangle, and an extension joins Achiasi on this central line with Kotoku on the Accra–Kumasi line. The port of Tema is also linked to the system by a short extension running from Achimota, near Accra. The Takoradi–Kumasi line is also joined by two other branch lines, one running from Tarkwa out to the gold-mining town of Prestea and the other running from Dunkwa to the bauxite-mining town of Awaso.

Apart from the fact that its length is limited, rail transport is much less popular than road transport; railways are primarily used for the transport of freight, especially minerals and logs, although even this declined after the introduction of heavy motor transport.

The road system

Motor transport, now widespread and popular, was introduced in the towns about 1912 and spread quickly to the cocoa-producing areas. While the railways are owned by the state, motor transport is almost entirely in private hands; the state's activities are confined to municipal bus services and express coach and freight haulage services between the larger towns.

Road quality ranges from first-class paved (asphalt-surfaced) roads to third-class unsurfaced roads. First-class roads run between the large urban centres; they include the coastal Accra–Sekondi-Takoradi–Axim road; the Accra–Kumasi road, which continues northward to Tamale; and the Accra–Ho, Accra–Keta, Cape Coast–Kumasi, and Sekondi-Takoradi–Kumasi roads. The best road in Ghana is a concrete-surfaced highway, with two lanes in each direction, running from Accra to Tema. Second-class roads are narrower than first-class roads and have a base of swish (sun-dried earth) rather than quarried stone.

The creation of Lake Volta occasioned an interruption on the Kumasi–Tamale road, where the Yeji ferry, formerly only a few yards wide, became a crossing of almost seven miles. A number of other ferries across the Volta completely disappeared.

Air transport. Small airports located at Takoradi, Kumasi, and Tamale are used solely for domestic services, while the Kotoka International Airport at Accra handles both domestic and international flights. The Accra runway is long enough to accommodate large jet aircraft, and the airport has a large terminal building. Domestic services are operated by a state-owned corporation, which also operates a West African service linking the coastal states as well as an international service to the Middle East, Europe, and the United Kingdom.

While air transport is popular in Ghana, the maintenance of the national airways is costly and requires a large annual governmental subsidy.

Water transport. Sea transport, formerly important, has dwindled with the expansion of air services. Most goods entering and leaving the country are, however, carried by sea. There are two modern harbours, Takoradi (opened in 1928) and Tema (opened in 1961). Takoradi specializes in exporting timber, manganese, and bauxite, while Tema specializes in cocoa. Both ports also handle passengers. In terms of tonnage, Tema and Takoradi handle about equal amounts of cargo. Ghana has its own national shipping company, the Black Star Line, Ltd., but ships from many other countries also use Ghanaian ports; traffic is mostly with Europe, the United States, and the Far East. The Ghana Shippers' Council was established by the government to provide for mutual consultation among shippers, shipowners, railways, and ports authorities. The Ghana Railway and Ports Authority is responsible for the country's port operations under the Ministry of Transport and Communications.

Launch services on Lake Volta are being expanded; a launch service is also maintained on the lower reaches of the Volta between Ada and Amedika.

ADMINISTRATION AND SOCIAL CONDITIONS

The coup of 1972

The structure of the government. On January 13, 1972, the armed forces overthrew the government of Ghana in a bloodless coup. The constitution was suspended, the office of the presidency abolished, and the National Assembly dissolved. The military leaders then established a National Redemption Council (NRC) as the main organ of government. At its formation, the membership of the NRC was tribally balanced, with four Akan, four Ewe, two Ga, and two Northern members. A civilian National Advisory Council later was established to advise on political matters.

In October 1975, however, the NRC was superseded as the highest legislative and administrative organ by a seven-member Supreme Military Council (SMC); at the same time, the NRC was reorganized to include the members of the SMC and certain other civilian and military members. Control of both organs was military.

Government under the constitution of 1969. The constitution provided for a unicameral parliamentary form of government with a president as head of state and a prime minister as head of government. The president was elected for a term of four years (with the possibility of re-election for one further term) by an electoral college consisting of the National Assembly members and representatives of the various houses of chiefs and district councils throughout the country. The prime minister, who was appointed by the president, was either the leader of the party with the largest plurality in the National Assembly or—if no party had a clear majority—the member of the National Assembly most likely to command the support of the rest of the members.

Assisting the prime minister was a Cabinet of eight to 17 members. There was also provision for the appointment of as many as 10 non-Cabinet ministers, subject to an overall limit of 21 ministers. All ministers were appointed by the president from among members of the National Assembly on the nomination of the prime minister.

The National Assembly sat in Accra and, together with the president, made up Parliament. It consisted of 140 to 150 members elected by universal adult suffrage for terms of five years, although there was provision for an extension, in the event of an emergency, and for an earlier dissolution by the president under certain circumstances. The minimum voting age for elections, as well as for membership of Parliament, was 21. Proceedings of the National Assembly were presided over by the speaker, who was elected by the members. During the absence of the president or in the event of any temporary vacancy in that office, the speaker acted as head of state.

The president was assisted and advised in some constitutional functions by a Council of State made up of the prime minister, the speaker of the National Assembly, the leader of the opposition, the president of the National House of Chiefs, and up to as many as 12 other persons appointed by the president. There was also provision for the appointment of an ombudsman and of a National Security Council, consisting of the prime minister and a number of other persons (including certain specified Cabinet ministers).

While the National Assembly constituted the legislature, the executive branch consisted of the president, the Council of State, the National Security Council, and the Cabinet; the various courts of the land constituted the judiciary (see below).

Local government. Apart from the capital district of Accra, Ghana is divided into eight regions—Ashanti, Brong-Ahafo, Central, Eastern, Northern, Upper, Volta, and Western regions. After 1972, when the constitution was suspended, each region was administered by a regional commissioner, who was an army officer. The structure of local administration was reorganized to include 37 district councils and 270 local councils.

Under the constitution of 1969, though it was essentially unitary in character, the underlying philosophy of local government was to delegate a considerable amount of power and responsibility to regional, district, and local councils in the management of their affairs. Certain civil service departments, such as education, health, public works, social welfare, parks and gardens, and fire services, were decentralized at the regional level; the rest were controlled centrally from Accra.

Local councils were made up of elected and traditional members, the traditional members being chiefs who constituted not more than one-third of the membership. Above these in the hierarchy were the district councils. Here, again, were elected and traditional members, with traditional members forming one-third and elected members two-thirds of each council. The two councils were responsible for the administration of their areas, subject to certain limitations, and were empowered to levy rates and taxes.

Regional councils

At the top were the regional councils, their memberships composed of representatives of all the district councils, chiefs from the regional house of chiefs, and certain heads of civil service ministries and departments. The chairman of each regional council was appointed by the prime minister; council functions included coordination and regional planning.

The political process. After the military coup of 1972 the lively political life of Ghana was brought under tight control. All political parties in existence before January 13, 1972, were proscribed, and the law prohibiting a one-party state was repealed. Members of the Progress Party, which had supported the former prime minister, were imprisoned for one year, and public meetings were prohibited.

Justice. Under the constitution of 1969 the judicial system was based chiefly on the English model, but Ghanaian customary law was recognized as well as English common law. The administration of justice was handled by various courts divided into two groups: the superior courts, consisting of the Supreme Court, the Court of Appeal and the High Court; and inferior courts, consisting, in descending order, of the circuit courts, the district courts, and other courts provided by law, such as the juvenile courts. After the military government took control, the Supreme Court was abolished, and the Court of Appeal, headed by a chief justice, became the highest court in Ghana.

The adjudicating authorities in chieftaincy and purely traditional matters were the regional and national houses of chiefs. Appeals from decisions of the National House of Chiefs were made directly to the Supreme Court.

Subversion Decree

After an abortive countercoup in July 1972 the government imposed the retroactive Subversion Decree. Under the authority of a military tribunal, the death penalty could be exacted for 10 offenses including subversive political activity, robbery, theft of specified items, and damaging public property. The decree was extended to include profiteering in 1973; in that year another decree was promulgated, prohibiting the spreading of rumours.

The armed forces. The Ghana armed forces consist of the army, the navy, and the air force, as well as such ancillary services as medical, transport, engineering, and workshop corps. All of the services are organized along British lines and equipped with relatively light but modern weapons. The headquarters is in Accra, where there is a military academy for the training of officers for all three services. Instruction at the academy is provided by Ghanaian officers and personnel from the United Kingdom and Canada; the armed forces are wholly officered by Ghanaians.

The navy has bases at Takoradi and Tema, and the main base of the air force is at Takoradi. Apart from its normal training duties, the navy carries out regular patrols for fishery protection and control of smuggling. The air force, organized in 1959, is equipped with a small but effective range of planes and helicopters. Besides its purely military functions, it assists the government with various development projects involving reconnaissance and mapping surveys and helps with quick delivery of supplies in times of emergency.

Social conditions. *Educational services.* Ghana has one of the best developed educational systems in the whole of tropical Africa, but the cost is high, representing about one-fifth of the total budget.

In April 1974 the government began implementation of a new educational system at the primary and secondary levels, consisting of a pre-primary cycle for ages four to six; a basic first cycle, including six years of primary education and three years of junior secondary; and a second cycle of variable length, leading to either secondary vocational or commercial certification programs or to a senior secondary upper course leading to university studies.

The first cycle is free and compulsory; for the first three years education is to be in the predominant local language, with provision for education in at least one other Ghanaian language and English, the latter to be the language of instruction from the fourth year of the primary cycle.

Teacher training and technical education are approximately equivalent to secondary education, though they tend to attract pupils who are not aiming at university

careers. The Tarkwa School of Mines offers a three-year diploma course in mining and related subjects.

The universities

University education is provided at three institutions: the University of Ghana, at Legon, near Accra; the University of Science and Technology at Kumasi; and the University of Cape Coast. Until 1971 university education was free for all qualified Ghanaians, but in the early 1970s a scheme was introduced to enable students to meet part of the cost themselves.

The enrollment in all schools, especially in secondary schools, has soared dramatically since Ghana achieved internal self-government. In the mid-1970s, there were about 1,500,000 pupils and students, from primary to university level. Nonetheless, it is estimated that as many as 30 to 50 percent of the children of elementary school age are still not in school because of lack of facilities.

In the 1960s an attempt was made to abolish tuition fees in all government-supported primary and middle schools, the aim being to make elementary education both free and compulsory. The attempt had to be abandoned indefinitely because of financial difficulties. From September 1973 students in primary, middle, and secondary schools and in teacher training colleges paid an annual fee for their textbooks and school materials.

There is a growing number of private schools at both elementary and secondary levels.

Health and welfare services. Major health problems are communicable diseases, poor sanitation, and poor nutrition. The main emphasis of government health policy is on improved public health, and since independence many improvements have been made in nutrition and in maternal and child care. Many of the endemic diseases, such as malaria, pneumonia, and diseases of the gastroenteritis group, which took a heavy toll of life, have been brought under a considerable measure of control as a result of improved hygiene, better drugs, and education. The results are to be seen in the decline in infant mortality and the rise in the general life expectancy.

Hospitals and clinics provided by the government and by various Christian missions exist in most parts of the country. Supplementary services consist of health centres, dispensaries, and dressing stations. Considerable progress has been made in the quantity and quality of health facilities and medical personnel, but rapid population growth continues to impose great pressures on the available facilities. In addition to the large number of doctors in the public service, many private practitioners operate their own clinics and hospitals. Registered doctors and dentists are supported by a paramedical staff of nurses, midwives, and pharmacists as well as by auxiliaries.

Rural community development

Government programs of rural community development are assisted by village improvement projects undertaken with the participation of the inhabitants. Welfare and economic aspects outside urban areas are handled by the national government. In the urban areas, welfare services concentrate on casework, probation work, youth activities, and guidance through voluntary organizations.

Housing. With the rapid growth of population and the movement of large numbers of people from the rural to urban areas, housing has become an acute problem, especially in the large cities where the problem is both quantitative and qualitative. In the rural areas the problem is mainly qualitative. There is distinct overcrowding in the urban areas, where the number of persons per house averages between 21 and 23, as compared with 10 in rural areas. Most housing is provided by private individuals, and the main role of the Ministry of Works and Housing is to supplement these efforts and to meet such special needs as the provision of low-cost housing and suitable local building materials.

Police services. The police service is directed by the Ministry of Internal Affairs. The training of police officers is conducted at the Ghana Police College in Accra. Each region has a regional police committee presided over by the regional chief executive.

The police force is divided into the general police, who are literate and are qualified to perform all of the functions belonging to the police; and the escort police, who are illiterate and are chiefly employed for patrol, escort, and guard duties. An armoured car squadron, established in 1959, can deal promptly with internal disturbances or threats to security. There is also a Women's Corps, mainly concerned with matters affecting women and children. Like the police service, the prison service comes under the Ministry of Internal Affairs, but the two are separate.

Wages and cost of living. Although there is a minimum wage for workers of somewhat less than $1 (U.S.) per day, the gap in wages between the lowest paid and well-paid workers is still wide. This, coupled with rising living costs, imposes severe hardships on a large section of the working population. The government is trying to lower the cost of living not only by increasing the supply of locally produced food staples but also by controlling the prices of other essential goods.

Egalitarian tendencies

Society. Ghanaian society is without sharp class distinctions. Insofar as traditional authority is based on a system of hereditary chieftaincy, it is possible to speak of aristocratic classes within the various tribal groups, but the institution of chieftaincy is essentially democratic in operation and the authority of chiefs is broadly based.

Another important social characteristic is that land is usually owned by families, militating against the emergence of a small and powerful landed class wielding economic control over a landless class. These inherent egalitarian tendencies of the society have been further heightened by economic and social mobility, depending on education and individual enterprise.

Cultural life and institutions. *The cultural milieu.* Ghana has a rich indigenous culture marked by great regional variety. Culturally, the peoples of Ghana have many affinities with their French-speaking neighbours, but each tribal group has distinctive cultural attributes. In all parts of the country the cultural heritage is closely linked with religion and the institution of chieftaincy. Various festivals and rites are centred on chieftaincy and the family and are occasioned by such events as harvest, marriage, birth, puberty, death and funerals.

The arts. Ghana's arts include dance and music, plastic art (especially pottery and wood-carving), gold- and silverwork, and textiles, most notably the richly coloured, handwoven kente cloth of the Akans and Ewes.

Despite the country's wealth in traditional art forms, indigenous art is in keen competition with various art forms of foreign origin, especially in those areas in which the end product is intended for practical household or personal use, such as pottery, carving, gold- and silversmithing, and weaving. Consequently, only the unique and most indispensable of these forms have managed to survive without special public support or patronage. The increased national self-consciousness generated in Ghana as in other African countries by the independence movement, however, has been instrumental in fostering and popularizing many art forms. Another and equally important factor has been the growth of tourism.

Cultural institutions. Apart from small indigenous groups of craftsmen who provide for the needs of the various chiefs' stools and skins throughout the country—a stool is the traditional symbol of office for chiefs in southern Ghana, and a skin is the equivalent symbol in the north—there are few properly established cultural institutions. The most outstanding are the National Cultural Centre, based in Kumasi, and the Arts Council of Ghana, based in Accra and with branches all over the country. The National Cultural Centre is primarily concerned with the cultural heritage of Ashanti, while the Arts Council is a nationwide organization concerned both with the preservation of indigenous Ghanaian culture in all of its forms and with its development and improvement in the light of contemporary local and world trends. Dance, music, drama, painting, and sculpture all come within the purview of the council.

A third cultural institution is the Ghana Museum and Monuments Board, based in Accra. The board has two museums (an ethnological museum and a science museum) in Accra and is responsible for the care and maintenance of buildings and relics of historical importance, such as the forts and castles, and for the preservation of

important art treasures throughout the country. The forts, built by various European powers, mostly between the 14th and 18th centuries, are all, except the Kumasi fort (1897), located on the coast. The oldest fort, São Jorge da Mina, at Elmina, was built by the Portuguese in 1482.

Press and broadcasting. The press in Ghana has a history dating back only to the late 19th century. Up until the time of independence, the press was wholly in private hands, and freedom of expression on political matters was severely restricted by the harsh operation of sedition laws. Following independence, however, the government established a number of state-owned papers, the role of which was to support and explain government policies. In 1957 the government established the Ghana News Agency as a corporation for the supply of home and foreign news to the local press and radio. Under the Nkrumah government, privately owned newspapers were subjected to severe pressures from government, and most of them ceased publication. Press freedom was later restored, although the two most influential newspapers were still controlled by the government in the 1970s. Press restrictions were reinstituted by the military government in 1972.

Ghana News Agency

Radio broadcasting in Ghana dates back to pre-independence days, while television broadcasting began only in 1965. Both radio and television broadcasting are controlled by the government. There is an external radio broadcasting service aimed chiefly at other African countries, but the television service is wholly domestic and generally confined to the larger urban centres and their immediate surroundings. On the home service, commercial and educational programs are available on both radio and television. The languages used by radio and television are mainly English and six Ghanaian languages —Twi, Ewe, Ga, Nzima, Hansa, and Dagbani.

Prospects for the future. Although Ghana is a comparatively small country in size and population, it enjoys significant variety in both its natural resources and its human material. Despite the long association with Britain and the many institutional and cultural innovations that this association brought about, thanks to the British policy of encouraging indigenous institutions and to the inherent vitality of these institutions themselves, the country has succeeded in retaining a surprisingly large measure of its own culture, particularly through the institution of chieftaincy and the even more pervasive institution of the family and lineage system.

Since independence in 1957, new and stronger pressures against the traditional ways of life have been felt because of the need to modernize and to raise living standards for a rapidly growing population, the survival of which depends on how well diverse social elements can be welded together into a single nation. As the first tropical African country under British rule to achieve independence, Ghana had bright prospects at the time of the British withdrawal for developing into a model nation with a strong economy. The high hopes entertained at the outset were dashed within 10 years after independence. Between 1967 and 1972 the country stood on the verge of new economic and political developments, but the problems confronting it were immense. Economic crises and an internal struggle for power precipitated the military coup of 1972. The next year the government published its Charter of Redemption, which announced its aims as including national solidarity, economic development, self-reliance, and service to the people.

BIBLIOGRAPHY. J. BRIAN WILLS (ed.), *Agriculture and Land Use in Ghana* (1962), the most comprehensive and authoritative account of the agricultural resources of the country and the problems involved in their development, including a particularly useful bibliography; E.A. BOATENG, *A Geography of Ghana*, 2nd ed. (1966), a concise and readable account of the systematic and regional geography of Ghana, fully illustrated with plates and diagrams; and "Ghana," in S.P. CHATTERJEE (ed.), *Developing Countries of the World*, ch. 15 (1968), a succinct account of the geographical aspects of the country's socioeconomic problems; WALTER BIRMINGHAM, I. NEUSTADT, and E.N. OMABOE (eds.), *A Study of Contemporary Ghana*, 2 vol. (1966–67), a collection of expert studies on selected aspects of the country's economic and social conditions up to the time immediately preceding the 1966 coup; GHANA, CENTRAL BUREAU OF STATISTICS, *Economic Survey*, an annual survey that provides a succinct statement of the country's economy supported with a wealth of statistical information; GHANA, CENSUS OFFICE, *1960 Population Census of Ghana* (1962–70), the most comprehensive work of its kind, and an indispensable source of statistical information on the population and socio-economic characteristics; the corresponding volumes of the *1970 Population Census of Ghana* began to appear in 1971. GHANA, PLANNING COMMISSION, *One Year Development Plan, 1970–71* (1970), provides useful background information on various aspects of the economy and its problems; *Seven-Year Plan for National Reconstruction and Development: Financial Years, 1963/64–1969/70* (1964), mainly a historical document but contains descriptive and statistical information on economic and social conditions; *Ghana Official Handbook, 1969* (1970), a valuable source of information on governmental structure and administration at that time; *Constitution of the Republic of Ghana* (1969), contains the provisions of the country's constitution of 1969; K.A. BUSIA, *The Position of the Chief in the Modern Political System of Ashanti* (1951), an authoritative account of the traditional status and functions of the chief and the character of the institution of chieftaincy; W.E.F. WARD, *A History of Ghana*, rev. 3rd ed. (1966), an important pioneering study of the traditional states of Ghana and their evolution into the modern state of Ghana; F.R. IRVINE, *Woody Plants of Ghana* (1961), a detailed and authoritative study of the major plants and their uses; A.H. BOOTH, *Small Mammals of West Africa* (1960), a useful pioneering work, fully illustrated; G.S. CANSDALE, *West African Snakes* (1961), by an acknowledged authority; J.H. ELGOOD, *Birds of the West African Town and Garden* (1960), a small but comprehensive study.

(E.A.B./Ed.)

Ghazālī, al-

Abū Ḥāmid Muḥammad ibn Muḥammad aṭ-Ṭūsī al-Ghazālī, prominent Islāmic jurist, theologian, and mystic, was one of the most powerful minds and important figures in the history of Islām, through whose efforts many Greek philosophical conceptions and methods were brought into the mainstream of Islāmic thought, while heretical aspects of philosophy were rejected. His abandonment of a brilliant career as a professor in order to lead a kind of monastic life won him many followers and critics among his contemporaries. Western scholars have been so attracted by his account of his spiritual development that they have paid him far more attention than they have other equally important Muslim thinkers.

Life

He was born at Ṭūs (near Meshed in eastern Iran) in AH 450 (AD 1058) and was educated there, then in Jorjān, and finally at Nishapur (Neyshābūr), where his teacher was al-Juwaynī, who earned the title of *imām al-ḥaramayn* (the *imām* of the two sacred cities of Mecca and Medina). After the latter's death in AH 478 (AD 1085), al-Ghazālī was invited to go to the court of Niẓām al-Mulk, the powerful vizier of the Seljuq sultans. The vizier was so impressed by his scholarship that in AH 484 (AD 1091) he appointed him chief professor in the Niẓāmīyah college in Baghdad. While lecturing to more than 300 students, he was also mastering and criticizing the Neoplatonist philosophies of al-Fārābī and Avicenna (Ibn Sīnā). He passed through a spiritual crisis that rendered him physically incapable of lecturing for a time. In November AH 488 (AD 1095) he abandoned his career and left Baghdad on the pretext of going on pilgrimage to Mecca. Making arrangements for his family, he disposed of his wealth and adopted the life of a poor Ṣūfī, or mystic. After some time in Damascus and Jerusalem, with a visit to Mecca in November AH 489 (AD 1096), al-Ghazālī settled in Ṭūs, where Ṣūfī disciples joined him in a virtually monastic communal life. In AH 499 (AD 1106) he was persuaded to return to teaching at the Neẓāmiyeh college at Nishapur. A consideration in this decision was that a "renewer" of the life of Islām was expected at the beginning of each century and his friends argued that he was the "renewer" for the century beginning in AH 500 (September, AD 1106). He continued lecturing in Nishapur at least until AH 504 (AD 1110) when he returned to Ṭūs, where he died December 18, AH 505 (AD 1111).

Works and teachings

Over 400 works are ascribed to al-Ghazālī, but he probably did not write nearly so many. Frequently the same work is found with different titles in different manuscripts; but many of the numerous manuscripts have not yet been carefully examined. Several works have also been falsely ascribed to him, and others are of doubtful authenticity. At least 50 genuine works are extant, some relatively short.

Al-Ghazālī's greatest work is *The Revival of the Religious Sciences* (*Iḥyā' ʿulūm ad-dīn*). In 40 "books" he explained the doctrines and practices of Islām and showed how these can be made the basis of a profound devotional life, leading to the higher stages of Ṣūfism or mysticism. The relation of mystical experience to other forms of cognition is discussed in *Mishkāt al-anwār* (*The Niche for Lights*). Al-Ghazālī's abandonment of his career and adoption of a mystical, monastic life is defended in the autobiographical work *al-Munqidh min aḍ-ḍalāl*, (*The Deliverer from Error*).

His philosophical studies began with treatises on logic and culminated in the *Tahāfut* (*The Inconsistency—or Incoherence—of the Philosophers*), in which he defended Islām against such philosophers as Avicenna who sought to demonstrate certain speculative views contrary to accepted Islāmic teaching. As preparation for this he published an objective account of *Maqāṣid al-falāsifah* (*The Aims of the Philosophers*; *i.e.*, their teachings). This book was influential in Europe and was one of the first to be translated from Arabic to Latin (12th century).

Most of his activity was in the field of jurisprudence and theology. Toward the end of his life he completed a work on general legal principles, *al-Mustaṣfā* (*Choice Part* or *Essentials*). His compendium of standard theological doctrine (translated into Spanish), *al-Iqtiṣād fī al-ʿtiqād* (*The Just Mean in Belief*), was probably written before he became a mystic, but there is nothing in the authentic writings to show that he rejected these doctrines, even though he came to hold that theology—the rational, systematic presentation of religious truths—was inferior to mystical experience. From a similar standpoint he wrote a polemical work against the militant sect of the Assassins (Ismā ʿīlīs), and (if it is authentic) a criticism of Christianity, as well as a book of *Counsel for Kings* (*Naṣīḥat al-mulūk*).

BIBLIOGRAPHY. W.M. WATT, *Muslim Intellectual: A Study of al-Ghazālī* (1963), deals with his life and the development of his thought. Still of value is D.B. MACDONALD, "The Life of al-Ghazālī with Special Reference to his Religious Experience and Opinions," *Journal of the American Oriental Society*, 20:71–132 (1899). M. SMITH, *Al-Ghazālī the Mystic* (1944), is good for his mysticism, but treats some dubious works as authentic. On this point, see W.M. WATT, "The Authenticity of the Works Attributed to al-Ghazālī," *Journal of the Royal Asiatic Society*, pp. 24–45 (1952).

(W.M.W.)

Ghāzān, Maḥmūd

Maḥmūd Ghāzān was the seventh and most gifted of the Il-Khans (subordinate *khāns*), the Mongol dynasty that ruled in Iran from AD 1256 to 1353 as the nominal vassals of the Great Khan (of Mongolia, later of China). Ghāzān's reign (1295–1304) was marked by the establishment of Islām as the state religion and by wars with the Mamlūk sultans of Egypt.

By courtesy of the Edinburgh University Library



Mahmūd Ghāzān receiving the nobles of Khorāsān, illumination from the Mongol manuscript *Jāmiʿ at-tawārīkh*, c. 1307. In the University of Edinburgh Library (MS. Or.20).

Birth and early life

Ghāzān was born November 5, 1271, at Abaskun near the present-day port of Bandar-e Shāh on the southeastern shore of the Caspian Sea, his father, Arghun, the later Il-Khan (1284–91), being then only 13 years old. Ghāzān's early childhood was spent largely in the company of his grandfather, the Il-Khan Abagha (1265–82), and he was brought up in the Buddhist faith that both his father and his grandfather professed. Upon his father's accession to the throne in 1284, Ghāzān was appointed viceroy of the provinces of northeastern Persia, where he resided for the next ten years, defending the frontier against the Chagatai Mongols of Central Asia and then against his own lieutenant Nawrūz, who had risen in revolt and made common cause with the Chagatai. Ghāzān's relations with Arghun's successor, Gaykhatu (1291–95), were cool; those with Baydū, the latter's cousin, who dethroned him and usurped the throne, came to open war. After a first encounter, followed by a truce and parley, Ghāzān spent the summer of 1295 in the mountains north of present-day Tehrān, where, on the advice of Nawrūz, with whom he was now reconciled, he declared himself a convert to Islām, and his example was followed by the troops under his command. It was thus at the head of a Muslim force that he resumed the attack against Baydū, who, deserted by his supporters, was captured and executed on the very day of Ghāzān's entry into the Il-Khanid capital of Tabriz.

Ghāzān was formally enthroned on November 3, 1295, and during the first year of his reign he had to cope with a number of revolts against his authority. These were suppressed with the utmost severity, no fewer than five princes of the blood being executed for their complicity. Nawrūz himself, who had helped raise Ghāzān to the throne, was soon to pay with his life for suspected collusion with the Mamlūks. Though now the Muslim head of a Muslim state, Ghāzān took up the hereditary quarrel of his family with these champions of Islām. In 1299–1300 he invaded Syria, defeated the Egyptian army at Homs, and made a triumphal entry into Damascus; but upon his return to Persia early in 1300 the country was re-occupied by the Mamlūks. In the autumn of the same year he returned to the attack; but poor weather rendered military operations impossible, and the campaign was abandoned before contact could be made with the enemy. For a third campaign he sought an alliance with the Christian West. In a letter to Pope Boniface VIII dated April 12, 1302, he refers to a detailed plan for the invasion of Syria, which he had previously proposed to the princes of Europe and continues:

> As for now, we are making our preparations exactly in the manner [laid down in that plan]. You too should prepare your troops, send word to the rulers of the various nations and not fail to keep the rendezvous. Heaven willing, we [*i.e.*, Ghāzān] shall make the great work [*i.e.*, the war against the Mamlūks] our sole aim.

The campaign to which Ghāzān here alludes was launched in the spring of 1303 without European aid. The Mongols advanced through Syria without meeting serious resistance until they were halted and decisively defeated south of Damascus. A fourth campaign was prevented by an illness that attacked Ghāzān in the autumn of 1303; he recovered for a while but then suffered a relapse and died on May 11, 1304.

Breadth of accomplishments

Ghāzān's accomplishments were in no way restricted to his activities on the battlefield. A man of great intellectual curiosity, he was conversant with natural history, medicine, astronomy, and chemistry and was also an adept in several handicrafts. "No one surpassed him," says the Byzantine historian Pachymeres, "in making saddles, bridles, spurs, greaves and helmets; he could hammer, stitch and polish, and in such occupations employed the hours of his leisure from war." Besides his native Mongolian, he is said to have had a knowledge of the Arabic, Persian, Hindi, Kashmiri, Tibetan, Chinese, and Frankish (*i.e.*, probably French) languages.

It was at his suggestion and with his assistance that his vizier Rashīd ad-Dīn composed a celebrated history of the Mongols, afterward expanded to embrace all the peoples of Asia and Europe with which their conquests

had brought them in contact. Rashīd ad-Dīn, Ghāzān's great minister, was perhaps the real author of the fiscal reforms that go under his master's name and that were designed to protect the sedentary population from the extortions of the nomad aristocracy. These measures, coupled with the adoption of Islām, must have played their part in welding the Mongols and Persians (like the Normans and English) into a single nation, and the Il-Khans might have ended, like the Plantagenets, by becoming a truly national dynasty. In fact, Ghāzān himself, by his ruthless elimination of princely rivals, must have contributed to the extinction of the Il-Khanids, who survived his death by little more than 30 years. In appearance he was short and ill-favoured, unlike his father, Arghun, who was tall and handsome. Commenting on his bravery in the battle at Homs, the Armenian prince Haithon observes at one point in his eyewitness account of the scene:

> And the most remarkable thing of all was that within a frame so small, and ugly almost to monstrosity, there should be assembled nearly all those high qualities which nature is wont to associate with a form of symmetry and beauty. In fact among all his host of 200,000 Tartars you could scarcely find one of smaller stature or of uglier and meaner aspect than this Prince.

BIBLIOGRAPHY. J.A. BOYLE (ed.), *Cambridge History of Iran*, vol. 5, *The Saljuq and Mongol Periods* (1968)—the personal history of Ghāzān is in Boyle's chapter on the Il-Khans, and his fiscal reforms in the chapter by I.P. PETRUSHEVSKY; these are based on Oriental, chiefly Persian, sources, of which the most important is the *Jāmiʿ at-tawārīkh* ("Universal History") of Rashīd ad-Dīn. There is a Russian translation by A.K. ARENDS (1957) of vol. 3 of this source, which deals with the Il-Khans. BERTOLD SPULER, *Die Mongolen in Iran*, 3rd ed. (1968), also based on Persian sources, including Rashīd ad-Dīn, is the only monograph on the Il-Khans.

(J.A.Bo.)

Ghent and Bruges, History of

Ghent and Bruges (Brugge) were the two most important towns of the county of Flanders in medieval times. Both cities are located in what is now Belgium. Ghent, provincial capital of Oost-Vlaanderen (East Flanders), stands at the junction of the Leie (Lys) and Scheldt rivers; Bruges, the provincial capital of West-Vlaanderen (West Flanders), is about 15 miles directly east of the coastal city of Ostend.

Settlement and growth. Ghent and its countryside were inhabited in Roman times. Ghent became important when the abbeys of St. Bavo and St. Peter were founded there (*c.* AD 630). The urban settlement was destroyed by Viking invasions in the 9th century, particularly in 879–881. Shortly afterward Count Baldwin II built a fortification, partly conserved today in the Castle of the Counts, that attracted a civilian settlement (*portus*) near its walls. Ghent in the 11th century became a commercial and industrial centre, with a cloth industry based on locally supplied wool.

The area of Bruges was also inhabited in Roman times. The town of Bruges grew as a seaport in the 4th–8th centuries AD and was the centre of the *pagus Flandrensis*, the Frankish name for the region around Bruges; Count Baldwin II made it the seat of his government in 892, and it remained the centre of the county's administration in succeeding centuries. Situated near the sea and being fortified, the town attracted a civilian settlement, which in time developed important relations with the English across the channel.

Urban autonomy. A great political crisis broke out when Count Charles the Good was murdered in Bruges in 1127. The townspeople helped crush the conspiracy; and, thanks to the initiative of Bruges and Ghent, Thierry of Alsace became the new count in 1128. Both Ghent and Bruges acquired urban magistrates (*scabini, schepenen*) as well as urban privileges during the time of the conspiracy.

In the 13th century the counts had to accept a large degree of urban autonomy, and power became concentrated in the hands of small urban aristocracies. In Bruges, beginning in 1241, government was in the hands of the merchants of the Flemish Hanse of London; and in Ghent the "39 aldermen," belonging to about 100 families, ruled from 1228.

These urban patricians (burghers) lived in fortified houses of stone and employed a considerable proletariat. They were so jealous of their autonomy that Count Philip of Alsace built the great Castle of the Counts in Ghent (beginning *c.* 1178) to maintain his authority there. The burghers built hospitals, such as the Byloke in Ghent (1228), and other public institutions for the welfare of the people.

Growth as commercial centres. Bruges became an international commercial centre; it kept its connections with the receding sea through canals and outports, such as Damme (1180) and Sluis (late 13th century). A yearly fair, dominated by the English wool trade, was established at Bruges in 1200, and relations with England became an important factor in the period that followed. Commerce, which had been carried out overseas by merchants from Bruges, was handled, beginning in the 13th century, by foreign merchants resident in the town. Ghent established a link to the sea through the 13th-century Lieve Canal. It became mainly an industrial centre; and its luxury cloth, produced from English wool, gained a European reputation.

Dominance over Flanders. Toward the end of the 13th century the existing equilibrium in Flanders was broken. Count Guy of Dampierre attacked the power of the urban patriciate—in 1280 he suspended the 39 aldermen in Ghent—and was himself attacked by King Philip IV the Fair, of France, who wanted to bring the country under his direct rule. Various social groups in the towns also turned against their local oligarchic rulers; the latter sided with France, whereas the count sought the support of the common people. At the Battle of the Golden Spurs (1302) a Flemish army, containing contingents from Bruges and Ghent, defeated the French. Ghent, Bruges, and Ypres came to dominate the county, and their old oligarchies lost supremacy.

The Battle of the Golden Spurs

The central authority was in jeopardy, and the county nearly split into three city-states. From 1323 to 1328 Bruges supported a peasants' revolt, which was suppressed with French help. Jacob van Artevelde, captain general of Ghent from 1338 to 1345, defying the pro-French Count Louis of Nevers, made a treaty with England to keep Flanders out of the Anglo-French Hundred Years' War and to safeguard the supply of English wool. Philip van Artevelde, son of Jacob, rose against Count Louis of Male in 1379 but was beaten by a French army at Roosebeke in 1382; after a prolonged resistance Ghent made peace with the new count, Philip the Bold, duke of Burgundy, at the Peace of Tournai in 1385.

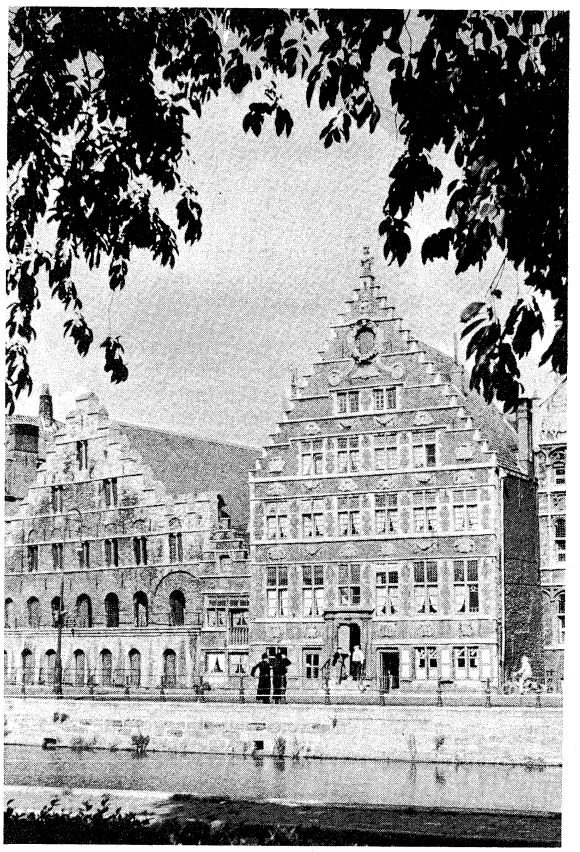
Decline of Ghent and Bruges. International commerce continued to flourish in Bruges. But the cloth industry in Ghent declined because of English and Italian competition; this led to economic hardship and social unrest. Rivalries between the craft guilds often turned into bloody street battles.

Under Burgundian rule. Under the counts of the House of Burgundy, central authority was re-established. Bruges, which accepted this more readily than did Ghent, became the artistic capital of Flanders through the patronage of the Burgundian court. In 1452 Ghent rose against Philip the Good but was beaten at the Battle of Gavere in 1453 and had to submit. After initial trouble Ghent accepted the stern rule of Charles the Bold, duke of Burgundy from 1467 to 1477, but rose, together with Bruges and other towns, in 1477 against his daughter Mary, who then revoked many centralizing measures. Both Ghent and Bruges revolted against her husband, Maximilian of Austria, particularly after Mary's death in 1482. In 1485 Bruges, which had imprisoned Maximilian, surrendered to his troops; but fresh rebellions broke out in 1487–89, 1490, and 1492.

The silting up of the Zwin estuary, the troubles under Maximilian, and the attraction of Antwerp sealed the decline of commercial Bruges in the 16th century; the town was gradually deserted by the "nations" of foreign merchants.

In 1539 Ghent rebelled against the emperor Charles V; but it was severely punished in 1540, when a new constitution was imposed. In 1566 iconoclasts destroyed many of Ghent's art treasures. From 1578 to 1584 the city was a Calvinist republic involved in a revolt against King Philip II of Spain. After it surrendered to the Spaniards under Alessandro Farnese in 1584, it lost its old vigour, partly through mass emigration of Protestant sympathizers. Bruges also fell to Farnese in 1584.

By courtesy of the Tourist Office of the city of Ghent, Belgium



Guild houses in the old port of Ghent. The house on the left dates from about 1200; the others are late 17th century.

In modern times. Ghent recovered during the 17th century, but the wars of Louis XIV (1670s) were disastrous for the city. It revived with the development of a modern cotton industry, modelled on that of England. A university, built in 1817, and the Ghent-Terneuzen Sea Canal, built in 1825–27, both under the initiative of King William I of the Netherlands, greatly contributed to industrial and intellectual development. During the 19th century Ghent played a great role in the Flemish movement—which tried to promote the Flemish language and culture. Bruges in the 19th century was an impoverished town, but its historic buildings exerted a romantic attraction. The Bruges-Zeebrugge Sea Canal was built in 1895–1907, but industrial development was slow. In Bruges, as in Ghent, there was considerable foreign investment in industry after World War II.

BIBLIOGRAPHY. A. DUCLOS, *Bruges: Histoire et souvenirs* (1910); M. LETTS, *Bruges and its Past*, 2nd ed. (1926); J.A. VAN HOUTTE, *Bruges: Essai d'histoire urbaine* (1967), and "The Rise and Decline of the Market of Bruges," in *Economic History Review*, 19:29–47 (1966), an authoritative study; R. DE ROOVER, *Money, Banking and Credit in Mediaeval Bruges* (1948), and *The Bruges Money Market Around 1400*, with a statistical supplement by H. SANDY (1968), important for financial history; J. MARECHAL, *Geschiedenis van de Brugse Beurs* (1949), the best work on the Bruges bourse; V. VERMEERSCH, *Brugges Kunstbezit* (1969), an excellent illustrated survey; F. DE POTTER, *Gent. van den oudsten tijd tot heden*, 8 vol. (1882–1901, 1933); V. FRIS, *Histoire de Gand* (1913, 2nd ed., 1930); F. BLOCKMANS, *Het Gentsche stadspatriciaat tot omstreeks 1302* (1938), a fundamental study of the social history to 1302; M.E. DUMONT, *Gent, een stedenaardrijkskundige studie*, 2 vol. (1951), a description of the modern city; H. PIRENNE, *Bibliographie de l'histoire de Belgique*, 3rd ed. (1931); V. FRIS, *Bibliographie de l'histoire de Gand*, 2 vol. (1907–21). Current bibliography for Bruges may be found in the yearly review in the *Handelingen v.d. Société d'Emulation;* and for Ghent and Bruges in the yearly review of works on Belgian history in the *Revue belge de philologie et d'histoire.*

(R.C.v.C.)

Ghiberti, Lorenzo

Lorenzo Ghiberti was the leading bronze sculptor in Florence in the first half of the 15th century. An early success and a recipient of valuable public commissions, Ghiberti was already a prominent artist of 40 years of age when he, Donatello, and Nanni di Banco each created a new sculptural style that formed the basis of most Italian sculpture of the Renaissance. Admiring antique art, Ghiberti aimed at making his figures lifelike but created sculptures with a personal mark of lyrical grace and the technical perfection learned in his early training as a goldsmith. Whether he was working in the earlier International Gothic style or in that of the Renaissance, Ghiberti's sense of balance and good taste was sought after and prized by both his contemporaries and the following generations.

Brogi—Alinari



Self-portrait from the "Gate of Paradise," bronze sculpture by Lorenzo Ghiberti, 1425–52. The Baptistery, Florence.

Early years and first commissions. Ghiberti was born around 1378. His mother had married Cione Ghiberti in 1370, and they lived in Pelago, near Florence; at some point she went to Florence and lived there as the common-law wife of a goldsmith named Bartolo di Michele. They were married in 1406 after Cione died, and it was in their home that Lorenzo Ghiberti grew up. It is not certain which man was Ghiberti's father, for he claimed each as his father at separate times. But throughout his early years, Lorenzo considered himself Bartolo's son, and it was Bartolo who trained the boy as a goldsmith. Lorenzo also received training as a painter; as he reported in the autobiographical part of his writings, he left Florence in 1400 with a painter to work in the town of Pesaro for its ruler, Sigismondo Malatesta.

Competition for the Florence Baptistery doors

Ghiberti returned quickly to his home city when he heard, in 1401, that a competition was being held for the commission to make a pair of bronze doors for the Baptistery of the cathedral of Florence. He and six other artists were given the task of representing the biblical scene of Abraham's sacrifice of Isaac in a bronze relief of quatrofoil shape. The entry panels of Ghiberti and of Filippo Brunelleschi are the sole survivors of the contest. In this first known youthful work, Ghiberti's art showed characteristics whose enormous appeal brought him suc-

cess: a graceful and lively composition executed with a mastery of the goldsmith's art that makes each superbly finished detail a pleasure to contemplate. In 1402 Ghiberti was chosen to make the doors by a large panel of judges; their decision brought Ghiberti, in his early 20s, the recognition and prominence that marked his entire career. The contract was signed in 1403 with Bartolo di Michele's workshop—overnight the most prestigious in Florence—and in 1407 Lorenzo legally took over the commission.

The work on the doors lasted until 1424, but Ghiberti did not devote himself to this alone. He created designs for the stained-glass windows in the cathedral; he regularly served as architectural consultant to the cathedral building supervisors, although it is unlikely that he actually collaborated with Brunelleschi on the construction of the dome as he later claimed. The Arte dei Mercanti di Calimala, the guild of the merchant bankers, gave him another commission, around 1412, to make a larger than life-size bronze statue of their patron saint, John the Baptist, for a niche on the outside of the guilds' common building, Or San Michele. The job was a bold undertaking, Ghiberti's first departure from goldsmith-scale work; it was, in fact, the first large bronze in Florence. Ghiberti successfully finished the "St. John" in 1416, adding gilding in the following year. The technical achievement and the modernity of its style brought Ghiberti commissions for two similarly large bronze figures for guild niches at Or San Michele: the "St. Matthew" in 1419 for the bankers' guild and the "St. Stephen" for the wool guild in 1425.

Niche figures at Or San Michele

These last two commissions brought Ghiberti into open competition with the newly prominent younger sculptors Donatello and Nanni di Banco, who had made stone statues for Or San Michele after Ghiberti's first figure there. The "St. John" was a frail figure enveloped by flowing draperies. It is characteristic of the style art historians call International Gothic, which swept Europe in the late 14th century and was quite new in Florence in the early 15th century. Ghiberti's "St. John" combined the soft draperies and closely observed, small-scale details in a sculpture larger than life. Donatello's "St. Mark" and "St. George" and Nanni di Banco's "St. Philip" and "Quattro Coronati" ("Four Crowned Saints") were as large as Ghiberti's figure but were designed with monumental proportions to match their scale; the boldness and strength of the weighty new classical figures constituted a challenge that Ghiberti met with success in his next sculptures, maintaining his pre-eminent position as a leading artist in Florence.

Increasing artistic success. The teens and '20s were years of flourishing expansion for Ghiberti and his firm. He had completed a great deal of the modelling and casting of the panels for the Baptistery doors by 1413, and he was in control of a smoothly functioning workshop with many assistants. In 1417 Ghiberti was asked to make two bronze reliefs for the baptismal font of the cathedral in Siena; he was so busy that he only finished them, under pressure from the Sienese authorities, ten years later. In 1419, when Pope Martin V was in Florence, Ghiberti was called on as a goldsmith to fashion a morse and mitre for the Pontiff; unfortunately these pieces, like other examples of Ghiberti's art in rare stones and precious metals, have disappeared. During these years, too, Lorenzo found a wife—Marsilia, the 16-year-old daughter of Bartolomeo di Luca, a wool carder. She soon bore him two sons: Tommaso was born in 1417 and Vittorio the next year; his sons later joined Ghiberti in his business, and Vittorio continued its operation after his father's death. Ghiberti's artistic success also had its financial rewards; a surviving tax return of 1427 lists property in Florence, land out of town, and a substantial amount of money invested in government bonds to his credit. Over the years, his real estate and monetary holdings continued to grow. In addition to being well paid, Ghiberti was a businessman who managed his affairs shrewdly. He was a well-to-do member of Florentine society and a rich man among the artists of his time.

Success of Ghiberti's workshop

Ghiberti was actively involved with and interested in other artists and their work; some (Donatello, Paolo Uccello, Michelozzo, Benozzo Gozzoli) had worked for a time in his workshop as young assistants. Ghiberti's association with the painter Fra Angelico is documented: Ghiberti designed the frame for his "Linaiuoli Altarpiece." In his commentaries, Ghiberti exaggerates only a bit when he proudly claims that "few important things were done in our city which were not devised or designed by my hand"; among his undocumented works may be noted some half-dozen floor tombs and sarcophagi, but the vast extent to which Ghiberti's providing of designs and models influenced Florentine art is hard to measure. He appears to have shared his knowledge and talent generously and freely; long before the completion of his second pair of doors (the "Gates of Paradise") in 1452, the fund of figures and models assembled in connection with this work, which the public saw only later, was open to painters of frescoes in the Chiostro Verde (green cloister) of SS. Annunziata and to the sculptor Luca della Robbia, working on a marble singing gallery for the cathedral. Naturally, the impact of the "Gates" increased after they were installed.

When he was 45 years old, Ghiberti finished the first doors. They are the effort of over 20 years of work and the major sculptural complex of the International Gothic style in Italy. They show some changes in the latest parts, however, to a more classical style that emphasizes the bodies of figures more than the elegant draperies that enfold them. Ghiberti created expressive, strong faces based on examples he knew of ancient Roman art—portrait busts and carved sarcophagi. Because of the success of the first doors, a contract was soon signed with the Calimala for a second pair, but the political and financial fortunes of the city and the guild did not permit work to get underway for about five years.

Development of humanistic theories and practices. Following the completion of the first doors, Ghiberti embarked on a decade of intense exploration of new ways of forming pictorial space and making gracefully active and lifelike figures. His works of the late 1420s show him able to make space increasingly intelligible in a series of clearly receding planes; using shallow relief, Ghiberti depicted volumes of bodies and deep architectural spaces. Examples of these are the reliefs in Siena; the Dati Tomb (the bronze plaque for the floor tomb of the Dominican general Leonardo Dati); and the two shrines in Florence, "Cassa di S. Zenobius" (a bronze casket with relief panels of stories from the saint's life) and "Shrine of SS. Protus, Hyacinth, and Nemesius" (a bronze container for the relics of three martyrs). It is likely that at this time Ghiberti encountered Leon Battista Alberti, a young Humanist scholar, who, inspired by the new art in Florence, was composing theoretical treatises on the visual arts. Their mutual belief that beauty was synonymous with the conception they shared of antique art makes it difficult to know whether or not Alberti's ideas in *De pictura* (*On Painting*) precede the three panels of the second door (Isaac, Joseph, and Solomon), which are the visual equivalent of those ideas. The beauty of antique art meant for both Alberti and Ghiberti an idealization of nature; capturing its essence meant revealing life by depicting movement, life's most salient visible characteristic. For the representation of a realistic spatial setting for these naturalistic figures, Alberti's treatise sets forth a perspective system for projecting such spaces onto the picture plane of a painting or bas relief. Ghiberti's three panels seem an embodiment of the Humanist's formulations for Renaissance pictorial art, and it is clear that any assessment of his art must account for the incorporation of the new theory as well as for the beauty and charm of these works. Ghiberti was himself so proud that he claimed to have made, in all ten panels,

Contact with Alberti and growing interest in antique art

> architectural settings in the relation with which the eye measures them, and real to such a degree that . . . one sees the figures which are near appear larger, and those that are far off smaller, as reality shows it.

Ghiberti's writings, *I Commentarii* (probably completed around 1447), shed more light on his Humanist interests.

The commentaries are composed of three books. The first, a history of art in ancient times, is Ghiberti's digest of writings of Latin authors he had read on the subject; in it he reveals his belief that the inseparability of practice and theory is responsible for the excellence of ancient art. The second book records the art of the immediate past, and Ghiberti expresses his admiration for certain Sienese painters and for a late 14th-century northern goldsmith named Gusmin who is known only through Ghiberti's pages; this book includes an autobiography, in which Ghiberti establishes his place in the history of art. The last book was apparently more theoretical, but in the surviving manuscript it is fragmentary. The commentaries demonstrate Ghiberti's confidence in his position as an important leader in the Florentine Renaissance—one interested in recapturing the art of the ancients, in studying that art as a Humanist scholar would, and one who developed a new style *all'antica* in which he freely created art works with a grace and beauty that have been found winning since their invention. Ghiberti died at Florence on December 1, 1455.

MAJOR WORKS

SCULPTURE: Competition relief for Baptistery doors, Florence (1401; Bargello, Florence); north door reliefs (1403–24; Baptistery, Florence); "St. John the Baptist" (1412–16; Or San Michele, Florence); reliefs showing St. John the Baptist before Herod and the Baptism of Christ (1417–27; Baptistery font, Siena); "St. Matthew" (1419–22; Or San Michele, Florence); "Tomb of Leonardo Dati" (1423–27; Sta. Maria Novella, Florence); "St. Stephen" (1425–28; Or San Michele, Florence); "Gates of Paradise," bronze door (1425–52; Baptistery, Florence); "Cassa di S. Zenobius" (1430s; Cathedral, Florence).

STAINED GLASS: "Assumption" (*c.* 1425; cathedral window, Florence).

BIBLIOGRAPHY. RICHARD KRAUTHEIMER, "Ghiberti and Master Gusmin," *Art Bulletin*, 29:25–35 (1947), a discussion of the literary and stylistic evidence for the connection of Ghiberti's early art with that of French goldsmiths; RICHARD KRAUTHEIMER and TRUDE KRAUTHEIMER-HESS, *Lorenzo Ghiberti* (1956, reprinted 1970 with new introduction), by far the most readable and comprehensive presentation and analysis of Ghiberti's life, art, and fame by eminent scholars (with bibliography); ULRICH MIDDELDORF, "Additions to Lorenzo Ghiberti's Work," *Burlington Magazine* 113:72–79 (1971), a study giving more examples of Ghiberti's widespread activity; JOHN POPE-HENNESSY, *An Introduction to Italian Sculpture*, 5 vol. (1955–63, reprinted 1970), a compendium of photographs, text, and scholarly notes on Ghiberti, his predecessors, and followers; FILIPPO ROSSI, "The Baptistery Doors in Florence," *Burlington Magazine* 89:334–341 (1947), details of the restoration of the doors to their present state; CHARLES SEYMOUR, *Sculpture in Italy 1400–1500* (1966), a study of Ghiberti in the general context of his century's sculptural art.

Ghiberti's writings: JULIUS VON SCHLOSSER, *Lorenzo Ghiberti's Denkwürdigkeiten* (*I Commentarii*), 2 vol. (1912), the only complete printing of the text of Ghiberti's Italian, with an analysis in German; excerpts in English translation appear in ELIZABETH GILMORE HOLT (ed.), *A Documentary History of Art*, 2nd ed. (1957); and ROBERT GOLDWATER and MARCO TREVES (eds.), *Artists on Art* (1945).

(C.Lo.)

Ghirlandajo, Domenico

A leading artist in Florence during the early Renaissance, Domenico Ghirlandajo was an extremely skillful fresco painter. His specialty was large and clearly organized frescoes with a strong narrative content and usually a large number of portraits.

Domenico di Tommaso Bigordi was born in Florence in 1449, the son of a goldsmith. The nickname "Ghirlandajo" (or "Ghirlandaio") is derived from his father's skill in making garlands. Domenico probably began as an apprentice in his father's shop, but almost nothing is known about his training as a painter or the beginnings of his career. The earliest works attributed to him, dating from the early 1470s, show strong influence from the frescoes of Andrea del Castagno, who died when Ghirlandajo was about eight years old. Giorgio Vasari, the biographer of Renaissance artists, recorded in his *Lives* (1550) that Ghirlandajo was a pupil of the Florentine painter Alesso Baldovinetti, but Baldovinetti was only four or five years older than Ghirlandajo himself. He worked in fresco on large wall surfaces in preference to smaller scale paintings executed on wood panels, although he used them for the altarpieces that were the centrepieces of the fresco cycles in his major undertakings. He never experimented with oil painting, although most Florentine painters of his generation began to use it exclusively in the last quarter of the 15th century.

Early works and influences

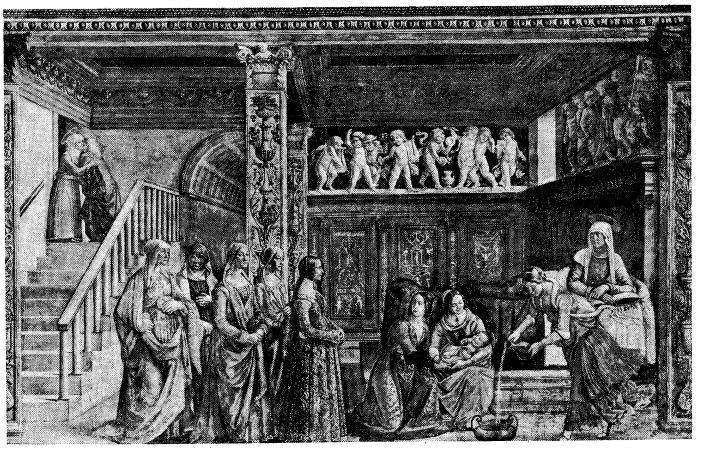
The village church of Cercina, near Florence, has a fresco of three saints, now thought to be Ghirlandajo's earliest work, but there is general agreement that some frescoes in the Church of Ognissinti in Florence, almost certainly dating from around 1472–73, show his style at its earliest developed stage. One of them represents the "Pietà" and depicts several members of the Vespucci family as mourners, thus already introducing Ghirlandajo's characteristic combination of portrait figures in contemporary dress with a specifically religious subject. Something of the passion for minute detail shown by the early Flemish painters can be found in Ghirlandajo's work at this period; his fresco "St. Jerome," also in Ognissanti and dated 1480, may even be an enlarged version in fresco of an oil painting by the Flemish painter Jan van Eyck, which had found its way to Florence. The "St. Jerome" fresco is particularly important because it is a companion piece to one of "St. Augustine" by Botticelli; the difference between the two reveals Ghirlandajo's rather pedestrian and anecdotal style.

The first major commissioned works were the frescoes the "Life of Sta. Fina," painted in 1475 in the church at San Gimignano, near Florence, which derive from Fra Filippo Lippi's slightly earlier fresco cycle in the cathedral at Prato and contain a number of portrait heads arranged, rather stiffly, in the symmetrical type of composition that was to become increasingly identified with him. Even then he was already employing assistants; in his later works he clearly could only complete large commissions in the comparatively short time allotted by the extensive use of highly trained assistants working simultaneously on different parts of the frescoes. In 1481–82 he received an important commission in the Vatican for a fresco, nominally representing the calling of the first Apostles, in the Sistine Chapel. Its style is reminiscent of the frescoes by Masaccio of about 1427, which had been the great innovating works of the early 15th century in Florence but by then must have seemed somewhat old-fashioned. The principal feature of this fresco is the group of portraits of the Florentine colony in Rome, who are represented as witnesses of the biblical event. It has been suggested that the inclusion of these Florentines in a fresco painted for the Vatican had political significance, because the Florentine government had recently accused Pope Sixtus IV of complicity in the conspiracy of the Pazzi, another powerful Tuscan banking family, to murder the leading members of the Florentine Medici family.

Study of Roman antiquities

Ghirlandajo must have used his stay in Rome to study Roman antiquities at first hand, for many details of triumphal arches, ancient sarcophagi, and similar antique elements occur in his works throughout the rest of his career. A sketchbook filled with drawings of such antiquities (now in El Escorial, near Madrid) seems to be the work of a member of his shop.

Late in his short life, Ghirlandajo and his assistants, including his brothers Davide and Benedetto, produced two major fresco cycles. The earlier, a series of frescoes and an altarpiece painted in tempera, was executed for the Sassetti Chapel in Sta. Trinità in Florence. Commissioned by Francesco Sassetti, an agent of the Medici bank, they were executed between about 1482 and 1485. They represent scenes from the life of St. Francis of Assisi, Sassetti's patron saint. Once more, the frescoes contain many details of the buildings and customs of the period—for example, the original front of the church of Sta. Trinità itself—and, in particular, there are numerous portraits of members of the Sassetti family shown together with some of the leading members of the Medici family, in what may appear to have been a closer intimacy than was actually the case. The altarpiece, dated



"The Birth of the Virgin" (with a portrait of Giovanna Tornabuoni, daughter of the patron of the series, followed by other Florentine patrician ladies in contemporary dress), one of the fresco series by Domenico Ghirlandajo, 1485–94. In Sta. Maria Novella, Florence.
Brogi—Alinari

1485, contains further evidence of Ghirlandajo's interest in classical antiquity, for it shows the "Adoration of the Shepherds" with a Roman triumphal arch in the background and a Roman sarcophagus in place of the traditional manger. This painting in tempera has several direct references to contemporary Flemish paintings, especially the enormous altarpiece painted in oil by Hugo van der Goes, which had been commissioned in Flanders by Tommaso Portinari, another agent of the Medici bank, and which arrived in Florence in the late 1470s.

Greatest fresco cycle

Ghirlandajo's last and greatest fresco cycle was painted for another Medici banker, Giovanni Tornabuoni, and represents scenes from the life of the Virgin and of St. John the Baptist. Ghirlandajo signed the contract on September 1, 1485, for these very large frescoes on the walls of the choir of Sta. Maria Novella in Florence. The altarpiece was still incomplete when he died in Florence on January 11, 1494, but his assistants, among whom was very probably the boy Michelangelo, had completed the frescoes by about 1490. Even more than in the Sassetti Chapel the narrative scenes contain a wealth of detail showing patrician interiors and contemporary dress; as a result they are one of the most important sources for current knowledge of the furnishings of a late-15th-century Florentine palace.

The frescoes in Sta. Maria Novella are overcrowded with detail, so that the compositions fail to make their full impact. Some of his smaller panel paintings, particularly the portrait of Giovanna Tornabuoni, have a simplicity that makes them far more striking than the frescoes. This portrait of Giovanna, who died in 1490, must have been used for the portrait of her in the frescoes of Sta. Maria Novella. The portrait representing an old man with a strawberry nose with his grandchild (Louvre, Paris) is perhaps Ghirlandajo's finest painting, notable for its tenderness and humanity, as well as a simplicity and directness of handling.

The vicissitudes of his reputation have been considerable. In the late 19th century, because of the high degree of realism in his work, he was ranked as a leading Florentine painter of the 15th century, but during much of the 20th century the greater imaginative power of Botticelli or Filippino Lippi made Ghirlandajo's paintings seem dull. Since the 1960s the honesty and truth of his works have brought him back into critical favour, although, significantly enough, in his lifetime Ghirlandajo never received a major commission from the Medici family, or from any other leading patrons.

MAJOR WORKS

"SS. Barbara, Jerome and Anthony Abbot" (early 1470s; Cercina, Italy); frescoes over Vespucci Altar (*c.* 1472–73; Church of Ognissanti, Florence); "Life of Sta. Fina" (1475; Chapel of Sta. Fina, Collegiata, San Gimignano); "Madonna and Child Enthroned, with SS. Michael, Raphael, Justus, and Zenobius" (*c.* 1480; Uffizi, Florence); "Last Supper" (1480; Church of Ognissanti, Florence); "St Jerome" (1480; Church of Ognissanti, Florence); "Christ Calling SS. Peter and Andrew" (1481–82; Sistine Chapel, Vatican, Rome); "Roman Heroes" (1482; Palazzo Vecchio, Sala dei Gigli, Florence); "Adoration of the Shepherds" (1485; Sta. Trinità, Sassetti Chapel, Florence); "Madonna and Saints," formerly the Sta. Maria Novella altarpiece (1485–94, unfinished; mostly in Alte Pinakothek, Munich); scenes from the lives of the Virgin and St. John the Baptist (1486–90; Choir, Sta. Maria Novella, Florence); "Adoration of the Magi" (1488; Ospedale degli Innocenti, Florence); "Giovanna Tornabuoni" (1488; Thyssen-Bornesmisza Collection, Castagnola, Switz.); "The Visitation" (1491; Louvre, Paris).

BIBLIOGRAPHY. The only modern monograph is JAN LAUTS, *Domenico Ghirlandajo* (1943). The principal source is VASARI'S *Vite* . . . (1550; 2nd ed., 1568); several English translations exist. See also the article by GIUSEPPE MARCHINI in the *Encyclopedia of World Art*, vol. 6, col. 319–325 (1962), with full bibliography; and the characterization of Ghirlandajo's art by FREDERICK HARTT in *A History of Italian Renaissance Art*, pp. 303–308 (1969).

(P.J.Mu.)

Giacometti, Alberto

Alberto Giacometti, Swiss painter and sculptor who lived mostly in Paris, was one of the outstanding artists of the 20th century. At a time when avant-garde artists aimed at rendering nonfigurative or expressive qualities rather than achieving resemblance to reality, he worked for the unattainable goal of equalling reality by rendering a portrait—whether drawing, painting, or sculpture—so that it would be perceived by the spectator with the impact it would have were it a living person. To do this, he introduced into the art of sculpture a new concept of rendering distance. Massless and weightless, his figures and heads are immediately seen from a specific frontal point of view and therefore perceived as situated in distance and space. Giacometti had such intellectual integrity—for example, living in a shabby studio in Montparnasse even after fame and fortune had reached him—that he became for his contemporaries, especially those of the postwar generation, an almost legendary figure during his lifetime.

Born on October 10, 1901, at Borgonovo in the Bergell Valley, Switzerland, Giacometti displayed precocious talent and was much encouraged by his father, Giovanni, a Postimpressionist painter, and by his godfather, Cuno Amiet, a Fauvist painter. He spent a happy childhood in the nearby village of Stampa, to which he returned regularly until his death. His brother Diego became known as a furniture designer and shared Giacometti's life as his model and aid. Another brother, Bruno, became an architect.

Giacometti left secondary school in Schiers in 1919 and then went to Geneva, where he attended art classes during the winter of 1919–20. After a time in Venice and Padova (May 1920) he went to Florence and Rome (fall 1920–summer 1921), where rich collections of Egyptian

Stay in Italy

art taught him that the impact of ancient and primitive hieratic styles—which adhere to fixed, conventional types and frontal or rigid figures—could be used as an equivalent for the force of reality.

Between 1922 and 1925, Giacometti studied at the Académie de la Grande-Chaumière in Paris. Although he owed much to his teacher, Émile-Antoine Bourdelle, his style was very different. It was related to the Cubist sculpture of Alexander Archipenko and Raymond Duchamp-Villon and to the Post-Cubist sculpture of Henri Laurens and Jacques Lipchitz. An example is "Torso" (1925). He was also inspired by African and Oceanic art, as in "The Spoon-Woman" (1926). His first important personal achievements were flat, slablike sculptures, such as "Observing Head" (1927/28), which soon made him popular among the Paris avant-garde.

©Karsh—Rapho Guillumette



Giacometti, photograph by Yousuf Karsh, 1965.

Any resemblance to reality had been abandoned in the period 1925–29, when he created mannered figures, such as "Cubist Composition" (1926) and "Three Figures Outdoors" (1929). The trend continued in the period 1930–32, in works in which emotions and erotic themes were given Surrealist sculptural form ("Suspended Ball" and "The Palace at 4 A.M."). In 1933–34 Giacometti attempted metaphorical compositions using the themes of life and death ("The Invisible Object" and "1 + 1 = 3"). At this time he was disturbed by the thought that his serious works of art had as little reference to reality as the merely decorative vases and lamps that he made to earn a living. Breaking definitely with the Surrealist group in 1935, he began to work after nature again; what had started as mere studies became a lifelong adventure and a turning point in art history: the phenomenological approach to reality—that is, the search for the given reality in what one sees when one is looking at a person.

Break with Surrealism

Around 1940, Giacometti arrived at matchstick-sized sculptures: figures and heads seen frontally as ungraspable appearances of reality far away in space. Around 1947, his massless, weightless image of reality was expressed in a skeletal style, with figures thin as beanstalks. From 1947 to 1950 he did compositions related to his work of the early 1930s—"Tall Figures"; "City Square"; "Composition with Seven Figures and a Head (The Forest)"; and "Chariot"—and became rapidly known, especially in the United States, through two exhibitions (1948 and 1950) at the Pierre Matisse Gallery in New York and an essay on his art by the French Existentialist writer Jean-Paul Sartre.

The evolution of his art continued, taking the form of a search for ways to challenge, actually to equal, reality in sculpture as well as in painting. An artwork was to become an almost magical evocation of reality in an imaginary space, as in heads of Diego and figures after his wife Annette (1952–58), executed like apparitions on gray canvases or on space-delimiting bases. Finally, the artwork was invested with the power of acting on the spectator like a double of reality in real space, as in portraits of Caroline or Elie Lotar, his models and friends in the last years (1958–65), which are heads and busts gazing intently and made only with lines of force, without contour lines or surfaces. At this point, the phenomenological approach was superseded; he felt that reality is no longer dependent on being perceived by someone; reality simply is. Like the characters of Samuel Beckett's novels and plays (Giacometti had done the set for *Waiting for Godot* in 1963), his figures represent a view of man in which space and time have their origin in the core of each being.

Figures like Beckett characters

Alberto Giacometti died on January 11, 1966, in Chur (Switzerland) of an inflammatory heart condition, without having carried out the final composition of the work he had been concerned with since the early 1930s, the metaphor of the totality of life.

MAJOR WORKS

SCULPTURE: Many of these works exist in editions of more than one cast and may therefore be seen in more than one collection. The Kunsthaus in Zürich and the Galerie Beyeler in Basel have the most comprehensive collections of Giacometti's work on loan from the Alberto Giacometti Foundation. Other important collections are in the Museum of Modern Art, New York and in the Fondation Maeght, Saint-Paul, France. "Torso" (1925); "The Couple" (1926); "The Spoon-Woman" (1926); "Cubist Composition" (1926); "Observing Head" (1927–28); "Man" (1929); "Reclining Woman Who Dreams" (1929); "Three Figures Outdoors" (1929); "Suspended Ball" (1930–31); "Disagreeable Object" (1931); "Hand Caught by a Finger" (1932); "Woman with Her Throat Cut" (1932); "The Palace at 4 A.M." (1932–33 Museum of Modern Art, New York); "No More Play" (1933); "Cubist Head" (1934–35); "Nude" (1933–34); "The Invisible Object" (1934); "1 + 1 = 3" (1934); "Woman with the Chariot I" (1942–43); "Man Pointing" (1947); "Head of a Man on a Rod" (1947); "Tall Figure I" (1947); "City Square" (1948); "Three Men Walking" (1949); "Walking Quickly Under the Rain" (1949); "Tall Figure" (1949); "Composition with Seven Figures and a Head (The Forest)" (1950); "Between Two Houses" (1950); "The Cage" (1950–51); "Four Women on a Base" (1950); "Chariot" (1950); "Dog" (1951); "Head" (1952); "Woman" (1953); "Nine Standing Figures (Venice)" (1956); "Tall Figures" (1960); "Walking Man I" (1960); "Monumental Head" (1960); "Bust of Annette" (1962); "Figure Standing" (1964); "Head of Diego" (1965); "Bust of Elie Lotar" (1965).

PAINTINGS: "Self-Portrait" (1921; Kunsthaus, Zürich); "The Artist's Mother" (1950; Museum of Modern Art, New York); "Caroline" (1961; private collection).

LITHOGRAPHS: "Paris sansfin" (1958–65).

BIBLIOGRAPHY. The most comprehensive collection of Giacometti's work is in the Alberto Giacometti-Foundation, Zürich, with fully illustrated catalog (1971). A comprehensive and systematic bibliography is in the monograph by Hohl (below). Bibliographies are also in exhibition catalogs, *Alberto Giacometti,* Museum of Modern Art, New York (1965); and Musée de l'Orangerie, Paris (1969); and in JAMES LORD, *Alberto Giacometti: Drawings* (1971). The most important autobiographical statements are the "Letters to Pierre Matisse" in the exhibition catalogs of the Pierre Matisse Gallery, New York (1948 and 1950); CARLTON LAKE, "The Wisdom of Giacometti (Interview)" in *The Atlantic Monthly,* 216:117–126 (1965); and JAMES LORD, *A Giacometti Portrait* (1965), day-by-day notes of 18 sittings with firsthand material on Giacometti's life and way of working. Many autobiographical quotations are translated into English in the monograph by Hohl. Literary articles on Giacometti include: JEAN-PAUL SARTRE, "The Search for the Absolute," in catalog *Giacometti,* Matisse Gallery, New York (1948; reprinted in *Situations III,* 1949), fundamental for the existentialist interpretation of his work; and "The Paintings of Giacometti," in catalog *Giacometti,* Maeght Gallery, Paris (1951; reprinted in *Art and Artist,* 1956); JEAN GENET, *L'Atelier d'Alberto Giacometti* (1958), fascinating personal views; JACQUES DUPIN, *Alberto Giacometti* (1962), enormously helpful, with 236 reproductions; and LUIGI CARLUCCIO, *Alberto Giacometti: le copie del passato* (1967; Eng. trans., *Giacometti: A Sketchbook of Interpretive Drawings,* 1967), which contains 144 drawings after other masters and one of the last autobiographical texts.

Some important critical studies are DAVID SYLVESTER, "Perpetuating the Transient," in catalog *Giacometti*, Arts Council, London (1955); and "The Residue of a Vision," in catalog *Giacometti*, Tate Gallery, London (1965); HILTON KRAMER, "Giacometti," *Arts Magazine*, 38:52–59 (1963), very valuable article against the existentialist interpretation of Giacometti's work; HERBERT LUST, *Alberto Giacometti: The Complete Graphics* (1971), useful for the reproductions; JAMES LORD, *Alberto Giacometti: Drawings* (1971), considerate introduction, chronology, and bibliography; and REINHOLD HOHL, *Alberto Giacometti* (1972), the most comprehensive monograph, with many quotations, a special dates and documents section, a systematic bibliography, and 196 plates.

(Re.H.)

Gibbon, Edward

Unquestionably the greatest English historian of the 18th century, Edward Gibbon combined enormous erudition with the philosophical ideals of the Enlightenment to produce the first history in English that achieved both sound scholarship and broad philosophical scope. His outstanding work, *The History of the Decline and Fall of the Roman Empire*, presents a continuous narrative from the 2nd century AD to the fall of Constantinople in 1453. A masterpiece of prose style, it has since been corrected in detail but never entirely supplanted; it continues to influence men's judgment of antiquity and the Middle Ages.



Gibbon, oil painting by Henry Walton, 1774.
In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

Gibbon was born at Putney, Surrey, on April 27, 1737. After the reform of the calendar in 1752, his birthday fell on May 8. His grandfather, Edward, had made a considerable fortune and his father, also Edward, was able to live an easygoing life in society and Parliament. He married Judith, a daughter of James Porten (pronounced Porteen), whose family had originated in Germany. Edward, too, had independent means throughout his life. He was the eldest and the only survivor of seven children, the rest dying in infancy.

Early life. His own childhood was a series of illnesses and more than once he nearly died. Neglected by his mother, he owed his life to her sister, Catherine Porten, whom he also called "the mother of his mind," and after his mother's death in 1747 he was almost entirely in his aunt's care. He early became an omnivorous reader and could indulge his tastes the more fully since his schooling was most irregular. He attended a day school in Putney and, in 1746, Kingston grammar school, where he was to note in his *Memoirs* "at the expense of many tears and some blood, [he] purchased a knowledge of Latin syntax." In 1749 he was admitted to Westminster School. He was taken in 1750 to Bath and Winchester in search of health and after an unsuccessful attempt to return to Westminster was placed for the next two years with tutors from whom he learned little. His father took him on visits to country houses where he had the run of libraries filled with old folios. He noted his 12th year as one of great intellectual development and says in his *Memoirs* that he had early discovered his "proper food," history. By his 14th year he had already covered the main fields of his subsequent masterpiece, applying his mind as well to difficult problems of chronology. The keynote of these early years of study was self-sufficiency. Apart from his aunt's initial guidance, Gibbon followed his intellectual bent in solitary independence. This characteristic remained with him throughout his life. His great work was composed without consulting other scholars and is impressed with the seal of his unique personality.

Intellectual independence

In his *Memoirs* Gibbon remarked that with the onset of puberty his health suddenly improved and remained excellent throughout his life. Never a strong or active man, he was of diminutive stature and very slightly built and he became corpulent in later years. The improvement in his health apparently accounts for his father's sudden decision to enter him at Magdalen College, Oxford, on April 3, 1752, about three weeks before his 15th birthday. He was now privileged and independent. Any expectations of study at Oxford were soon disappointed. The authorities failed to look after him intellectually or spiritually or even to note his absences from the college. Left to himself, Gibbon turned to theology and read himself into the Roman Catholic faith. It was a purely intellectual conversion. Yet he acted on it and was received into the Roman Catholic Church by a priest in London on June 8, 1753.

Exile. His father, outraged because under the existing laws his son had disqualified himself for all public service and office, acted swiftly, and Edward was dispatched to Lausanne and lodged with a Calvinist minister, the Rev. Daniel Pavillard. Though the change was complete, and Gibbon was under strict surveillance, in great discomfort, and with the scantiest allowance, he later spoke of this period with gratitude. To Pavillard he owed kindly and competent instruction and the formation of regular habits of study. He mastered the bulk of classical Latin literature and studied mathematics and logic. He also became perfectly conversant with the language and literature of France, which exercised a permanent influence on him. These studies made him not only a man of considerable learning but a stylist for life. He began his first work, written in French, *Essai sur l'étude de la littérature* (*An Essay on the Study of Literature*). Meanwhile, the main purpose of his exile had not been neglected. Not without weighty thought, Gibbon at last abjured his new faith and was publicly readmitted to the Protestant communion at Christmas 1754. "It was here," Gibbon says somewhat ambiguously, "that I suspended my religious enquiries, acquiescing with implicit belief in the tenets and mysteries which are adopted by the general consent of Catholics and Protestants."

Further education at Lausanne

In the latter part of his exile Gibbon entered more freely into Lausanne society. He attended Voltaire's parties. He formed an enduring friendship with a young Swiss, Georges Deyverdun, and also fell in love with and rashly plighted himself to Suzanne Curchod, a pastor's daughter of great charm and intelligence. In 1758 his father called Gibbon home shortly before his 21st birthday and settled an annuity of £300 on him. On the other hand, he found that his father and his second wife were implacably opposed to his engagement, and he was compelled to break it off. ("I sighed as a lover, I obeyed as a son.") He never again thought seriously of marriage. After a natural estrangement he and Mlle Curchod became lifelong friends. She was well known as the wife of Jacques Necker, the French finance minister under Louis XVI. During the next five years Gibbon read widely and considered many possible subjects for a historical composition. He published his *Essai* first in French (1761) and later in English (1764). From 1760 until the end of 1762, his studies were seriously interrupted by his service on home defense duties with the Hampshire militia. With the rank of captain he did his duty conscientiously and later claimed that his experience of men and camps had been useful to him as a historian.

Choosing a subject. Gibbon left England on January 25, 1763, and spent some time in Paris, making the ac-

quaintance of several Philosophes, Denis Diderot and Jean Le Rond d'Alembert among others. During the autumn and winter spent in study and gaiety at Lausanne, he gained a valuable friend in John Baker Holroyd (later Lord Sheffield), who was to become his literary executor.

Visit to Rome

In 1764 Gibbon went to Rome, where he made an exhaustive study of the antiquities and, on October 15, 1764, while musing amid the ruins of the Capitol, was inspired to write of the decline and fall of the city. Some time was yet to pass before he decided on the history of the empire.

At home, the next five years were the least satisfactory in Gibbon's life. He was dependent on his father and although nearly 30 had achieved little in life. Although bent on writing a history, he had not settled on a definite subject. Impressed by the supremacy of French culture in Europe, he began in that language a history of the liberty of the Swiss but was dissuaded from continuing it. He and Deyverdun published two volumes of *Mémoires littéraires de la Grande Bretagne* (1768–69). In 1770 he sought to attract some attention by publishing *Critical Observations on the Sixth Book of the Aeneid.*

His father died intestate in 1770. After two years of tiresome business, Gibbon was established in Bentinck Street, London, and concentrated on his Roman history. At the same time he entered fully into social life. He joined the fashionable clubs and was also becoming known among men of letters. In 1775 he was elected to the Club, the brilliant circle that the painter Sir Joshua Reynolds had formed round the writer and lexicographer Dr. Samuel Johnson. Although Johnson's biographer, James Boswell, openly detested Gibbon, and it may be inferred that Johnson disliked him, Gibbon took an active part in the Club and became intimate with Reynolds and the actor David Garrick. In the previous year he had entered Parliament and was an assiduous, though silent, supporter of Lord North.

The "Decline and Fall." The first quarto volume of his history, published on February 17, 1776, immediately scored a success that was resounding, if somewhat scandalous because of the last two chapters in which he dealt with great irony with the rise of Christianity.

Critical reception to his great work

Reactions to Gibbon's treatment of Christianity have displayed various phases. Both in his lifetime and after, he was attacked and personally ridiculed by those who feared that his skepticism would shake the existing establishment. In the 19th century he was hailed as a champion by militant agnostics. Gibbon himself was not militant. He did not cry with Voltaire, "*Écrasez l'Infâme!*" ("Crush the Infamy!") because in his England and Switzerland he saw no danger in the ecclesiastical systems. His concern was past history. One may say, however, with confidence, that he had no belief in a divine revelation and little sympathy with those who had such a belief. While he treated the supernatural with irony, his main purpose was to establish the principle that religions must be treated as phenomena of human experience. In this his successors have followed him and added to the collateral causes of Christianity's growth those that he had overlooked or could not know of, such as the various mystery religions of the empire and particularly the Mithraic cult. Although Gibbon's best known treatment of Christianity is found mainly in the 15th and 16th chapters, no less significant are later chapters in which he traced the developments of theology and ecclesiasticism in relation to the breakup of the empire.

Gibbon went on to prepare the next volumes. Meanwhile, he was assailed by many pamphleteers and subjected to much ridicule. His ugliness and elaborate clothes made him an easy target. For the most part he ignored his critics. The historians David Hume and William Robertson recognized him as their equal if not their superior. Only to those who had accused him of falsifying his evidence did he make a devastating reply in *A Vindication of Some Passages in the Fifteenth and Sixteenth Chapters of the Decline and Fall of the Roman Empire* (1779).

In the same year he obtained a valuable sinecure as a commissioner of trade and plantations. Shortly after that he composed *Mémoire justificatif* (1779), a masterly state paper in reply to continental criticism of the British government's policy in America. In 1781 he published the second and third volumes of his history, bringing the narrative down to the end of the empire in the West. Gibbon paused at this point to consider continuing his history. In 1782, however, Lord North's government fell, and soon Gibbon's commission was abolished. This was a serious loss of income. To economize he left England and joined Deyverdun in a house at Lausanne. There he quietly completed his history in three more volumes, writing the last lines of it on June 27, 1787. He soon returned to England with the manuscript, and these volumes were published on his 51st birthday, May 8, 1788. The completion of this great work was acclaimed on all sides.

The Decline and Fall is thus comprised of two divisions, equal in bulk but inevitably different in treatment. The first half covers a period of about 300 years to the end of the empire in the West, about AD 480. In the second half nearly 1,000 years are compressed. Yet the work is a coherent whole by virtue of its conception of the Roman Empire as a single entity throughout its long and diversified course.

Unity of the work

Gibbon imposed a further unity on his narrative by viewing it as an undeviating decline from those ideals of political and, even more, intellectual freedom that he had found in classical literature. The material decay that had inspired him in Rome was the effect and symbol of moral decadence. However well this attitude suited the history of the West, its continuance constitutes the most serious defect of the second half of Gibbon's history and involved him in obvious contradictions. He asserted, for example, that the long story of empire in the East is one of continuous decay, yet for 1,000 years Constantinople stood as a bulwark of eastern Europe. The fact is that Gibbon was not only out of sympathy with Byzantine civilization; he was less at home with Greek sources than with Latin and had no access to vast stores of material in other languages that subsequent scholars have assembled. Consequently there are serious omissions in his narrative, as well as unsatisfactory summaries.

Nevertheless this second half contains much of Gibbon's best. With all its shortcomings, it marshals with masterly lucidity the successive forces that eventually overthrew Constantinople. Many of his most famous chapters occur there. These include sections on Justinian, the Trinitarian controversies, the rise of Islām, and the history of Roman law. There is, in addition, a brilliant and moving story of the last siege and capture of Constantinople and, finally, the epilogue of chapters describing medieval and Renaissance Rome, which gives some hope that the long decline is over and that mankind has some prospect of recovering intellectual freedom. The vindication of intellectual freedom is a large part of Gibbon's purpose as a historian. When toward the end of his work he remarks, "I have described the triumph of barbarism and religion," he reveals epigrammatically his view of the causes of the decay of the Greco-Roman world. They can hardly be disputed. But there is the further question of whether the changes brought about are to be regarded as ones of progress or retrogression. Writing as a mid-18th-century "philosopher," Gibbon saw the process as retrogression, and his judgment remains of perpetual interest.

Last years

Returning to Lausanne, Gibbon turned mainly to writing his memoirs. His happiness was broken first by Deyverdun's death in 1789, quickly followed by the outbreak of the French Revolution and the subsequent apprehension of an invasion of Switzerland. He had now become very fat and his health was declining. In 1793 he suddenly returned to England on hearing of Lady Sheffield's death. The journey aggravated his ailments and he died in a house in St. James's Street, London, on January 16, 1794. His remains were placed in Lord Sheffield's family vault in Fletching Church, Sussex.

Assessment. Modern knowledge of history, in Gibbon's field alone, has increased conspicuously. Economic, social, and constitutional history have grown up. The

study of coins, inscriptions, and archaeology generally has brought in a great harvest. Above all, the scientific examination of literary sources, so rigorously practiced now, was unknown to Gibbon. Yet he often exhibits a flair and an acumen that seem to anticipate these systematic studies. He had genius in large measure, as well as untiring industry and accuracy in consulting his sources. Though he was unsympathetic to Christianity, his sense of fairness and probity made him respectful of honest opinion and true devotion, even among those with whom he disagreed. These qualities, expressed with his command of historical perspective and his incomparable literary style, justify a modern historian's dictum that, "whatever else is read Gibbon must be read too," or the conclusion of the great Cambridge historian J.B. Bury:

> That Gibbon is behind date in many details and in some departments of importance, simply signifies that we and our fathers have not lived in an absolutely incompetent world. But in the main things he is still our master above and beyond "date."

BIBLIOGRAPHY

Works: The Decline and Fall, original ed., vol. 1 (1776), vol. 2–3 (1781), vol. 4–6 (1788). Numerous later editions, especially W. SMITH'S reissue of DEAN MILMAN'S ed. with GUIZOT'S notes, 8 vol. (1854 and 1872), and J.B. BURY'S ed., 7 vol. (1896–1900 and 1909–14), with valuable introduction and commentaries—abridged in 1 vol. by D.M. LOW (1960); *Miscellaneous Works*, ed. by LORD SHEFFIELD, 2 vol. (1796), 5 vol. (1814 and 1815) containing the *Memoirs*, a large selection of Gibbon's letters, and a number received by him from others; *Memoirs of My Life and Writings*, reissues of Sheffield's edition by G. BIRKBECK HILL (1900) and J.B. BURY (1907, 1923)—definitive edition by GEORGES A. BONNARD (1969); *Manuscript Drafts*, ed. by JOHN MURRAY (1896); *Letters*, ed. by R.E. PROTHERO, 2 vol. (1896); *Correspondence*, minor works and unpublished material, ed. by J.E. NORTON, 3 vol. (1956), the fullest collection of all the letters known at that date—invaluable both for its information and style.

Journals: Gibbon's Journal to January 28, 1763, ed. by D.M. LOW (1929), contains an account of Gibbon's service in the militia and contemporary social life and reading, preceded by a brief relation of his early years; *Le Journal de Gibbon à Lausanne, 1763–1764* (1945) and *Gibbon's Journey from Geneva to Rome 1764* (1961), both ed. by G.A. BONNARD.

Bibliography: H.M. BEATTY in vol. 7 of J.B. BURY'S ed. of *The Decline and Fall* (1914); JANE E. NORTON, *A Bibliography of the Works of Edward Gibbon* (1940).

Biography and criticism: J.C. MORRISON, *Gibbon* (1878); G.M. YOUNG, *Gibbon* (1932); D.M. LOW, *Edward Gibbon, 1737–1794* (1937); G. GIARIZZO, *Edward Gibbon e la cultura europea del settecento* (1954); E.J. OLIVER, *Gibbon and Rome* (1958); HAROLD L. BOND, *The Literary Art of Edward Gibbon* (1960); SIR GAVIN DE BEER, *Gibbon and His World* (1968).

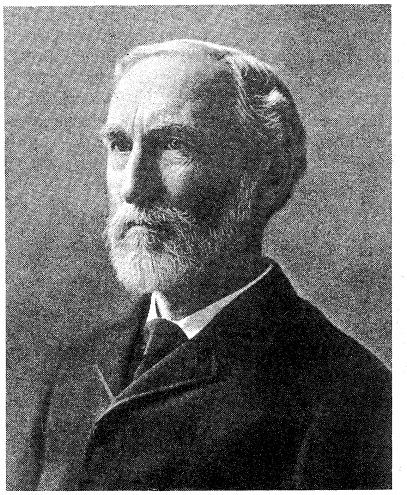
(D.M.Lo.)

Gibbs, J. Willard

Josiah Willard Gibbs was a theoretical physicist and chemist, considered by many in the latter part of the 20th century to be the greatest scientist yet produced by the United States. His application of thermodynamic theory converted a large part of physical chemistry from an empirical into a deductive science. His work also enabled the quantities involved in many chemical industrial processes to be forecast precisely by calculation, thus greatly increasing efficiency and economy.

Early years

Gibbs was born at New Haven, Connecticut, on February 11, 1839, the fourth child and only son of Josiah Willard Gibbs, Sr., professor of sacred literature at Yale. There were college presidents among his ancestors and scientific ability in his mother's family. Facially and mentally, Gibbs resembled his mother. He was a friendly youth but withdrawn and intellectually absorbed. This circumstance and his delicate health kept him from participating much in student and social life. He was educated at the local Hopkins Grammar School and entered Yale in 1854, where he won a succession of prizes. After graduating, he pursued research in engineering. His thesis on the design of gearing was distinguished by the logical rigour with which he used geometrical methods of analysis. In 1863 Gibbs received the first doctorate of



Gibbs.
By courtesy of Yale University

engineering to be conferred in the United States. He was appointed a tutor at Yale in the same year. He devoted some attention to engineering invention.

Gibbs lost his parents rather early, and he and his two older sisters inherited the family home and a modest fortune. In 1866 they went to Europe, remaining there nearly three years while Gibbs attended the lectures of European masters of mathematics and physics, whose intellectual technique he assimilated. He returned more a European than a U.S. scientist in spirit—one of the reasons why general recognition in his native country came so slowly. He applied his increasing command of theory to the improvement of James Watt's steam-engine governor. In analyzing its equilibrium, he began to develop the method by which the equilibriums of chemical processes could be calculated.

He was appointed professor of mathematical physics at Yale in 1871, before he had published his fundamental work. His first major paper was "Graphical Methods in the Thermodynamics of Fluids," published in 1873. It was followed in the same year by "A Method of Geometrical Representation of the Thermodynamic Properties of Substances by Means of Surfaces" and in 1876 by his most famous paper, "On the Equilibrium of Heterogeneous Substances." The importance of his work was immediately recognized by the Scottish physicist James Clerk Maxwell in England, who made a model of Gibbs's thermodynamic surface with his own hands and sent it to him.

He remained a bachelor, living in his surviving sister's household. In his later years he was a tall, dignified gentleman, with a healthy stride and ruddy complexion, performing his share of household chores, approachable and kind (if unintelligible) to students. He died at New Haven on April 28, 1903, at the age of 64.

Ignorance of his work

Gibbs was highly esteemed by his friends, but U.S. science was too preoccupied with practical questions to make much use of his profound theoretical work during his lifetime. He lived out his quiet life at Yale, deeply admired by a few able students but making no immediate impress on U.S. science commensurate with his genius. He never even became a member of the American Physical Society. He seems to have been unaffected by this. He was aware of the significance of what he had done and was content to let posterity appraise him.

The contemporary historian Henry Adams called Gibbs "the greatest of Americans, judged by his rank in science." His application of thermodynamics to physical processes led him to develop the science of statistical mechanics; his treatment of it was so general that it was later found to apply as well to quantum mechanics as to the classical physics from which it had been derived.

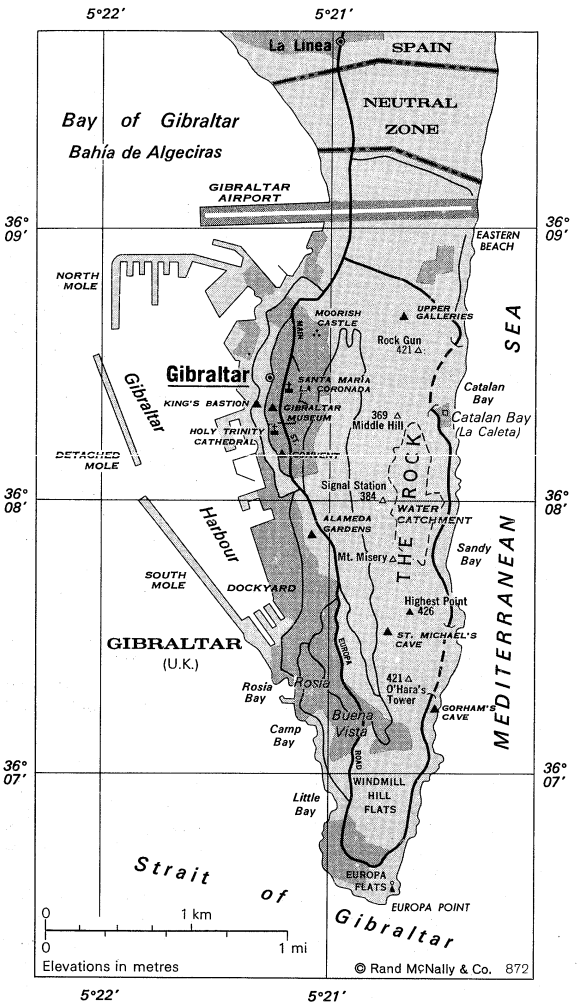
BIBLIOGRAPHY. The most considerable biography yet available is LYNDE PHELPS WHEELER, *Josiah Willard Gibbs* (1970), containing an extensive bibliography. One of the best personal accounts of Gibbs is E.B. WILSON, "Reminiscences of

Gibbs by a Student and Colleague," *Scient. Mon.*, 32:211–227 (1931). A biography for the general reader is contained in J.G. CROWTHER, *Famous American Men of Science* (1969). An appraisal of his work is F.G. DONNAN and A.E. HAAS (eds.), *A Commentary on the Scientific Writings of J. Willard Gibbs* (1936). His papers have been brought together as *The Collected Works of J. Willard Gibbs*, ed. by W.R. LONGLEY and R.G. VAN NAME, with a biographical sketch by H.A. BUMSTEAD (1928).

(J.G.Cr.)

Gibraltar

The British colony of Gibraltar occupies a narrow peninsula of Spain's southern Mediterranean coast, just northeast of the Strait of Gibraltar. Three miles long and three-quarters of a mile wide, it is connected to Spain by a low, sandy isthmus that is one mile long. Its name is derived from the Arabic Jabal Ṭāriq (Mt. Tarik), honouring Ṭāriq ibn Ziyād, who captured the peninsula in AD 711. The colony's total area is 2.25 square miles (5.8 square kilometres).



GIBRALTAR

Gibraltar was possibly one of the "Pillars of Hercules" that defined the limits of navigation for the ancient Mediterranean world. Since the 18th century, it has been a symbol of British naval strength, commonly known as "the Rock." With the opening of the Suez Canal in 1869, Gibraltar increased in strategic importance, and its position as a provisioning port was greatly enhanced. Since World War II the British military garrison and naval dockyard have continued to be an important part of Gibraltar's economy, and naval operations of the North Atlantic Treaty Organization (NATO) often use the port facilities. Attempts are being made, however, to diversify the colony's economy through its commercial port and increased tourist facilities. Gibraltar's historic heritage is reflected in the social composition and unique cultural traditions of the colony and also, less happily, in the disputes between Britain and Spain over its political status.

The natural environment. *Relief.* The peninsula consists of a limestone and shale ridge known as the Rock. It rises abruptly from the isthmus to 1,380 feet (421 metres) at Rock Gun, which is its northernmost summit. Its greatest height, 1,396 feet, is attained near its southern end. The Rock shelves down to the sea at Europa Point, which faces Ceuta (a Spanish *plaza* in Morocco), 14 miles across the strait. From the Mediterranean, Gibraltar appears as a series of sheer, inaccessible cliffs, fronting the sea, that are broken by huge sand slopes above Catalan and Sandy bays. The slope is more gradual on its western side and is occupied by tier upon tier of houses that stretch for some 300 feet above the old defensive walls. Above the houses, the steep, stony hillside is baked brown in summer and bursts into abundant vegetation after the first autumn rains. Higher up limestone cliffs almost isolate the Upper Rock, which is covered with a tangle of wild trees. The Rock

Water supply. Gibraltar has no springs or rivers, and 34 acres of the eastern sand slopes have been sheeted over to provide a rain-catchment area. The water is stored in 13 tanks (with a total capacity of 16,000,000 gallons) that were blasted into the Rock. The rainwater is then blended with water pumped from wells on the isthmus or distilled from the sea. When these sources threaten to run short, water is imported from northern Europe. Seawater is supplied for sanitation purposes.

Soils. Most of the peninsula's soil has been formed by the disintegration of limestone. There is red sand beneath and to the south of the city, and fawn-coloured sand slopes cover over half of the eastern side of the Rock. The eastern sands are believed to have been windblown from the Sahara.

Vegetation and animal life. The Gibraltar candytuft is a flower native only to the Rock. There are more than 500 species of small flowering plants, and wild olive and pine trees grow on the Upper Rock. Mammals include rabbits, foxes, and Barbary apes. The only wild monkeys in Europe, the apes have roamed the Rock for hundreds of years and were long a symbol of British presence in Gibraltar. Although free to wander at will, they are generally to be seen on the Upper Rock. Migratory birds are common, and Gibraltar is the home of the only specimens of Barbary partridge in Europe. The Barbary apes

The human imprint. *Settlement.* The old walled city of Gibraltar stands on the western face of the Rock. Because of the sheer face of the cliffs above 300 feet, the town has been forced to spread onto the isthmus to the north and southward toward Europa Point, and some 100 acres have been reclaimed from the sea. There is also a small village at Catalan Bay.

The people. The 1970 population total of 27,965 included 2,783 aliens and 1,003 visitors but excluded British servicemen. Only Gibraltarians—those born in Gibraltar before 1925 and their descendants—have the right to live in the colony. All others must obtain residence permits. Most Gibraltarians are of mixed Genoese, British, Spanish, Maltese, and Portuguese descent. The alien community includes Indian shopkeepers and their families and workers from Morocco.

The majority of the population belongs to the Roman Catholic Church, which established a bishopric in Gibraltar in 1910. The Anglican bishopric, created in 1842, also covers communities in southern Europe. The small Jewish community is of Sefardic descent. The birth rate averages 20 per 1,000 and the mortality rate eight per 1,000 over the course of a decade. Spanish is the language of most homes, although most Gibraltarians are bilingual in English and Spanish.

The economy. Because of lack of space, there is no agriculture. There is a small amount of light industry—tobacco, beverages, canning—but the main sources of income are tourism, provisioning of ships and military personnel, and the re-export trade. Since the 1960s, large-scale expansion of hotel and beach facilities has been funded largely by loans from the United Kingdom. The

port facilities occupy most of the western shore and a portion of land reclaimed from the sea. The naval dockyard and auxiliary installations face a harbour which is formed by three moles, parts of which are now used commercially. There is also a small commercial repair yard and yacht marina. Of the re-export commodities, fuel for visiting ships is the most important, while import duties and the price of water are maintained at a low level to encourage provisioning.

The port

Customs and income tax produce most of the colony's revenue. Aid from Britain has been received since Spain started its restrictions; before that Gibraltar balanced its own budget. Principal expenditures include medical services, education, and housing.

Transportation. Spain closed its road to Gibraltar in 1969, restricting access to the colony to sea and air service. Passenger and cargo vessels stop at the port, and a car ferry crosses daily to Tangier, Morocco. There are regular flights linking Gibraltar to London and Tangier. Over 5,000 vehicles use the 35 miles of roads and some of the 30 miles of tunnels within the Rock. A cable car ascends to the central summit of the ridge.

Administration and social conditions. *Government.* Gibraltar is a colony of the United Kingdom that is self-governing in all matters but defense. Gibraltarians of both sexes over 18 and British civilians resident for over a year are entitled to vote every four years to elect a House of Asssembly of 15 members. A government is formed from the party or coalition of parties gaining the majority of seats, with a chief minister presiding over a Council of Ministers. Instead of a city council, one minister is responsible for municipal affairs. The judicial system includes a Court of Appeal, the Supreme Court, the Court of First Instance, and the Magistrates' Court.

Defense forces

In addition to the United Kingdom forces on the Rock, the volunteer Gibraltar Regiment serves as an infantry and coast artillery unit. The police force is responsible for immigration, residential permits, ambulance service, and traffic control in addition to other police duties.

Health and welfare. A compulsory-insurance scheme for industrial workers provides benefits during unemployment and old age or after an accident. The government maintains a general hospital and one for mental illnesses. Payments are scaled to the patient's income or are abolished for members of lower income groups. The Royal Naval Hospital serves visiting warships and the garrison.

Education. Education is compulsory between the ages of five and 15. All instruction is in English and includes religious training. Fifteen primary and seven secondary schools, all tuition-free, accept Gibraltarians and the children of British service personnel. There are also private primary schools, a technical college, and a special commercial school for girls.

The media. There are two English-language daily newspapers, an English–Spanish biweekly, and a bilingual weekly. The Gibraltar Broadcasting Corporation runs television and radio broadcasts. Similar services from Spain reach many homes.

Cultural life. Art exhibitions are frequent, amateur drama thrives, and enthusiastic groups engage in cave exploration, bird watching, and archaeological research. The Gibraltar Museum houses natural-history, archaeology, and military collections, and the British Council has a periodical and record library. Football, hockey, cricket, yachting, and fishing are popular sports.

Prospects for the future. Gibraltar's future revolves around the question of its political status. Spain claims the right to sovereignty over the peninsula, the United Kingdom wishes to retain sovereignty, and Gibraltar desires the closest links with Britain. Tension has been high since the mid-1960s, when, in a series of moves, the Spanish government withdrew all Spanish workers from Gibraltar, closed the frontier, attempted to restrict air and sea traffic, and terminated telegraph and telephone services with and through Spain. In the plebiscite of 1967 Gibraltar voted by over 99 percent to remain with the United Kingdom, and by the early 1970s no solution to the problem had been reached.

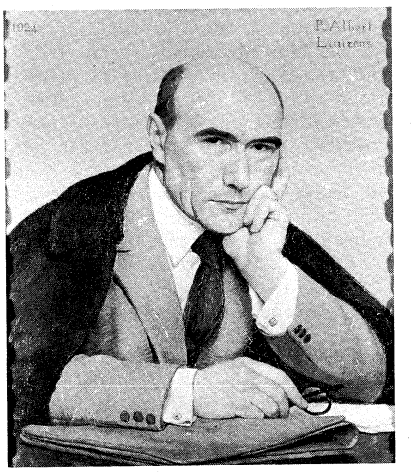
BIBLIOGRAPHY. ALLEN ANDREWS, *Proud Fortress: The Fighting Story of Gibraltar* (1958); DOROTHY ELLICOTT, *From Rooke to Nelson: 101 Eventful Years in Gibraltar, 1704–1805* (1965), historical text from the naval angle, and *Bastion Against Aggression: How Gibraltar Helped Spain During the Peninsular War* (1968); B.H.T. FRERE, *A Guide to the Flora of Gibraltar* (1910), standard work; H.W. HOWES, *The Gibraltarian: The Origin and Development of the Population of Gibraltar from 1704* (1950); E.R. KENYON, *Gibraltar Under Moor, Spaniard and Briton*, rev. ed. (1938), history with special reference to existing antiquities; JACK RUSSELL, *Gibraltar Besieged, 1779–1783* (1965), recent research into the Great Siege.

(D.M.E.)

Gide, André

For most of his life a controversial figure, André Gide, one of the most notable writers of his time, was also one of its outstanding personalities. Long regarded as a revolutionary for his open support of the claims of individual liberty of action in defiance of conventional morality, before his death he was widely recognized as first and foremost a moralist, interested rather in humanity than in individuals. It is for his integrity and nobility of thought, his purity and harmony of style, that, as a moralist in the great 17th-century French tradition, he takes his place among the masters of French literature.

Giraudon, © A.D.A.G.P. 1970



Gide, oil painting by P.A. Laurens, 1924. In the National Museum of Modern Art, Paris.

Heritage and youth. Born in Paris on November 22, 1869, Gide was the only child of Paul Gide and his wife, Juliette Rondeaux. His father was of southern Huguenot peasant stock; his mother, a Norman heiress, although Protestant by upbringing, belonged to a northern Catholic family long established at Rouen. To this mixed heredity Gide was to attribute the conflict in his character between asceticism and hedonism, spirituality and sensuality.

When he was eight he was sent to the École Alsacienne in Paris, but his education was much interrupted by ill health. After his father's early death in 1880, his well-being became his mother's chief concern, and, often kept at home, he was taught by indifferent tutors and by his mother's governess. Most of his holidays were spent in Normandy: in the summer he and his mother stayed at La Roque, where she had inherited a castle dating back to the 16th century. Gide used it as a setting for two of his tales, *The Immoralist* and *Isabelle*. His closest friends were his cousins Jeanne, Valentine, and Madeleine Rondeaux, with whom he stayed in Rouen and at their country house at Cuverville, a village near the coast. A pleasant 18th-century house, it is the setting of *Strait Is the Gate*, Gide's most classically perfect work.

The turning point of Gide's life

At Rouen, when he was 13, Gide underwent the emotional experience that changed his life. One day he came upon his 15-year-old cousin Madeleine weeping as she knelt at prayer. She told him that she had discovered her mother's infidelity and knew that she must keep it secret.

Gide fell in love with her and realized that he had found the object of his lifelong devotion. The scene provides a pivot for the action of *Strait Is the Gate* and is described in his autobiography, *If It Die* . . . .

Gide returned to the École Alsacienne to prepare for his *baccalauréat* examination, and, after passing it in 1889, he decided to spend his life in writing, music, and travel. Secretly he had determined to enshrine his devotion to Madeleine in literary form and so to win her love. His first work, however, the autobiographical tale *The Notebooks of André Walter*, published anonymously at his own expense, pleased neither Madeleine nor the public. Written, like most of his later works, in the first person, it uses the confessional form in which he was to achieve his greatest successes.

Symbolist period. In the same year a school friend, the writer Pierre Louÿs, introduced him to the poet Stéphane Mallarmé's famous "Tuesday evenings," the centre of the French Symbolist movement, and for a time he was influenced by its aesthetic theories. His next work, "Narcissus" (1891), exemplifies the poetic purpose of the symbol in its story of Narcissus, who, seeking to know himself by gazing at his reflection in the River of Time, learns that truth and perfection can only be apprehended imaginatively through symbols. Two "tales" published in 1893 also belong to Gide's Symbolist period: *Urien's Voyage*, in which he first reveals a characteristic irony and satiric power; and "The Lovers' Attempt," illustrating the difference between "dispassionate love" and desire.

In 1893 Gide paid the first of many visits to North Africa, hoping to find release from dissatisfaction with the restrictions imposed by his upbringing. There, during recovery from a serious illness, he composed the paean of praise to life in all its fullness that was to form part of *Fruits of the Earth*. But after he returned to France, his relief at having shed the shackles of convention evaporated in what he calls the "stifling atmosphere" of the Paris salons. Had he been unable to satirize this in *Marshlands*—a brilliant parable of animals who, living always in dark caves, lose their sight because they never use it—he would, he claimed, have been driven to the verge of suicide.

Marshlands was completed in 1894, and in a resurgence of the mood in which he had begun *Fruits of the Earth*, Gide returned to North Africa. There he met Oscar Wilde and Lord Alfred Douglas, who encouraged him to admit the nature of his suppressed homosexual desires and to defy traditional morality. With a new freedom and serenity, he composed the greater part of *Fruits of the Earth*, a work in which he appeals to the young to escape from convention, to "act without pausing to consider whether the action is good or evil; love without troubling whether it is good or evil." But before it was finished, he was recalled to France because of his mother's illness. She died in May 1895. Grief shattered the joy he had found in Africa, and he abandoned *Fruits of the Earth* to write *Saül*, a dramatic work diametrically opposed to the former in intent.

Mayor of La Roque

In October he married his cousin Madeleine Rondeaux, who had earlier refused him. Early in 1896 he was elected mayor of the commune of La Roque—at 27 the youngest mayor in France. He took his duties seriously but managed to complete *Fruits of the Earth*. It was published in 1897 and fell completely flat, although after World War I it was to become Gide's most popular and influential work. In the postwar generation, its call to each individual to express fully whatever in him is most completely his own evoked an immediate response, and it remains the one of Gide's works most enjoyed by the young.

Great creative period. *Prometheus Misbound*, a return to the satirical style of *Urien's Voyage* and *Marshland*, is Gide's last discussion of man's search for individual values. His next tales, with which his great creative period begins—*The Immoralist*, *Strait Is the Gate*, and "The Pastoral Symphony"—reflect his attempts to achieve harmony in his marriage in their treatment of the problems of human relationships. They mark an important stage in his development: adapting treatment and style to his concern with psychological problems. In them he first achieves a mastery of classical construction and a pure, simple style.

During most of this period Gide was suffering deep anxiety and distress. Although his love for Madeleine had given his life what he called its "mystic orientation," he found himself unable, in a close, permanent relationship, to reconcile this love with his need for freedom and for experience of every kind. Floundering in doubts and scruples, he tried to find in religion a test of the ultimate values and a source of inner peace and security. But here again, certainty evaded him; although, from *Strait Is the Gate* onward, his works were widely read and many praised them highly, they were also severely criticized for immorality.

The Vatican Swindle marks the transition to the second phase of Gide's great creative period. He called it not a tale but a *sotie*, by which he meant a satirical work in which all the characters are fools or madmen. In it he explores the various ways in which all men are fools. This, the first of his works to be violently attacked for anticlericalism, lost him many Catholic friends, among them the writer Paul Claudel; paradoxically, when writing this work, Gide had come nearest to embracing Catholicism.

Literary critic

In the early 1900s he had already begun to be widely known as a literary critic, and in 1908 he was foremost among those who founded *La Nouvelle Revue Française*, the literary review that was to unite progressive French writers until World War II.

During World War I Gide worked in Paris, first for the Red Cross, then in a soldiers' convalescent home, and finally as assistant director of the Foyer Franco-Belge, giving shelter to refugees. In 1916 he returned to Cuverville, his home since his marriage, and began to write again.

The war had intensified Gide's anguish, and early in 1916 he had begun to keep a second *Journal* (published in 1926 as *Numquid et tu?*) in which he recorded his search for God. Finally, however, unable to resolve the dilemma (expressed in his statement "Catholicism is inadmissible, Protestantism is intolerable; and I feel profoundly Christian"), he resolved to achieve his own ethic, and by casting off his sense of guilt to become his true self. Now, in a desire to liquidate the past, he began his autobiography, *If It Die* . . . , an account of his life from birth to marriage that is among the great works of confessional literature. In 1918 his friendship for the young Marc Allégret caused a serious crisis in his marriage, when his wife in jealous despair destroyed her "dearest possession on earth"—his letters to her.

After the war a great change took place in Gide, and his face began to assume the serene expression of his later years. By the decision involved in beginning his autobiography and the completion in 1918 of *Corydon* (a Platonic dialogue in defense of homosexuality begun earlier), he had achieved at last an inner reconciliation. In 1923 the birth of his only child Catherine, whose mother, Elisabeth van Bysselberghe, was the feminist daughter of a lifelong friend, gave him new cause for hope. *Corydon*'s publication in 1924 was disastrous; he was violently attacked, even by his closest friends. Although he felt their betrayal keenly, it was less wounding than his own sense of self-betrayal for his long hesitation before publishing the book.

Gide called his next work, *The Coiners*, his only novel. He meant by this that in conception, range, and scope it was on a vaster scale than his tales or his *soties*. It is the most complex and intricately constructed of his works, in which, he said, he aimed to achieve in fiction what Bach achieved in music in his *Art of the Fugue*. Its structure is fugal; themes rise, are submerged, and rise again in new patterns and formations; and the numerous plots and characters are linked with a musician's skill by a "main theme"—a character who is involved in all the novel's actions. Like *The Vatican Swindle*, *The Coiners* is transitional: in it are gathered up all of Gide's favourite themes, and it looks forward to the compassionate, objective concern for humanity of his final phase. Gide's awareness of the novel's difficulties is shown by his pub-

lication of the diary he had kept while writing it, *The Logbook of the Coiners.*

Journey to Africa

In 1925 Gide set off for French Equatorial Africa. He had planned a long absence to recover from the hostility aroused by *Corydon*. When he returned he published *Travels in the Congo*, in which he criticized French colonial policy. His journey had released him from personal obsessions and had given him energy to attack the world's problems. He became the champion of social outcasts, demanding more humane conditions for criminals and equality for women; and condemning the exploitation of Africans under colonial rule and the lot of the underprivileged.

For a time it seemed to him that he had found a faith in Communism. In 1936, full of hope, he set out on a visit to the Soviet Union; his disillusionment is expressed in *Return from the U.S.S.R.* and *Afterthoughts on the U.S.S.R.*

Late works. In his *Journal* for 1938 a thick black line is ruled across the pages between the entries for April and for August as a symbol of mourning for the death of his wife. After a long estrangement they had been brought together by her fears for his safety in Africa and by her final illness. To him she was always the great—perhaps the only—love of his life. With the outbreak of World War II, Gide began to realize the value of tradition and to appreciate the past. In a series of imaginary interviews written in 1941 and 1942 for *Le Figaro*, he expressed a new concept of liberty, declaring that absolute freedom destroys both the individual and society: freedom must be linked with the discipline of tradition.

Gide's new concept of liberty

From 1942 until the end of the war Gide lived in North Africa. There he wrote "Theseus," his last great work and literary testament. In it, through the meaning he gives to the story of Theseus, he symbolizes his realization of the value of the past: Theseus returns to Ariadne only because he has clung to the thread of tradition. In Theseus' final statement of faith, Gide's own voice can be heard:

> It is without regret that I draw near to my solitary death. I have enjoyed the benefits of the earth. I am glad to think that after me, and thanks to me, men will find themselves happier, better and more free. For the good of humanity to come I have accomplished my work. I have lived.

In June 1947 Gide received the first honour of his life: the Doctor of Letters of the University of Oxford. It was followed in November by the Nobel Prize for Literature. In 1950 he published the last volume of his *Journal*, which took the record of his life up to his 80th birthday in 1949. All Gide's writings illuminate some aspect of his complex character. He is seen at his most characteristic, however, in the *Journal* he kept from 1889, a unique work of more than a million words in which he records his experiences, impressions, interests, and moral crises during a period of more than 60 years. Here the reader meets a man of rare humility and simplicity with an unusually acute psychological perception. In its depth of self-knowledge, its humanity, and its perfection of style, his *Journal* is Gide's lasting memorial. After its publication he resolved to write no more; but he recorded his random thoughts in a notebook, published after his death as *So Be It, or, The Chips Are Down*. More significant perhaps was his letter, written a month before his death, to a Japanese victim of the Nagasaki atomic bomb who had asked him what attitude a man should take on the eve of the conquest of the world by conformism and machines. Only the individual, with his hatred of falsehood, Gide answered, offered hope of salvation: "We are like unto one who, to light his way, follows a torch that he himself is carrying."

The *Journal*

Gide died on February 19, 1951, and was buried beside his wife in the country cemetery at Cuverville.

MAJOR WORKS

COLLECTIONS: *Oeuvres complètes*, 15 vol. (1932–39; index, 1954), still awaiting completion but including items unpublished elsewhere; *Romans, récits et soties, oeuvres lyriques* (1958), collected fiction, with commentary and textual notes; *Théâtre* (1947; *My Theater*, 1951); *The André Gide Reader* (1971).

VERSE AND PROSE POETRY: *Les Cahiers d'André Walter* (1891; *The Notebooks of André Walter*, 1968); *Le Traité du Narcisse* (1891; "Narcissus," in *The Return of the Prodigal*, 1953); *Les Poésies d'André Walter* (1892); *La Tentative amoureuse* (1893, "The Lovers' Attempt," in *The Return of the Prodigal*, 1953); *Le Voyage d'Urien* (1893; *Urien's Voyage*, 1964); *Les Nourritures terrestres* (1897; *Fruits of the Earth*, 1949); *El Hadj* (1899; Eng. trans. in *The Return of the Prodigal*, 1953); *Amyntas* (1906; Eng. trans., 1958); *Le Retour de l'enfant prodigue* (1907; *The Return of the Prodigal*, 1953); *Les Nouvelles Nourritures* (1935; *New Fruits of the Earth*, 1949).

STORIES, SATIRES, AND FABLES: *Paludes* (1895; *Marshlands*, 1953); *Le Prométhée mal enchaîné* (1899; *Prometheus Misbound*, 1953); *L'Immoraliste* (1902; *The Immoralist*, 1930); *La Porte étroite* (1909; *Strait Is the Gate*, 1924); *Isabelle* (1911; Eng. trans. in *Two Symphonies*, 1931); *Les Caves du Vatican* (1914; *The Vatican Swindle*, 1925; *Lafcadio's Adventures*, 1927); *La Symphonie pastorale* (1919; "The Pastoral Symphony," in *Two Symphonies*, 1931); *L'École des femmes* (1929; *The School for Wives*, 1950); *Robert* (1929; "Robert," in *The School for Wives*, 1950); *Geneviève* (1936; "Geneviève; or the Unfinished Confidence," in *The School for Wives*, 1950); *Thésée* (1946; "Theseus," in *Two Legends: Oedipus and Theseus*, 1950).

NOVEL: *Les Faux-Monnayeurs* (1926; *The Counterfeiters*, 1927; also as *The Coiners*).

DRAMA: *Philoctète* (1899; "Philoctetes," in *The Return of the Prodigal*, 1953); *Le Roi Candaule* (1901; "King Candaules," in *My Theater*, 1951); *Saül* (1903; Eng. trans. in *The Return of the Prodigal*, 1953); *Bethsabé* (1912; "Bathsheba," in *The Return of the Prodigal*, 1953); *Oedipe* (1931; "Oedipus," in *Two Legends: Oedipus and Theseus*, 1950); *Persephone* (1934; Eng. trans. in *My Theater*, 1951); *Le Treizième arbre* (1935); *Robert, ou l'intérêt général* (1944–45); *Le Retour* (1946); *Les Caves du Vatican* (1950).

CRITICISM: *Prétextes* (1903; *Pretexts*, 1959); *Nouveaux prétextes* (1911); *Dostoïevsky* (1923; Eng. trans., 1925); *Incidences* (1924); *Le Journal des Faux-Monnayeurs* (1926; *The Logbook of the Coiners*, 1952); *Essai sur Montaigne* (1929; *Montaigne*, 1929); *Divers* (1931); *Interviews imaginaires* (1943; *Imaginary Interviews*, 1944); *Attendu que . . .* (1943); *L'Enseignement de Poussin* (1945; *Poussin*, 1947); *Poétique* (1947); *Préfaces* (1948); *Rencontres* (1948); *Éloges* (1948); *Notes sur Chopin* (1948; *Notes on Chopin*, 1949).

TRAVEL: *Voyage au Congo* (1927), and *Le Retour du Tchad* (1928; *Travels in the Congo*, 1929); *Dindiki* (1927); *Retour de l'U.R.S.S.* (1936; *Return from the U.S.S.R.*, 1937); *Retouches à mon retour de l'U.R.S.S.* (1937; *Afterthoughts on the U.S.S.R.*, 1937).

JOURNAL: *Journal, 1889–1939* (1939); *Journal, 1939–49* (1954), including most other autobiographical works; *Journals of André Gide*, 4 vol. (1947–51).

LETTERS: *Correspondence* with Francis Jammes (1948), Paul Claudel (1949), Charles du Bos (1950), Paul Valéry (1955), François Mauriac (1971).

OTHER WORKS: *Souvenirs de la Cour d'Assises* (1914; *Recollections of the Assize Court*, 1941); *Morceaux choisis* (1921); *Corydon* (1924; Eng. trans., 1950); *Si le grain ne meurt* (1926; *If It Die . . .*, 1935); *Numquid et tu . . . ?* (1926; Eng. trans. in *Journals*, vol. 2, 1948); *Un Esprit non prévenu* (1929); *L'Affaire Redureau* (1930); *La Séquestrée de Poitiers* (1930); *Jeunesse* (1945); *Et Nunc Manet in Te* (1947, 1951; Eng. trans., 1952; U.S. title, *Madeleine*); *Feuillets d'automne* (1949; *Autumn Leaves*, 1950); *Littérature engagée* (1950); *Ainsi soit-il, ou, les jeux sont faits* (1952; *So Be It, or, The Chips Are Down*, 1959).

TRANSLATIONS: Gide's works have been translated in most western and various non-western languages, and nearly all are available in English. Specially noteworthy are the fine English translations of most of his major works by his friend Dorothy Bussy, and of his *Journals* by Justin O'Brien.

BIBLIOGRAPHY. ARNOLD NAVILLE, *Bibliographie des écrits d'André Gide depuis 1891 jusqu'en 1952* (1962), is a complete bibliography of Gide's writings. The Bibliothèque Jacques Doucet, Paris, possesses unpublished manuscripts and correspondence. Other manuscript material and letters are in the Bibliothèque Nationale, Paris. The Association des Amis d'André Gide publishes *Cahiers André Gide* (1969–).

Biography: PIERRE DE BOISDEFFRE, *Vie d'André Gide* 2 vol. (1970–71), the standard full-scale biography; JEAN DELAY, *La Jeunesse d'André Gide*, 2 vol. (1956–57; abridged Eng. trans., *The Youth of André Gide*, 1963), a sound, important psychoanalytical study, with new documentation; JEAN LAMBERT, *Gide familier* (1958), a study of Gide in his old age, by his son-in-law; CLAUDE MAHIAS, *La Vie d'André Gide* (1955), a pictorial biography; GEORGE D. PAINTER, *André*

Gide: A Critical Biography (1968), a comprehensive short study that provides a good introduction to Gide's life and work; JEAN SCHLUMBERGER, *Madeleine et André Gide* (1956), on Gide's marriage.

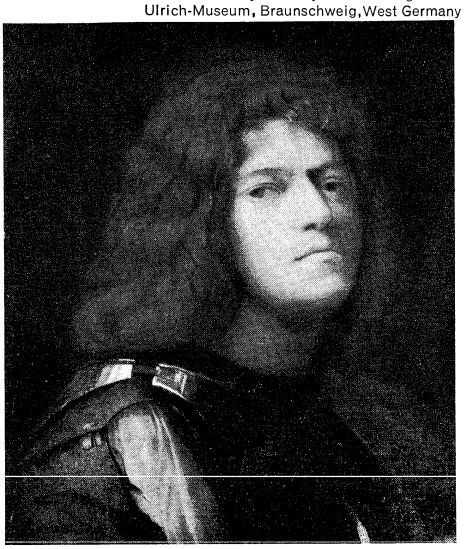
Criticism: GERMAINE BREE, *Gide* (1963); WALLACE FOWLIE, *André Gide: His Life and Art* (1965), an interesting account of the development of Gide's thought; ALBERT J. GUERARD, *André Gide*, 2nd ed. (1969), a critical study by a specialist on Gide; JUSTIN O'BRIEN, *Portrait of André Gide* (1963); ENID STARKIE, *André Gide* (1953), a sympathetic brief guide based on a long friendship with Gide.

(En.S.)

Giorgione

The works of Giorgione of Castelfranco mark the transition from the early Renaissance to the High Renaissance style. Despite the brevity of 34 years of life, his poetic and imaginative art left its mark in the first quarter of the 16th century upon virtually all of the Venetian painters who were his contemporaries.

By courtesy of the Herzog Anton Ulrich-Museum, Braunschweig, West Germany



Giorgione, self-portrait, oil painting. In the Herzog Anton Ulrich-Museum, Braunschweig, West Germany.

Giorgione was probably born in 1477 in Castelfranco, north of Venice. Nothing is known about his personal life except the legends reported by the biographer and Mannerist artist Giorgio Vasari in the two editions (1550 and 1568) of his *Vite de' più eccellenti pittori, scultori ed architettori italiani . . .* (*Lives of the Most Eminent Italian Painters, Sculptors and Architects . . .*). His name is given in two surviving documents of 1507 and 1508 as Zorzi da Castelfranco (in Venetian dialect); *i.e.*, Giorgio of Castelfranco. The form Giorgione (or Zorzon), which is customarily used today, first appears in the 1528 inventory of the Grimani Collection. His name means tall or big George, implying that he was a large man. Tradition holds that he was handsome and amorous. Correspondence dated October 25, 1510, between the celebrated Renaissance patroness of the arts Isabella d'Este of Mantua and her agent Taddeo Albano at Venice mentions Giorgione's death as having occurred recently, probably caused by the plague that was raging in Venice at that time. Vasari's biography is the earliest. It emphasizes the artist's humble origin, his elevated mind, and his personal charm, but this characterization undoubtedly was a product of Vasari's imagination, based upon the poetic quality of Giorgione's paintings.

That the young painter went to Venice to study about 1490 under Giovanni Bellini, the greatest Venetian master of the day, is undeniable. The technique, colour, and mood of Giorgione's pictures are clearly related to Bellini's late style.

Public commissions

Works. The commission of 1507 for a painting or paintings to be placed in the Audience Hall of the Ducal Palace at Venice was perhaps never completed, since no further notice of the work is recorded. Giorgione's principal public commission was the execution of frescoes on the exterior of the Fondaco dei Tedeschi (the German Exchange), where he painted the figures on the facade over the canal. The frescoes over the street were carried out by the young Titian, perhaps under Giorgione's direction. These works, documented in 1508, are lost, except for fragments that contain faint outlines of figures.

Aside from the works mentioned in specific documents, the notes on the art collections of Venice (*Notizie d'opere del disegno*), written between 1520 and 1543 by the Venetian patrician Marcantonio Michiel, contain references to pictures by Giorgione. This information occurs so shortly after the master's death that it is considered generally reliable. Of the 12 paintings and one drawing listed, five works have survived: "The Tempest," "The Three Philosophers," "Sleeping Venus," "Boy with an Arrow," and "Shepherd with a Flute."

"The Tempest" (*c.* 1505) is a milestone in Renaissance landscape painting with its dramatization of a storm about to break. Here is the kind of poetic interpretation of nature that the Renaissance writers Pietro Bembo and Jacopo Sannazzaro evoked. This feeling for nature is probably also intimately related to, though not directly derived from, the philosophical "naturalism" of the contemporary Venetian and Paduan Humanists grouped around the important Renaissance philosopher Pietro Pomponazzi. The meaning of the two people seated in the foreground of "The Tempest" has been the subject of numerous interpretations, none of them definitive. Michiel called them a soldier and a Gypsy. Some literary source of a romantic, Arcadian nature is generally assumed, since no Renaissance artist would include two mysterious figures devoid of meaning. The same kind of evocative literary theme involves the "Pastoral Concert" (*c.* 1510), the attribution of which is much debated, whether it is the work of Giorgione or the young Titian, or both.

The "Sleeping Venus" (*c.* 1510) was left unfinished at Giorgione's death. Michiel stated that the task of adding the landscape background fell to Titian. The picture itself validates this statement, for the landscape with buildings in the right distance is repeated in other works of Titian. Giorgione's "Sleeping Venus" inaugurates a long series of paintings of the goddess of love in Venetian art, particularly those of Titian. None, however, achieved so fully the expression of remoteness and unselfconscious beauty as this majestic and ideally conceived figure. "Judith" (*c.* 1505), though undocumented, evokes the same concept of universal beauty; she is more of a goddess than the avenger of her people.

Idealization of feminine beauty

Few religious paintings are mentioned in the early documentary sources. The panels representing the "Trial of Moses" and the "Judgment of Solomon" are generally agreed to number among the artist's first works (*c.* 1495–1500). Although the figures look slightly archaic, the beauty of the landscape setting with its soft melting distances unmistakably reveals the hand of the painter of "The Tempest." Most celebrated of his religious pictures is the "Madonna and Child with SS. Francis and Liberale" (*c.* 1504, Castelfranco). The composition of this painting forms an equilateral triangle in conformance with the search for geometric solutions characteristic of the Renaissance mind. Thoroughly in the spirit of the master are the landscape and the dreamy mood of the figures, who seem lost in a religious reverie. "The Holy Family" (*c.* 1508) and the "Adoration of the Shepherds" (*c.* 1508) are of equally fine quality. The latter is particularly noteworthy for its exquisitely adjusted colour.

"The Three Philosophers" (*c.* 1510) is one of the works Michiel saw and specifically identified as being by Giorgione. He stated, however, that it was completed by the Venetian painter Sebastiano del Piombo after the master's death. The composition and colour are so fully Giorgione's that Sebastiano could only have added a few finishing touches. In addition, the dreamy melancholy of the three men—who represent youth, maturity, and old age—embodies the spirit of the master. Though the notion of three ages of man is surely implied, little agreement prevails among critics as to whether the three magi,

three philosophers, or a literary source in ancient Roman legend is really intended.

The "Christ Carrying the Cross" (Scuola Grande di San Rocco, Venice) is widely disputed even today. Nevertheless, Vasari in 1568 specifically stated that the painter was Titian, correcting an error that he made in the edition of 1550 in attributing the picture to Giorgione. The canvas, much restored and repainted, has no more than archaeological interest. Other questioned paintings that seem to many 20th-century critics to be the works of Giorgione rather than Titian are "The Adulteress Brought Before Christ" (*c.* 1500), the "Madonna and Child with SS. Roch and Anthony of Padua" (*c.* 1505), and the "Madonna and Child in a Landscape" (*c.* 1504).

Significance as a portraitist

Influence and significance. In portraiture Giorgione made a most profound and far-reaching impression. Venetian painters such as Titian, Palma Vecchio, and Lorenzo Lotto so closely imitated him in the early 16th century that it is at times virtually impossible to distinguish between them. Nevertheless, the portrait of a "Youth" (*c.* 1504) is universally considered to be by Giorgione. The indescribably subtle expression of serenity and the immobile features, added to the chiselled effect of the silhouette and modelling, combine to make the "Youth" an unforgettable expression of Renaissance man. The same sort of exquisite refinement and sensibility characterizes the disputed portrait supposedly of the poet "Antonio Broccardo" (*c.* 1506). Accepted by all critics is the portrait of the so-called "Laura," on the back of which is an inscription giving the date as June 1, 1506, and Zorzi of Castelfranco as the painter. "La Vecchia" ("The Old Woman") bears an inscription *"col tempo"* ("with time") that clearly indicates an allegorical significance of the devastating effects of the passage of time. Despite the singularity of subject in Giorgione's career, the compositional formula and the mention of the picture in the 1569 inventory of Gabriele Vendramin's collection reinforce the attribution.

Giorgione's "Self Portrait as David" (*c.* 1510), recorded in an engraving of 1650 by the well-known German engraver Wenzel Hollar, can safely be considered a much damaged original that has been drastically cut down in size. The artist gave his own portrait more dramatic force by the frown upon his face and by turning the body inward at an angle to the parapet. Titian adopted the same arrangement in his portrait of a gentleman in blue (*c.* 1512, National Gallery, London), where the initials "TV" (Tiziano Vecellio) establish him as the painter rather than Giorgione, as was formerly believed. Despite considerable recent research, the short-lived master from Castelfranco still remains one of the most enigmatic of Renaissance painters. Yet the quality and charm of his paintings have made him as highly esteemed today as he was in his own time—a Venetian master of poetic mood created through idealized form, colour, and light.

MAJOR WORKS

"Judgment of Solomon" and "Trial of Moses" (*c.* 1495–1500; Uffizi, Florence); "The Adulteress Brought Before Christ" (cut down at the right; *c.* 1500; Museum and Art Gallery, Glasgow, Scotland); "Madonna and Child in a Landscape" (*c.* 1504; Hermitage, Leningrad); "Madonna and Child with SS. Francis and Liberale" (*c.* 1504; Cathedral, Castelfranco, Italy); "Portrait of a Youth" (*c.* 1504; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Boy with an Arrow" (*c.* 1505; Kunsthistorisches Museum Vienna); "Fugger Youth" (*c.* 1505; Alte Pinakothek, Munich); "Judith" (*c.* 1505; Hermitage, Leningrad); "Madonna and Child with SS. Roch and Anthony of Padua" (*c.* 1505; Prado, Madrid); "Shepherd with a Flute" (*c.* 1505; Hampton Court Palace, Surrey, England); "The Tempest" (*c.* 1505; Accademia, Venice); "Antonio Broccardo" (*c.* 1506; Museum of Fine Arts, Budapest, Hungary); "Laura" (1506; Kunsthistorisches Museum, Vienna); "Adoration of the Shepherds" ("Allendale Nativity," *c.* 1508; National Gallery of Art, Washington, D.C.); "Epiphany" (*c.* 1508; National Gallery, London); "The Holy Family" (*c.* 1508; National Gallery of Art, Washington, D.C.); "Nude Woman" (1508; fragment from the German Exchange; Accademia, Venice); "Portrait of a Young Man" (1508?; Fine Arts Gallery of San Diego, San Diego, California); "Pastoral Concert" (*c.* 1510; Louvre, Paris); "Self Portrait as David" (*c.* 1510; Herzog Anton Ulrich-Museum, Braunschweig, Germany); "Sleeping Venus" (*c.* 1510; Gemäldegalerie, Dresden; landscape by Titian); "The Three Philosophers" (*c.* 1510; Kunsthistorisches Museum, Vienna); "La Vecchia" ("The Old Woman," *c.* 1510; Accademia, Venice).

BIBLIOGRAPHY. GIORGIO VASARI, *Le Vite de' più eccellenti pittori, scultori ed architettori* . . . (1550 and 1568; exists in several English translations), the major source for Giorgione's life; CARLO RIDOLFI, *Le maraviglie dell'arte* (1648; modern annotated edition by DETLEV VON HADELN, 2 vol., 1914–24), an early source for Giorgione's life and works; MARCANTONIO MICHIEL, *Notizie d'opere del disegno* (*c.* 1532; Eng. trans. by G.C. WILLIAMSON and PAOLO MUSSI, *The Anonimo*, 1903), citations of pictures by Giorgione by a Venetian contemporary; GEORGE M. RICHTER, *Giorgio da Castelfranco, Called Giorgione* (1937), all documents and notices reproduced—catalogue raisonné and complete bibliography until 1936; BERNARD BERENSON, *Italian Pictures of the Renaissance*, 2 vol. (1957), lists of works and attributions; LIONELLO VENTURI, "Giorgione," *Encyclopedia of World Art*, vol. 6, col. 328–339 (1962), a general account, with extensive bibliography; TERISIO PIGNATTI, *Giorgione* (1969; Eng. trans., 1971), documentation and bibliography reproduced from Richter but brought up to date—the attributions are erratic; EDGAR WIND, *Giorgione's Tempesta* (1969), iconographic study of several works.

(H.E.W.)

Giotto

Giotto di Bondone was a 14th-century Italian painter who for more than six centuries has been revered as the first of the great Italian masters. Certainly, he achieved great personal fame in his own lifetime; in the *Divine Comedy*, Dante says of his relation to his reputed teacher, the Florentine artist Cimabue, that "Cimabue thought to hold the field in painting, but now Giotto has the cry, so that the fame of Cimabue is obscured." The mere fact that he was mentioned in Dante, whether or not in a particularly flattering context, was sufficient to establish and maintain this fame in 14th- and 15th-century Italy, and legends soon began to crystallize around his name. When, in 1550, the artist and biographer Giorgio Vasari published *Le vite de' più eccellenti pittori, scultori ed architettori italiani* . . . (*Lives of the Most Eminent Italian Painters, Sculptors and Architects* . . .), he naturally began his history of Italian art with Giotto as the man who, even more than Cimabue, broke away from the Middle Ages and ushered in the "good modern manner."

Anderson—Alinari



"The Lamentation," fresco by Giotto, *c.* 1305–06. In the Arena Chapel, Padua, Italy.

Birth date controversy

Giotto was born at Vespignano near Florence. Very little is known of his life. Much of his biography and artistic development must be deduced from the evidence of surviving works (a large portion of which cannot be attributed to him with certainty) and stories that originate for the most part from the late 14th century on. The date of Giotto's birth can be taken as either 1266/67 or 1276, and the ten years' difference is of fundamental impor-

tance in assessing his early development and is crucial to the problem of the attribution of the frescoes in the church of S. Francesco, in Assisi, which, if indeed by Giotto, are his great early works. It is known that Giotto died on January 8, 1337 (1336, old style); this was recorded at the time in the Villani chronicle. In about 1373, a rhymed version of the Villani chronicle was produced by Antonio Pucci, town crier of Florence and amateur poet, in which it is stated that Giotto was 70 when he died. This fact would imply that he was born in 1266/67, and it is clear that there was 14th-century authority for the statement (possibly Giotto's original tombstone, now lost). Vasari, however, gives 1276 as the year of Giotto's birth, and it may be that he was copying one of the two known versions of the *Libro di Antonio Billi*, a 16th-century collection of notes on Florentine artists. In the *Codex Petrei* version, a statement that Giotto was born in 1276 at Vespignano, the son of a peasant, occurs at the very end of the "Life" and may have been added much later, even, conceivably, from Vasari. In any case, whether Vasari or "Antonio Billi" first made the statement, it cannot have the same authority as that attached to Antonio Pucci, who was about 27 when Giotto died. Certainty of the date of Giotto's birth, if settled by new documents, could help to solve the problem of his work at Assisi, as well as the question of the origins of his style.

Pupil of Cimabue

Giotto has always been assumed to have been the pupil of Cimabue; two independent traditions, each differing on the particular circumstances, assert this, and it is probably correct. Furthermore, Cimabue's style was, in certain respects, so similar to Giotto's in intention that a connection seems inescapable. Cimabue was the most outstanding painter in Italy at the end of the 13th century; he tried, as no artist had before, to break through, with the power of reality and imaginative force, the stylized forms of medieval art. He did not fully succeed, but it seems almost certain that Giotto began his remarkable development with him, inspired by his strength of drawing and his ability to incorporate dramatic tension into his works. On the other hand, whatever Giotto may have learned from Cimabue, it is clear that, even more than the sculptor Nicola Pisano about 30 years earlier, he succeeded in an astonishing innovation that originated in his own genius—a true revival of classical ideals and an expression in art of the new humanity that St. Francis had in the early 13th century brought to religion.

In Giotto's works human beings are the exclusive subject matter, and they act with dedicated passion their parts in the great Christian drama of sacrifice and redemption. By comparison, all his predecessors and most of his immediate successors painted a puppet show with lifeless manikins tricked out in the rags of the splendid, hieratic, and impersonal art of Byzantium, which was to be entirely superseded by the urgent emotionalism of the Franciscan approach to Christianity.

The Assisi problem. The central problem in Giotto studies, the attribution of the Assisi frescoes, may be summed up as the question whether Giotto ever painted at Assisi and, if so, what? There can be no reasonable doubt that he did work at Assisi, for a long literary tradition goes back to the *Compilatio chronologica* of Riccobaldo Ferrarese, who wrote in or before 1319, when Giotto was alive and famous. Later writers down to Vasari expanded this and made it clear that Giotto's works were in the great double church of S. Francesco. By Vasari's time, several frescoes in both upper and lower churches were attributed to Giotto, the most important being the cycle of 28 scenes from the life of St. Francis of Assisi in the nave of the Upper Church and the "Franciscan Virtues" and some other frescoes in the Lower Church.

Revolutionary expression

The majority of these scenes, mostly narrative, are revolutionary in their expression of reality and humanity. In these frescoes, the emphasis is on the dramatic moment of each situation, and, with details of dress and background at a minimum, the inner reality of human emotion is intensified through crucial gestures and glances. In the 19th century, however, it was observed that all these frescoes, though similar in style, could not be by the same hand, and the new trend toward skepticism of Vasari's statements led to the position that rejected all the Assisi frescoes and dated the St. Francis cycle to a period after Giotto's death. This extreme view has been generally abandoned, and, indeed, a dated picture of 1307 can be shown to derive from the St. Francis cycle. Nevertheless, many scholars prefer to accept the idea of an otherwise totally unknown Master of the St. Francis legend, on the grounds that the style of the cycle is irreconcilable with that of the later Arena Chapel frescoes in Padua, which are universally accepted as Giotto's. This involves the idea that the works referred to (in Giotto's lifetime) by Riccobaldo cannot be identified with anything now extant and must have perished centuries ago, so that the early-15th-century sculptor Lorenzo Ghiberti, Vasari, and others mistakenly transferred the existing St. Francis cycle to Giotto. Five hundred years of tradition are thus written off.

Still more difficult, if Giotto did not paint the St. Francis frescoes, major works of art, then they must be attributed to a painter who cannot be shown to have created anything else, whose name has disappeared without trace, although he was of the first rank, and, odder still, was formed by the combined influences of Cimabue, the Florentine sculptor Arnolfo di Cambio, and the Roman painter Pietro Cavallini—influences which coalesce at Assisi and may be taken as the influences that formed Giotto himself.

Influences of Cimabue, Cavallini, and Arnolfo

Arising out of the fusion of Roman and Florentine influences in the Assisi frescoes, there was later a tendency to see the hand of Giotto, as a very young man, in the works of the Isaac Master, the painter of two scenes of "Isaac and Esau" and "Jacob and Isaac" in the nave above the St. Francis cycle. If this theory is accepted, it is easy to understand that Giotto, as a young man, made such a success of this commission that he was entrusted with the most important one, the official painted biography of St. Francis based on the new official biography written around 1266 by St. Bonaventura. In fact, the whole of today's mental picture of St. Francis stems largely from these frescoes. Clearly, a man born in 1276 was less likely to have received such a commission than one ten years older, if, as was always thought, the commission was given in 1296 or soon after by Fra Giovanni di Muro, general of the Franciscans. The works in the Lower Church are generally regarded as productions of Giotto's followers (there are, indeed, resemblances to his works at Padua), and there is real disagreement only over the "Legend of St. Francis." The main strength of the non-Giotto school lies in the admittedly sharp stylistic contrasts between the St. Francis cycle and the frescoes in the Arena Chapel at Padua, especially if the Assisi frescoes were painted 1296–*c.* 1300 and those of the Arena *c.* 1303–05; for the interval between the two cycles is too small to allow for major stylistic developments. This argument becomes less compelling when the validity of the dates proposed and the Roman period *c.* 1300 are taken into account. As already mentioned, the Assisi frescoes may have been painted before 1296 and not necessarily afterward, and the Arena frescoes are datable with certainty only in or before 1309, although probably painted *c.* 1305–06; clearly, a greater time lag between the two cycles can help to explain stylistic differences, as can the experiences that Giotto underwent in what was probably his second Roman period.

Roman period. Three principal works are attributed to Giotto in Rome. They are the great mosaic of "Christ Walking on the Water" (the "Navicella"), over the entrance to St. Peter's; the altarpiece painted for Cardinal Stefaneschi, (Vatican Museum); and the fresco fragment of "Boniface VIII Proclaiming the Jubilee," in S. Giovanni in Laterano (St. John Lateran). Giotto is also known to have painted some frescoes in the choir of old St. Peter's, but these are lost.

Attribution and criticism of the Roman works

These Roman works also pose problems in attribution and criticism. The attribution of the "Navicella" is certain; it is known that Cardinal Stefaneschi commissioned Giotto to do it. The mosaic, however, was almost entirely

remade in the 17th century except for two fragmentary heads of angels, so that old copies must be used for all stylistic deductions. The fresco fragment in S. Giovanni in Laterano was cleaned in the 20th century and was tentatively reattributed to Giotto on the basis of its likeness to the Assisi frescoes, but the original attribution can be traced only as far back as the 17th century. The "Stefaneschi Altarpiece," with its portrait of the Cardinal himself, must be one of the works commissioned by him. The fact that he commissioned Giotto to do the "Navicella" might suggest that this work is by Giotto as well, but the altarpiece is so poor in quality that it cannot be by Giotto's own hand. It may be observed that several works bearing Giotto's signature, notably the "St. Francis of Assisi" (Louvre, Paris) and the altarpieces in Bologna and Florence (Sta. Croce), are generally regarded as school pieces bearing his trademark, whereas the "Ognissanti Madonna," unsigned and virtually undocumented, is so superlative in quality that it is accepted as entirely by his hand.

During this period Giotto may also have done the "Crucifix" in Sta. Maria Novella and the "Madonna" in S. Giorgio e Massimiliano dello Spirito Santo (both in Florence). These works may be possibly identifiable with works mentioned in very early sources, and if so they throw light on Giotto's early style (before 1300). It is also possible that, about 1305, Giotto went to Avignon, in France, but the evidence for this is slender.

Paduan period. There is thus no very generally agreed picture of Giotto's early development. It is some relief, therefore, to turn to the fresco cycle in the chapel in Padua known as the Arena or Scrovegni Chapel. Its name derives from the fact that it was built on the site of a Roman amphitheatre by Enrico Scrovegni, the son of a notorious usurer mentioned by Dante. The founder is shown offering a model of the church in the huge "Last Judgment," which covers the whole west wall. The rest of the small, bare church is covered with frescoes in three tiers representing scenes from the lives of Joachim and Anna, the life of the Virgin, the Annunciation (on the chancel arch), and the life and Passion of Christ, concluding with Pentecost. Below these three narrative bands is a fourth containing monochrome personifications of the Virtues and Vices. The chapel was apparently founded in 1303 and consecrated on March 25, 1305. It is known that the frescoes were completed in or before 1309, and they are generally dated *c.* 1305–06, but even with several assistants it must have taken at least two years to complete so large a cycle.

Narrative skill

The frescoes are in relatively good condition, and all that has been said of Giotto's power to render the bare essentials of a setting with a few impressive and simple figures telling the story as dramatically and yet as economically as possible is usually based on the narrative power that is the fundamental characteristic of these frescoes. These dominating figures, simple and severe, similar to those in the Assisi cycle but placed in settings of more formal abstraction and rendered with more grandeur, are the quintessence of his style, and anatomy and perspective were used—or even invented—by him as adjuncts to his narrative gifts. He never attained to the skill that so often, in fact, misled the men of the 15th and 16th centuries. In the Padua frescoes the details are always significant, whereas it is a characteristic of the Assisi cycle that there occurs from time to time a delighted dwelling on details that are not absolutely essential to the story.

Sta. Croce frescoes. Documents show that Giotto was in Florence in 1311–14 and 1320; and it was probably during these years, before going to Naples (*c.* 1329), that he painted frescoes in four chapels in Sta. Croce belonging to the Giugni, Tosinghi-Spinelli, Bardi, and Peruzzi families. The Giugni Chapel frescoes are lost as are all the Tosinghi-Spinelli ones, except for an "Assumption" over the entrance, not universally accepted as by Giotto. The Bardi and Peruzzi chapels contained cycles of St. Francis, St. John the Baptist, and St. John the Evangelist, but the frescoes were whitewashed and were not recovered until the mid-19th century, when they were damaged in the process of removing the whitewash and then heavily restored. Much the same happened to a portrait of Dante in the Bargello, also in Florence, for which there is a traditional attribution to Giotto. Writers tended to take more or less account of these additions and restorations according to the view they held of the Assisi problem, but a prolonged cleaning and re-restoration of both chapels in the mid-20th century has demonstrated that the Bardi Chapel has few but splendid figures remaining, painted in true fresco, whereas the Peruzzi Chapel figures are now largely ghosts, since they were painted in a different technique. The older view, that the two cycles were contemporary, is no longer necessarily valid, and there is no evidence for the date of either cycle, except that both are probably later than the Arena Chapel frescoes.

Naples and the last Florentine period. In January, 1330, King Robert of Naples promoted Giotto to the rank of "familiar" (member of the royal household), which implies that he had been in Naples for some while, possibly since 1329, and he remained there until 1332–33. All the works he executed there have been lost, but traces of his style may be distinguished in the local school. On April 12, 1334, he was appointed *capomastro*, or surveyor, of the Duomo in Florence and architect to the city. This was a tribute to his great fame as a painter and not on account of any special architectural knowledge. On July 19 of the same year he began the campanile, or bell tower, of the Duomo. It was later altered but is known, in part at least, from a drawing in Siena. He may have designed some of the reliefs carved by Andrea Pisano on the campanile; certainly the bronze doors of the baptistery by Andrea show clear traces of Giotto's frescoes in Sta. Croce. Indeed the whole course of painting in Tuscany was dominated by his pupils and followers—by Taddeo Gaddi, Bernardo Daddi, Maso di Banco, Andrea Orcagna, and Pietro and Ambrogio Lorenzetti in Siena—but none of these really understood all of his innovations, and it was not until the Renaissance, with Masaccio and Michelangelo, that his true successors arose.

Architect to the city of Florence

MAJOR WORKS

The only works universally accepted as Giotto's are the fresco cycle in Padua, firmly datable in the first decade of the 14th century, and the two chapels in Sta. Croce, Florence, which used to be dated around 1320, but since their cleaning in the 1950s and 1960s are now placed anywhere in the second or third decade. The only panel painting universally accepted as Giotto's own work is "The Madonna in Glory" ("Ognissanti Madonna"; Uffizi, Florence), usually dated *c.* 1305–10.

The most controversial attribution is the fresco cycle at Assisi, but other important works in this category are the "Crucifix" in Sta. Maria Novella, Florence (*c.* 1295–1300); the "Dormition of the Virgin" (Staatliche Museen Preussischer Kulturbesitz, Berlin); and the polyptych now in the Museo dell'Opera, Sta. Croce, Florence.

Works that are certainly from Giotto's shop, not necessarily painted by his own hand, include panels in Bologna Pinacoteca Nazionale and the Vatican, and others in Boston (Isabella Stewart Gardner Museum) and the galleries of London, Munich, New York, and Washington. A fragment of the original mosaic from St. Peter's is now in the church at Boville Ernica, near Rome, and another is in the Vatican Museum.

BIBLIOGRAPHY. There is a full bibliography, from the 14th century to 1937, in ROBERTO SALVINI, *Giotto bibliografia* (1938). The earliest important biography is by GIORGIO VASARI in his *Vite*, published in Florence in 1550 and 1568: the 1550 edition has not been translated, but there are many English versions of the 1568 one. Full-scale monographs are ROBERTO SALVINI, *Tutta la pittura di Giotto*, 2nd ed. (1962; Eng. trans., *All the Paintings of Giotto*, 1963); and CESARE GNUDI, *Giotto* (1958; Eng. trans., 1959); and special studies include P. MURRAY, "On the Date of Giotto's Birth," in *Giotto e il suo tempo* (1971); ALASTAIR SMART, *The Assisi Problem and the Art of Giotto* (1971); LEONETTO TINTORI and MILLARD MEISS, *The Painting of the Life of St. Francis in Assisi* (1962); JAMES STUBBLEBINE (comp.), *Giotto: The Arena Chapel Frescoes* (1969); and LEONETTO TINTORI and EVE BORSOOK, *Giotto: The Peruzzi Chapel* (1965).

(P.J.Mu.)

Glaciation, Landforms Produced by

The form and distribution of landscape features on the earth's surface are the imprints of the geological processes that produced them. Those parts of the earth's land surface that were covered by glacier ice during the geologic past contain a variety of landforms testifying to the former presence of glaciers. These are, quite properly, called glacial landforms, but even regions beyond the former glacier boundary—either in the cold-dominated zones of polar latitudes or in latitudes in which more temperate climates prevail today—contain landforms that are wholly or partly the result of climatic conditions not existing today. These features also are attributable to glaciation and will be discussed and described in this article.

Landforms produced by glaciation are of two general types, erosional and depositional. Erosional landforms are those produced by the direct down-wearing action of the ice as it moved across the preglacial land surface. Depositional landforms are those formed directly by ice, either at the bottom of the glacier or at the margin, or by glacial meltwaters flowing in ice channels, subglacial tunnels, or along and beyond the glacier front.

Extent of Pleistocene glaciation

The extent of glaciation during the Pleistocene Epoch (2,500,000 to 10,000 years ago) has been determined principally from the geographic distribution of glacial landforms and deposits of glacial origin. Glacial landforms produced by past glaciations are found on all continents. Ice sheets or continental glaciers covered all of Canada and much of the United States as far south as the Ohio and Missouri rivers. Existing glaciers in Alaska and the Rocky Mountains were greatly enlarged. Many mountain valleys that contain no glaciers today were occupied by glaciers during the Pleistocene. These valley or alpine glaciers were confined to former river valleys, in contrast to the ice sheets that engulfed whole landscapes of continental proportions. Approximately one-third of the earth's total land area was covered by glaciers during the maximum ice advance.

In Europe the Scandinavian countries, England, northern Germany, and the Baltic areas were glaciated. The glaciers of the Alps were greatly enlarged and many modern river valleys, such as the Inn Valley in Austria, were filled with huge glaciers—literally rivers of ice.

In the Soviet Union both the European sector (west of the Ural Mountains) and much of Siberia were glaciated. The Arctic islands such as Novaya Zemlya and Franz Josef Land were nearly completely ice covered.

High mountain ranges throughout the world—the Himalayas, the Andes, and the volcanic peaks of Ruwenzori, Mt. Kenya, and Kilimanjaro in Equatorial Africa, for example—were more extensively glaciated during the Pleistocene than they are today. The southern Andes (Patagonia) was the scene of greatly expanded glaciers, and the Southern Alps in New Zealand were glaciated over an area that was perhaps 50 times larger than the area covered by existing mountain glaciers in that country. In the Pacific, Mauna Kea on the island of Hawaii supported glaciers on its summit, where there are no glaciers today.

Glacial landforms and geological history

The geological record shows that at least three times in the last third of geological history extensive glaciation occurred in many parts of the world where no glaciers exist today. The earliest widespread glaciation occurred during Precambrian time (prior to 570,000,000 years ago). The evidence for this lies mainly in glacially derived sedimentary materials called tillites and in glacial markings, such as grooves and scratches, on the older rocks on which the tillites occur. The Gowganda Glacial Tillite of central Canada indicates the presence of an ice sheet that covered several thousand square miles during Middle Precambrian time. Tillites in North America, Scandinavia, Greenland, India, China, Africa, and Australia testify to the possibility that the greatest glaciation the earth has ever experienced took place during the final phases of the Precambrian.

Tillites in South America, South Africa, Australia, India, and Antarctica seem to record the subsequent depositional activity of several large glaciers during the Permian Period (225,000,000 to 280,000,000 years ago). The Dwyka Tillite of South Africa lies on an older rock surface that bears the grooves and other markings of this inferred Permian glacial action. The widespread occurrence of tillites equivalent in age to the Dwyka and the direction of ice movement deduced from scratches and grooves have been used to support the hypothesis of continental drift (*q.v.*). Accordingly, the origin of the Dwyka and related tillites has been related to a single ice sheet of continental dimensions that existed when the continents of the Southern Hemisphere and India, and possibly China, were part of a single large landmass. It should be noted, however, that some authorities believe that the tillites reflect 30 or more separate ice masses, scattered throughout the Southern Hemisphere during the Permian and the preceding Carboniferous Period. In either case glacially derived landforms and deposits are of considerable value in reconstructing the gross geography of the earth at various times in its history.

The last of the great glaciations occurred during the Pleistocene Epoch (2,500,000 to 10,000 years ago). The maximum extent of the Pleistocene glaciers is much easier to establish than that of older glaciations because the Pleistocene glacial landforms and related deposits generally are not covered by extensive blankets of younger sediments. In some cases, the oldest Pleistocene deposits are overlain by younger glacial materials or have been severely dissected by postglacial streams.

The greatest success in the use of glacial landforms as a means of reconstructing the detailed history of a particular Pleistocene glacial event has been in the establishment of a glacier border during its retreat from a position of maximum advance. Studies of modern glaciers, both those confined to valleys (valley glaciers) and those covering broad expanses of the landscape (continental glaciers), reveal the development of ridges and hummocky topography at the ice border, where glacial sediments (till) and related glacial meltwater deposits are dumped along the ice front. When a glacier retreats from a position it has occupied for some time, perhaps several hundred years, the belt of ridged and hummocky terrain marking the former glacier margin is left behind. This feature is called an end moraine; a single end moraine may be traced for several hundred miles and plotted on a map to show the position of the glacier margin at the time the end moraine was formed. Parallel end moraines thus reveal successive positions of the retreating ice front during the general shrinking of a continental glacier in the waning phases of the Pleistocene. Tree stumps may become incorporated in a moraine and can be dated by geochemical methods. If it is assumed that the trees were killed by glacier overriding, then the absolute age of the dead trees is also the age of the moraine materials containing them.

Importance of glacial landforms and deposits to man

The landscapes produced by continental glaciers extend from the high latitudes to the more temperate middle latitudes, especially in Europe and North America. Glacial drift (ice-transported debris) provided an excellent source material for the development of fertile soils during postglacial time. Wherever the presence of such soils has coincided with favourable climatic conditions, as in the Upper Mississippi Valley or northern Europe, agricultural enterprises of extremely high productivity have emerged. Many farming and dairying industries are thus related to the distribution of glacially derived soils.

Glacial terrains of Pleistocene age generally are associated with abundant surface and subsurface water, the single most important natural resource for the sustenance of large populations and industrial centres. Modern rivers the origins of which are related to glaciation provide hydroelectric power in North America, in the fjord (drowned valley) regions of Scandinavia, and elsewhere in the world. The tens of thousands of lakes in Canada, the Great Lakes of the United States, and the lakes in Scandinavian countries are mostly of glacial origin, and are invaluable for recreation and other uses.

Groundwater supplies in the glaciated areas are contained in aquifers consisting of glacial sands and gravels.

Many municipalities and a multitude of industries derive their primary water supplies from relatively shallow groundwater reservoirs in the glacial landscape.

Earth materials of glacial origin are also of significant economic value. Sand and gravel, widely available in glaciated areas, are used for aggregate in the manufacture of concrete and for road material on secondary roads. Clays deposited in ancient glacial lakes are exploited for bricks, drain tile, and various clay products.

Glaciated landscapes thus provide man with a great variety of natural resources ranging from agricultural lands, water, and constructional materials to scenic lakes and picturesque expanses of wooded hills and valleys.

The direct and indirect effects of glaciation in terms of glacial landforms, features, and deposits are dealt with in this article. For further information on the geological time intervals that coincided with major glaciations, see PRECAMBRIAN TIME; PERMIAN PERIOD; and PLEISTOCENE EPOCH. The modern distribution of ice is covered in the articles ICE SHEETS AND GLACIERS; ICEBERGS AND PACK ICE; and ICE IN RIVERS AND LAKES; and certain glacially derived features or deposits are described in COASTAL FEATURES; CONTINENTAL SHELF AND SLOPE; WATERFALLS; PERMAFROST; VARVED DEPOSITS; and CONGLOMERATES AND BRECCIAS. See also CLIMATIC CHANGE for the evidence of climatic fluctuation through time; HYDROLOGIC CYCLE for a discussion of the importance of ice in the earth's water balance; and LANDFORM EVOLUTION for treatment of the general cause-and-effect relations that govern the configuration of landforms.

This article is divided into the following sections:

I. The processes of glaciation
  - Glacial erosion
  - Glacial transport
  - Glacial deposition
  - Glacial loading and unloading
  - Periglacial processes

II. Landforms produced by glacial erosion
  - Microrelief and small-scale features
  - Glaciated valleys and associated features
  - Glacially eroded rock basins in nonmountainous regions

III. Landforms produced by glacial deposition
  - Nonstratified drift and associated landforms
  - Stratified drift and associated landforms
  - Glaciolacustrine sediments and associated landforms

IV. Landforms produced by periglacial processes
  - Permafrost
  - Talus, rock glaciers, and blockfields
  - Patterned ground
  - Eolian deposits

V. Modern landforms and former glaciation
  - Modern lakes and Pleistocene forerunners
  - Coastal features and sea-level change
  - Drainage patterns and ice disruption
  - Climatic implications

I. The processes of glaciation

As a glacier moves over the land surface it modifies the terrain by removing material (glacial erosion) or by depositing debris carried within the ice mass or on the glacier surface (glacial deposition). Each of these major glacial processes produces distinctive landforms that remain long after the glacier has disappeared. In addition to direct glacial erosion and deposition, closely related geological events occur in nonglaciated regions as a direct consequence of the glacial ice, or the climatic conditions that caused the glaciers to grow and expand. These events include the depression of the earth's crust resulting from the superimposed weight of glacier ice, the creation of lakes at the edge of the glacier during its retreat, the distribution of wind-transported sand dunes and silt blankets, emerged and submerged shorelines resulting from sea-level fluctuations in response to waxing and waning of the ice sheets, and the development of characteristic features produced by frost action in cold-dominated regions that were not covered by glacier ice.

GLACIAL EROSION

Glaciers are made up of interlocked grains of ice crystals. Laboratory experiments on single ice crystals and on samples of glacier ice show that ice deforms plastically when stress is applied. In nature the stress is gravity, the basic cause of ice movement. Two kinds of glacier movement have been identified: internal deformation of the ice by slippage or shear along certain planes in the individual ice crystals, and sliding of the base of the glacier on its bed. The velocity of the glacier thus consists of two components, an internal one resulting from deformation or creep of the glacier, and a basal one resulting from the sliding of the glacier on its bed.

The basal ice in many glaciers is at or near the pressure-melting point of ice, so that freezing and thawing occur at various times at the plane of contact between the glacier and the rock or soil beneath it. This permits loose debris such as silt, sand, cobbles, and larger rock fragments to be incorporated as part of the moving glacier mass. These basal, non-ice constituents of the glacier are abrasive tools that scratch, polish, or groove the rock or frozen ground over which the glacier moves.

Glacial quarrying

The melting and refreezing of basal ice is a process that is capable of removing very large blocks of bedrock and incorporating them in the moving glacier mass. Many rocks of the earth's crust characteristically contain intersecting fractures or cracks, called joints, which define incipient angular blocks that can be incorporated into the base of a glacier through the melting-freezing process. Large-scale excavation of many blocks (the individual dimensions of which may range from a few feet to more than ten feet) is another form of glacial erosion known as glacial quarrying, or plucking.

The exact mechanism of glacial quarrying is not known because conditions at the base of a glacier cannot be directly observed. Tunnels, driven into glaciers along the ice-rock basal contact, have yielded very little information about quarrying. Generally, the excavation of rock by glacial quarrying seems to be restricted to rocks that are well-jointed. Massive, unjointed rocks, such as quartzite and others with few joints, are smoothed and polished by overriding glaciers. Glacially quarried surfaces, on the other hand, are rough and jagged because blocks were lifted out by the poorly understood quarrying process.

A good deal of controversy exists about the efficacy of glacial erosion. Some geologists admit that glaciers are capable of abrading rock surfaces but deny that glaciers produce any major landforms comparable to the large canyons cut by rivers. It is suggested instead that such features were produced by preglacial erosion and that glacial modification of them was insufficient to obliterate their preglacial forms.

The real question is not whether glaciers are capable of eroding the land surface but, rather, the magnitude of glacier erosion (see the later section of this article, *Landforms produced by glacial erosion*).

GLACIAL TRANSPORT

Material incorporated in a glacier by abrasion or plucking is transported by glacier flow until it is deposited directly by the ice, or until the glacier melts and leaves its load as a mantle over the landscape. Valley glaciers receive material from the valley walls confining them. Individual boulders and rock fragments, loosened by frost action from cliffs above the glacier surface, fall onto the glacier and are carried in conveyor-belt fashion as the ice stream flows down-valley. In some instances large landslides and debris flows may descend to the surface of a valley glacier, which then transports the rock and soil debris in a down-glacier direction.

Ablation and accumulation of glaciers

A glacier consists of two parts, the zone of accumulation and the zone of ablation. The former occurs in the upper reaches of the glacier where more snow falls each winter than is melted in the following summer. The latter is found at lower elevations on a glacier where, in addition to melting of the annual snowfall during the summer months, part of the glacier ice is destroyed by melting. If a glacier ends in a lake or the ocean, large masses of glacier ice become detached from the glacier terminus to become icebergs. Ablation thus is any process that removes ice from the glacier mass.

Rock debris that falls on a glacier in the zone of accumulation becomes buried by the snowfall of successive winters. A boulder that comes to rest on the glacier surface in the zone of accumulation has two components of movement as it is carried by the glacier. One component is down-valley, parallel to the general surface of the glacier. The other component is vertically downward as the boulder becomes covered by successive snow layers. Snow in the accumulation zone is transformed to glacier ice by a process of recrystallization as it is buried deeper and deeper each year. Debris falling on the glacier surface as airborne dust or boulders from a rockfall ultimately becomes imbedded in the glacier ice.

The flow lines of a glacier in the zone of accumulation are thus inclined downward toward the base of the glacier, whereas in the ablation zone, the flow lines have an upward component toward the glacier surface. Because of this pattern of flow, debris falling on the uppermost reaches of a glacier will be carried to the base of the glacier and then move upward until it reappears at the surface of the ice in the ablation zone.

The boundary between the ablation zone and the accumulation zone is called the snowline. It is best observed in late summer at the end of the ablation season. At that time the accumulation zone still has a residual layer of snow from the previous winter, whereas the ablation zone is strewn with debris released from the melting glacier ice.

If the amount of annual accumulation exceeds annual ablation over a period of years, then mass is added to the glacier. The glacier responds in two ways: it thickens, and it expands over a larger area. Valley glaciers expand in a down-valley direction and ice sheets expand by spreading outward in all directions. In both cases the glacier is said to be advancing.

If, on the other hand, ablation exceeds accumulation over a period of time, the glacier responds by thinning and by reducing its total area. The terminus of a valley glacier retreats in an up-valley direction while the edge of an ice sheet withdraws along the entire margin. During advance and retreat, the glacier continues to transport material in the direction of glacier flow.

From the foregoing it can be deduced that a vigorously advancing glacier may transport debris lying at its terminus by a "snowplow" action, thereby pushing material forward as the snout or margin advances. The snout is also the site of transportation of debris along thrust planes within the glacier itself. Thrust planes in some glaciers are exposed in crevasses near the glacier terminus or on the ablation surface, where the edges of the thrust planes occur as a series of parallel or subparallel bands emphasized by a concentration of dirt and stones that have been carried to the glacier surface by thrusting.

GLACIAL DEPOSITION

Material transported by glacier ice eventually comes to rest on the land surface. Glaciers that terminate in the ocean (tidewater glaciers), such as those discharging from Greenland, Alaska, or Antarctica, deposit material on the sea floor. If the terminus of a tidewater glacier is grounded on the sea floor, deposition is concentrated around the edges of the grounded portion as debris frozen in the glacier is released by melting. If the terminus is afloat, sediment released from the underside of the glacier will fall to the sea floor as melting progresses. These glacio-marine sediments are distinguished from other marine deposits by their heterogeneous detrital content.

Fluvioglacial processes at the ice margin

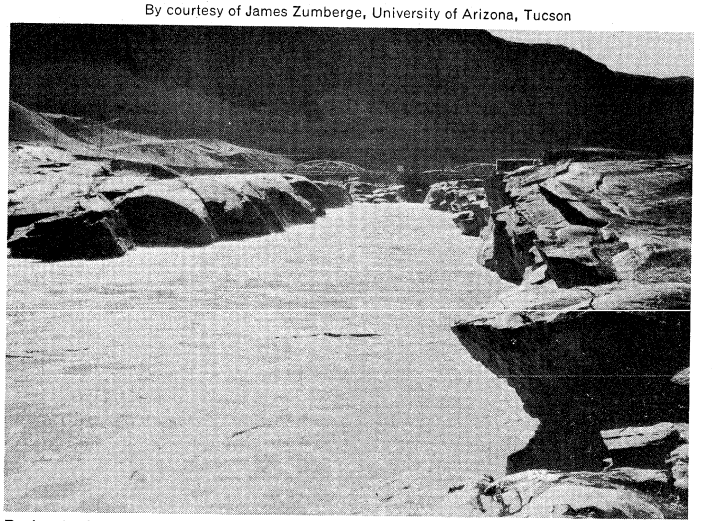
During the height of the ablation season, the terminus of a glacier on land is a chaotic mixture of melting ice, running water, and mounds of rock and soil debris recently released from the ice by ablation. These conditions change hourly; in late afternoon the discharge of glacial meltwater reaches a maximum, and large volumes of water can be seen flowing on the glacier surface in ice channels or debouching from tunnels in the ice. The flow of meltwater slackens after sunset, and by the early morning hours of the following day, the volume of meltwater being discharged from the glacier may be only a small fraction of the maximum flow of the preceding day.

These conditions give rise to a great variety of glacially derived sediments ranging from ablation debris blanketed over stagnant ice masses to sands and gravels deposited by glacial meltwaters a few miles down-valley from the glacier terminus. The extremely variable conditions of transportation in space and time give rise to extremely varied deposition of sediments. Ablation boulders and unsorted rock rubble may slump onto a deposit of well-sorted gravels, or boulders may fall from the ice front into a quiet meltwater pool where fine sand and silt are accumulating on the bottom.

If a glacier margin retreats year after year, these chaotic ice-marginal deposits will persist over a wide expanse of recently deglaciated terrain. Pauses in the retreat may result in a concentration of the ice-marginal deposits in the terminal zone. If the ice front advances, it may override the older debris-covered areas or push some of it into ridges and mounds that remain for centuries after the glacier has retreated or disappeared entirely.

Meltwater issuing from the ice front in rivulets on the ice, or torrential discharges from ice tunnels, generally converges to form a single channel or series of channels that divide and recombine in an anastomosing (braided) pattern. These braided rivers are choked with glacially derived sediment called outwash or glaciofluvial sediments.

By courtesy of James Zumberge, University of Arizona, Tucson



Bedrock channel near the head of Søndre Strømfjord, West Greenland, carved by silt-laden glacial meltwaters. The rock surface was smoothed by glacial erosion.

If a glaciofluvial river empties into a lake, the coarser fraction of the suspended load is deposited at the river mouth while the silt and clay, called rock flour, are deposited on the lake floor. Glacial meltwaters heavily charged with rock flour are distinctively grayish-white in colour and are referred to as *Gletschermilch* in the Alpine regions of Europe.

GLACIAL LOADING AND UNLOADING

The Earth's crust tends toward a condition of balance or isostatic equilibrium. If the crust is considered to be a series of large blocks extending to a constant depth, approximately 20 to 50 kilometres (12 to 30 miles), each of the blocks will stand at a different elevation above sea level because of its difference in density. This condition of crustal balance is constantly changing because material from the higher-standing blocks (the continents) is being eroded and transported to the lower blocks (the ocean basins).

If glacier ice accumulates to some appreciable thickness over a large part of the earth, the crust will be depressed about 300 metres (1,000 feet) for every 900 metres (3,000 feet) of ice. Crustal downwarping increases in magnitude from the marginal areas of a large ice sheet toward the zone of greatest thickness. Where the ice is thin, the crustal downwarping is zero, and where the ice is thickest, the depression of the crust is greatest.

Deformation of the crust because of glacial loading is not a permanent condition. After an ice sheet has melted away, the crust restores itself to a new condition of isostatic balance by rising in response to the glacial unloading.

Postglacial uplift in progress today

Crustal downwarping from glacial loading, and crustal uplift from glacial unloading are operative on the earth today. The crust beneath Greenland and the Antarctic ice sheets is downwarped several thousand feet because of the load of glacier ice on each. In contrast, crustal upwarping is going on in Scandinavia, North America, and the British Isles as a result of the disappearance of the ice sheets from those areas somewhat less than 10,000 years ago. The rate of uplift is determined by the change in the elevation of tide gauges with respect to mean sea level during historical time. The Scandinavian Peninsula, for example, is currently rising at differential rates: north of the Gulf of Bothnia the rate is about 88 centimetres (35 inches) per century, near Stockholm it is around 40 centimetres (16 inches) per century, and at Copenhagen it is zero. These rates are undoubtedly less than they were in the earlier stages of deglaciation, but it appears from other evidence that the land around the Baltic Sea may rise another 600 feet in the centre of the area of uplift before isostatic equilibrium is regained. The maximum uplift in Fennoscandia at that time will have been as great as 760 metres (2,500 feet). If it is assumed that the postglacial rebound equals the total amount of downwarping, and that the ratio of downwarping to ice thickness is about one to three, then the maximum ice thickness for the Fennoscandian region was about 2,300 metres (7,500 feet). This is comparable to the measured ice thickness of the Greenland Ice Sheet, which covers about the same area today as did the Scandinavian Ice Sheet during the Pleistocene.

The depression of the earth's crust by glacier ice produces some side effects in the nearby nonglaciated areas. Coastal regions are submerged, causing invasion of the land by the sea. River mouths are inundated by seawater and river gradients are reduced, giving rise to deposition of river sediments.

The postglacial rise of a previously glaciated land area is recognized in coastal regions such as Scandinavia, the eastern United States, and the Great Lakes region of southern Canada (Figure 1). The evidence comes from differentially uplifted or warped strandlines, which marked the shorelines of water bodies marginal to the retreating Pleistocene ice sheets. These strandlines, represented by beach ridge deposits, were horizontal when formed, but because of glacial rebound since deglaciation these linear features now rise toward the direction of greatest ice thickness. Accurate dating of successively younger strandlines provides the basis for determining postglacial rates of uplift. Crustal recovery patterns are depicted in maps showing isobases, lines connecting points of equal uplift in a given region, as around the Great Lakes or the Gulf of Bothnia and the Baltic Sea. The limit of uplift is defined by the zero isobase (no uplift) or "hinge line." In the Great Lakes region the hinge line has an east-west trend roughly from Milwaukee through Cleveland to New York City. South of that line, all strandlines of the ancestral Great Lakes are horizontal, a fact that implies no uplift. Northward from the zero isobase of the Great Lakes region, uplift is still in progress, but the rate appears to be declining. It has been shown that the maximum uplift north of the Great Lakes occurred at Golfe de Richmond at about latitude 56° N on the east coast of Hudson Bay. There, at least 300 metres (1,000 feet) of uplift has occurred since the load of the ice sheet was reduced by thinning at the end of Pleistocene time. Assuming that complete uplift represents one-third of the ice thickness, the uplift at Golfe de Richmond reflects only about 900 metres (3,000 feet) of ice. This value is too small in relation to maximum Pleistocene ice thickness for the Greenland and Antarctic ice sheets, at comparable distances from their margins. It must therefore be assumed that uplift in the Hudson Bay region is still in progress, and that as much as 600 metres (2,000 feet) of additional uplift can be expected to occur.

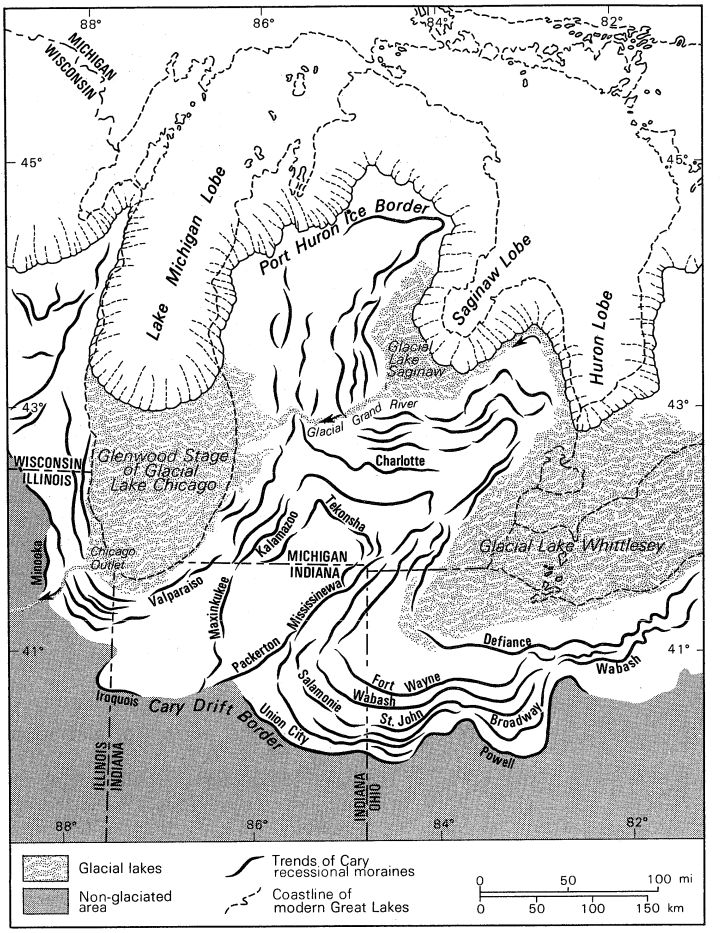


Figure 1: Upper Great Lakes region during Port Huron maximum, about 13,000 years ago.

Adapted from Wayne and Zumberge, "Pleistocene Geolbgy of Indiana and Michigan," in *The Quaternary of the United States*, by H.E. Wright, Jr., and David G. Frey (copyright © 1965 by Princeton University Press), fig. 4, p. 70, reprinted by permission

PERIGLACIAL PROCESSES

As originally used, the term periglacial was an adjective used to describe the climatic conditions prevailing in a nonglacial zone marginal to a large ice sheet. As used today, the term is synonymous with "cold-dominated" and is applicable to a rigorous climate in which the freeze-thaw process predominates. Generally speaking, the periglacial environment is characterized by perennially frozen ground (permafrost), low precipitation, strong winds, and vegetation ranging from the boreal forests (central Alaska), through taiga (northern Canada) and tundra (Arctic Alaska and Siberia), to vegetation-free areas (northern coast of Greenland and glacier-free areas of Antarctica).

The periglacial environment was undoubtedly more widespread during the Pleistocene than now, and many effects of this environment are preserved in various landforms that resulted from periglacial processes that are no longer active. The landforms resulting from these processes will be discussed later in this article. Whereas such features are not, in the strictest sense, landforms produced by glaciation, they are closely related to glacial climates and are commonly superimposed on terrains that originated through glaciation. Moreover, the presence of underground ice, and its thawing and refreezing are basic requirements for many of the landforms associated with the periglacial environment.

II. Landforms produced by glacial erosion

Glacial movement consists of two parts, internal flow or plastic deformation, and slippage of the basal ice along the ground surface over which it moves. Along the margins of valley glaciers, where glacier ice is in contact with the rock surface, some of the effects of a glacier moving over a bedrock surface can be assessed.

Erosion by glacier ice has been deduced from the study of areas formerly covered by glaciers, and from direct observations along the walls of glaciated valleys where the contact between ice and bedrock can be seen. The landforms resulting from the erosive action of glaciers range in size from very small forms of microrelief to large-scale phenomena with dimensions in hundreds or thousands of metres. These landforms are the product of glacial abrasion, glacial quarrying or plucking, or a combination of these two processes.

MICRORELIEF AND SMALL-SCALE FEATURES

Rock polish, striations, grooves, and friction cracks. Pure ice at its pressure-melting point has little abrasive ability on hard rock, yet many granite outcrops are highly polished as the result of the passage of glacier ice. Other rocks such as massive limestones and basaltic lavas exhibit small scratches and other markings that were made by glaciers. These features are caused by debris-laden ice during its passage over rock outcrops.

Determination of ice-movement directions

Striations are scratches on a rock surface, and when the orientations of a sufficient number of them are plotted on a map, they provide a means of determining the former direction of ice movement over a relatively large area. The mapping of glacial striae in Finland, New England, and parts of the Canadian Shield are examples of how this technique has been used to deduce the general flow pattern of Pleistocene ice sheets. Many glaciated rock surfaces contain striae that diverge in orientation. Although it is possible that crossing striae on a single rock outcrop represent two distinct periods of glaciation, it is more likely that they reflect local variations in the flow direction of the basal ice during a single glacial advance. Because striae may diverge as much as 90 degrees from the direction of regional ice movement, a few scattered observations of striae orientations are not reliable indicators of regional flow direction. Only when hundreds of striae directions are plotted on a map covering thousands of square kilometres will regional flow patterns emerge.

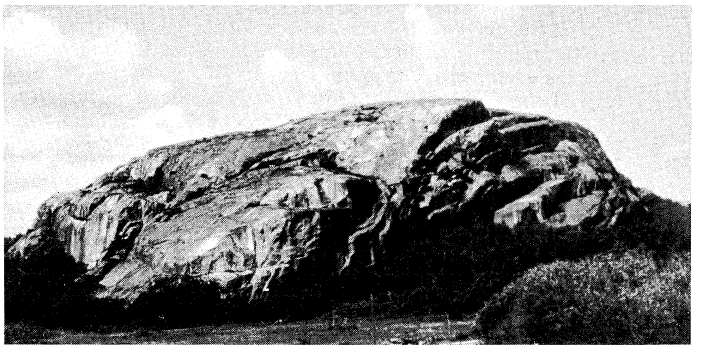
Glacially formed grooves in bedrock are less common than striae. Some grooves appear to be the result of a single angular rock fragment carried by glacier ice over a relatively soft and fine-grained rock surface such as limestone. The glacial grooves of Kelleys Island in western Lake Erie occur in limestone bedrock, and range from less than 0.3 metre (one foot) to more than one metre (three feet) deep and are three metres (ten feet) or more in length. The origin of these and other glacial grooves is in dispute because it is not known whether subglacial meltwater or a slurry of mud moving beneath the ice could have produced all or part of them. The presence of true striae and glacial polish along the walls of such grooves is a clear indication of the passage of ice along their axes, but it does not prove that the entire groove was caused by glacier abrasion. Some geologists have argued that fast-flowing subglacial meltwater or slurries of mud and other glacial-fluvial sediments scoured these grooves while the ice was present. Others suggest that grooves are nonglacial features the gross shapes of which were only slightly modified by the passage of glacier ice.

Some of the largest grooves associated with glaciated terrain are those in the Mackenzie Valley in the Northwest Territories of Canada. There the bedrock is furrowed with grooves, some of which are as much as 30 metres (100 feet) deep and more than 1.5 kilometres (one mile) long. The grooves occur with parallel alignment and are indifferent to the regional "grain" of the bedrock or surface drainage lines. It is believed, however, that the axes of the grooves coincide with the true direction of ice movement.

Crescentic and lunate fractures and chatter marks

Friction cracks are small-scale markings on rock surfaces formerly overridden by glacier ice. They appear to be the result of frictional contact between a glacier-carried boulder and a bedrock surface. Commonly, these features are curved linear cracks in plan view with either the convex or concave side oriented toward the direction of ice movement. In cross section, a friction crack consists of two intersecting rock crevices that are inclined at different angles with respect to the rock surface. The more steeply inclined crack always faces away from the direction of ice movement, regardless of the surface manifestation of the marking. This generalization has been verified by the orientation of striae on the same rock surface. Friction cracks have been given a number of names such as crescentic gouges, crescentic fractures, or lunate fractures, depending on specific characteristics of each. These cracks were formerly called chatter marks in reference to the kind of markings achieved on a hard surface by forcing another hard material across it.

By courtesy of James Zumberge, University of Arizona, Tucson



Roche moutonnée in southern Sweden. Ice moved from left to right.

Roches moutonnées. A roche moutonnée is a bedrock knob or protuberance that has been shaped by glacier overriding. These bedrock hills range in size from a few metres to several kilometres in the longest horizontal dimension, and about three metres to 30 (98 feet) or more metres high.

Roches moutonnées are a characteristic landform of a glaciated bedrock surface. They result from erosion by valley glaciers as well as by ice sheets. Commonly they are elongate in plan view with the long axis parallel to ice movement, but this is not by any means a universal relationship. More diagnostic than the plan view is the vertical profile parallel to the direction of ice movement. The side facing the oncoming ice (stoss side) is gently inclined, smoothed with glacial polish, and contains the usual marks of glacial abrasion. The side away from the moving ice (lee side), on the other hand, is steeply inclined, jagged, and rough, with little sign of abrasive action. The smooth stoss side is formed by glacier abrasion, and the jagged lee side is commonly regarded as the product of glacier plucking. This stoss-and-lee profile is the hallmark of a roche moutonnée, large or small, and is, along with glacial polish, one of the unique products of glacier erosion in bedrock.

GLACIATED VALLEYS AND ASSOCIATED FEATURES

Some of the most scenic landscapes in the world are valleys formerly occupied by glaciers. These valleys occur in most mountainous regions and bear witness to the climatic regime of the Pleistocene. A glaciated valley has a unique geometric form that sets it apart from a valley that was cut solely by running water. Its gross shape, as well as specific features on its walls and floor, reflects the sculpturing effect of a valley glacier (Figure 2).

A valley glacier comes into existence in mountainous regions in which temperature and snowfall are conducive to the annual net accumulation of snow. A succession of years in which more snow accumulates at high elevations than is lost by ablation eventually produces glacier ice that flows under the influence of gravity down the valleys carved by preglacial rivers. Gradually, the regime of running water is replaced by flowing ice as the glaciers grow in size and expand in a down-valley direction. These "rivers of ice" modify the former river valley by ice erosion and produce profound changes in both the longitudinal and cross-sectional valley profiles.

Cirques, arêtes, horns, and cols. The upper reaches of a glaciated valley are very steep and commonly begin in

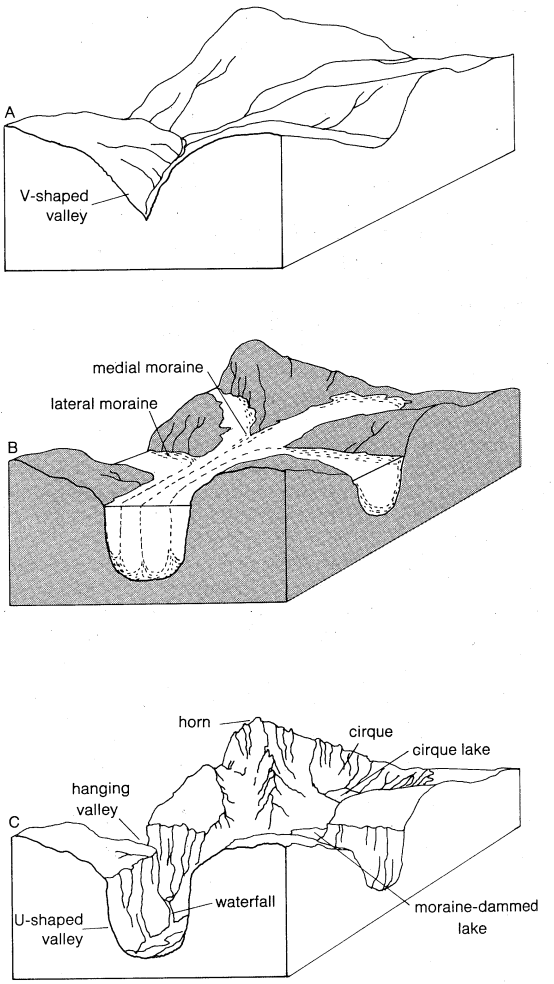


Figure 2: *Changes produced by valley glaciation.*
(A) Unglaciated topography, characterized by V-shaped valleys and rounded to rolling hills. (B) Area occupied by valley glaciers that produce deep erosion and down-valley transport of debris (moraines). (C) Area after glacier retreat: main valleys have become U-shaped; tributary, more shallow valleys now hang above main valleys, causing waterfalls; and upland terrain has become more rugged as a result of both frost action and glaciation.
From J. Zumberge, *Elements of Geology* (1963); John Wiley & Sons, Inc.

The process of cirque formation

a cirque, a rock-bound, amphitheatre-like feature that characterizes almost all headward reaches of glaciated valleys. The floor of a well-developed cirque contains a bedrock depression in which water may collect to form a lake. A cirque is universally regarded as a product of glacial erosion, but the actual mechanics of the process are uncertain. Studies on cirques that are still occupied by small glaciers indicate that frost action in the headwall region may be the most important process in the development of the cirque walls. The loosening of bedrock by this mechanism was originally thought to occur in the bergschrund, the opening between the glacier and the cirque wall. Temperature fluctuations in the bergschrund, above and below the freezing point, were supposed to have caused extensive frost shattering of the cirque wall. Blocks loosened from the rock wall of the bergschrund would fall to the bottom of the bergschrund and become incorporated into the glacier, which would transport them to the snout of the glacier over a period of years. This hypothesis is not supported by actual temperature measurements in the bergschrund itself. Such measurements in bergschrunds of Switzerland, Greenland, and Norway show that no appreciable freeze-thaw temperature fluctuations occur in bergschrunds to depths of almost 30 metres (98 feet). Only in very wide bergschrunds does the temperature fluctuate above and below the freezing point in response to diurnal or seasonal changes in air temperature. Water from melting snow or ice that enters these bergschrunds from above freezes to the rock face in the upper part of the bergschrund and acts as a protective covering for any further frost shattering. Some frost shattering occurs in the cirque walls above the bergschrund (*i.e.*, above the glacier surface), where rapid temperature fluctuations across the freezing point are commonplace. Frost-rived blocks produced there would tumble into the bergschrund or become lodged on the glacier surface; in either case they would be transported away from the cirque wall.

Frost shattering thus may be the dominant process in cirque-wall enlargement and retreat. Some frost-rived blocks undoubtedly are derived from the rock wall of the bergschrund, but the conditions for intense frost action are more favourable on the cirque walls above the glacier than in the bergschrund.

The excavation of the cirque floor is a more difficult problem to solve because glacier abrasion appears to be the chief means whereby the cirque floor is lowered. Glacial polish and striations on cirque floors confirm that abrasion did operate at one time during the development of the cirque basin. The possibility of periglacial conditions in the high mountains prior to the growth of glaciers, however, should not be overlooked. Intense frost action associated with such a climatic regime could have produced considerable frost rubble that would have been easily removable by glacial transportation at a later date. This explanation is incomplete, however, because of the difficulty in accounting for the rock sill on most cirque basins if glacial action is considered to be responsible only for removing preglacial frost rubble. This problem needs more study.

Studies of modern cirque glaciers (less than 300 metres [1,000 feet] long and about 50 metres [150 feet] thick), in which horizontal tunnels have been excavated, reveal a sliding action of the basal ice over the bedrock floors. This sliding action has an upward component in the lower reaches of the glacier where the bedrock deepening apparently is taking place. Although it should not be assumed that cirque glaciers move only by a sliding on their beds, there is sufficient motion of this type to account for considerable bedrock scour by glacial abrasion. Thus, in addition to removing any frost rubble or other regolith that formed prior to the onset of glacierization, a cirque glacier further deepens the bedrock by abrasion.

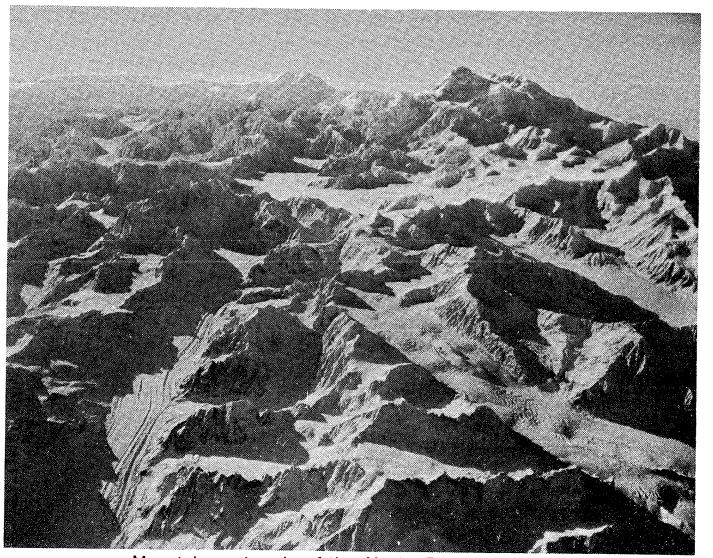
Headward extension of cirques and the formation of serrate divides

When a number of cirques on the flanks of a mountain range enlarge by headward extension, the sidewalls and the headwalls of adjacent cirques intersect. The headward erosion of cirques may continue after the glaciers contained in them have disappeared. Postglacial cirque enlargement is accomplished by continued frost action on the walls; the accumulation of frost rubble at the base of cirque cliffs bears witness to the continuation of the frost-shattering process even when no glaciers are present.

The headward extension of cirques, either during or after glaciation, produces the jagged ridges of glaciated mountains. These serrate divides can be observed in the Alps, Himalayas, Andes, and Alaska Range, and a number of other high mountains of the world. A single sharp crested ridge between the sidewalls of two cirques is called an arête. Two cirques that intersect on a major divide produce a low point on the divide called a col. Three cirques, each migrating headward toward a central point on a ridge produce a horn. The Matterhorn, near Zermatt in the Swiss Alps, is a classic example and the best known horn in the world.

Longitudinal profiles. Unlike valleys that are cut by streams, glacially sculptured valleys are characterized by a very irregular longitudinal profile consisting of alternating steps and rock basins. The true bedrock profile of a glacial valley may be masked or modified by various depositional features, but evidence from many sources indicates the geometry of the bedrock profile along the axis of a glacial valley.

The first inference drawn from the irregularity of longitudinal profile is that the efficacy of glacial erosion was not everywhere uniform, or that the bedrock forming the valley floor was not uniform in its susceptibility to glacial erosion. The question is whether a rock basin was



Mountainous terrain of the Alaska Range produced by glacial erosion and frost action.
By courtesy of the Air Force Cambridge Research Laboratories, Bedford, Massachusetts

produced where the glacier had an intrinsically greater capacity to erode, or where rocks were more susceptible to glacial erosion than those in which no basins occur.

There is reason to believe that both explanations account for the alternating rock steps and rock basins that characterize the longitudinal profile of a glacial valley. The analysis of flow characteristics in existing valley glaciers reveals two general types of flow, extending flow and compressional flow. Both may occur in different reaches of the glacier stream. Extending flow occurs in the steeper reaches and compressional flow is restricted to the flatter reaches. Extending flow is characterized by shear planes in the ice that curve downward and approach the bedrock floor tangentially. In compressive flow, the shear planes curve tangentially upward from the base of the ice toward the glacier surface in a downvalley direction. Compressive flow is conducive to the removal of bedrock material, whereas extending flow tends to force rock debris along the bedrock floor. Thus the glacier would tend to be a more efficient agent of bedrock erosion in areas in which compressive flow was dominant.

A glacier flowing in a valley inherits the longitudinal profile produced by the preglacial stream. This profile has a number of irregularities that are accentuated by glacial erosion. Small declivities in the original preglacial profile become enlarged through compressive flow as previously described, until rock basins are formed.

Differential rock resistance and weathered zones

Another factor to consider in explaining the sequence of rock basins and intervening steps and sills is the differential resistance of bedrock types that occurs along the valley. Even if the bedrock is of uniform composition, such as a granite, differences in the spacing of joints are reflected in the depth of preglacial weathering. Closely spaced joints permit deeper penetration of downward percolating surface water, which weakens and loosens mineral grains by chemical decomposition, and also enhance the efficacy of frost action prior to glaciation. Preglacial deep weathering along a valley thus guides the future course of glacial erosion. The more deeply weathered zones are the places at which rock basins are excavated by glacial ice, and the less deeply weathered zones become the intervening rock sills or "glacial steps."

The basin-and-step profile of a glacial valley is therefore controlled by at least three factors, the minor irregularities in the original stream-valley profile, differential preglacial weathering, and the action of extending or compressive flow of the glacier ice.

Valley cross sections. Stream-cut valleys normally have a V-shaped cross section, whereas glaciated valleys are U-shaped in cross section. The direct erosive action of a river is confined to the apex of the valley in which it flows. The stream of flowing water occupies only a relatively small segment of the valley. A stream of ice, on the other hand, may be in direct contact with more than half the cross-sectional perimeter of the valley, thereby putting flowing ice in direct contact with valley walls that were never directly in contact with the preglacial stream. Glacial erosion on the valley walls will, over a period of time, transform the V-shape into a U-shape.

V-shaped stream valleys are characterized by spurs that develop between adjacent entering tributaries or merely as a consequence of the curving course of the river as it incises its channel deeper. These spurs are truncated by glacier flow because an ice stream tends to straighten the inherited course of the preglacial stream valley. Vigorous ice erosion takes place not only on the valley floor but on the valley sidewalls as well.

Valley glaciers commonly occur as systems in which a master ice stream is fed by a number of tributary glaciers. The main ice stream and its tributaries form an ice-drainage pattern that is generally inherited from a former stream-drainage pattern. Under normal conditions of stream erosion, tributaries join the trunk stream at the same level. This "law of accordant junctions" was formulated long ago by geologists who recognized the regularity of the juncture of a tributary and channel it joined. Because the main glacier normally carries a much larger volume of ice than any of the tributary glaciers, it will have proportionally greater erosive power and will cut the trunk valley deeper than the inflowing tributary glaciers. After the glacier episode of erosion is terminated, and the glaciers are replaced by streams, the junction of tributary valleys with the main valley will no longer be at accordant elevations. The tributary valleys are left "hanging" above the floor of the trunk valley. Streams that occupy the hanging valleys plunge over steep cliffs where they debouch into the overdeepened main valley. Spectacular waterfalls in many formerly glaciated regions originated in this way.

Fjords. The mountainous coastal regions of the world that lie poleward of 45° north or south latitude contain steepwalled, long, and deep embayments known as fjords. Fjords abound on the Pacific coasts of southern Alaska and British Columbia, and the Baffin Bay coast of Baffin Island; in Greenland, Iceland, and Norway; in the Southern Hemisphere, the Chilean coast of Patagonia, South Island of New Zealand, and the coastal regions of Antarctica contain fjords. Many of those in Antarctica are still filled or partially filled with glacier ice.

Longitudinal profile of fjords

The longitudinal profile of a typical fjord reveals a very steep landward segment, the headwall, that descends rapidly to below sea level. From the inner headwall to the seaward entrance of the fjord, the longitudinal profile is similar to the profile of a glaciated valley because it contains deep rock basins separated by rock sills or thresholds. From the mouth of the fjord seaward, the profile rises to a rock platform or strandflat that is relatively shallow compared with the water depths over the inner rock basins. The great Sognafjorden of southern Norway, for example, is 204 kilometres (127 miles) long and reaches a maximum water depth of 1,300 metres (4,300 feet) behind the strandflat, which is only about 200 metres (650 feet) below sea level.

Fjords are generally considered to be glacially modified river valleys. The longitudinal profile of a fjord, however, is a feature that cannot be accounted for by river erosion because of the deeply submerged rock basins and intervening rock thresholds. The rock basins are too deep to be explained by stream erosion during periods of lower sea level because there is no independent evidence of sea level lowering of this magnitude. Moreover, rivers are incapable of excavating long and deep rock basins along their courses. Therefore, it is concluded by many geologists that glacier ice, not river water, was responsible for the creation of fjord profiles.

This explanation is challenged by a few geologists because the occurrence and distribution of many fjord systems appear to be related to major structural trends in the bedrock of the fjord regions. Intersecting fracture zones, for example, can account for the rectilinear pat-

tern of fjord distribution in many coastal regions. A few geologists have argued the extreme viewpoint that because the fjord axes coincide with linear joints, fractures, or faults in the bedrock, fjords are therefore of tectonic origin, *i.e.*, they represent primary downfaulted blocks produced by tensional stresses in the earth's crust. Although some rectilinear fjord systems do, in fact, reflect an intersecting system of faults or other linear tectonic features, there are many fjord patterns that bear no apparent relationship to such crustal lineaments. Furthermore, the fact that some fjord systems are aligned with structural elements of the crust does not necessarily prove a tectonic origin of the fjord troughs. A more likely explanation is that fault zones, closely spaced joints, and soft tilted rock layers in juxtaposition with hard layers would have been normal zones of weakness that were more vulnerable to the forces of rock weathering than the more massive intervening rock masses. Differential weathering and stream erosion during the millions of years prior to Pleistocene glaciation would have resulted in an etching of the landscape into a pattern controlled by the linear zones of more easily weathered and eroded materials. These preglacial valleys would then have been the natural troughs for the channelling action of valley glaciers during the Pleistocene.

Other critics of the glacier-erosion hypothesis for the origin of fjords argue against the efficacy of valley glaciers to erode such deep basins below sea level. They suggest that these rock basins were formed by valley glaciers in the early Pleistocene when the coastline stood relatively higher above sea level, and before the ice reached a thickness sufficient to cause depression of the crust. The rock basins in the fjords, according to this view, are submerged at such great depths because postglacial crustal uplift is not yet complete, and also because the level of the oceans is higher now than it was during the glacial maximum. Both explanations are true in essence, but neither separately nor in combination can they account for the maximum known depth of the submerged rock basins in fjords. Assuming that sea level is now about 90 metres (300 feet) higher than it was during the glacial maximum, and assuming further that the total amount of crustal depression in the fjord region of Norway was about 605 metres (1,980 feet), of which perhaps 450 metres (1,500 feet) have been recovered by late Pleistocene and postglacial uplift, it can be deduced that only about 240 metres (790 feet) of the present water depth of Sognafjorden is due to these factors (90 metres [295 feet] of sea-level rise plus 150 metres [492 feet] of glacial rebound still to occur). Using the same values for sea-level changes and the amount of total crustal subsidence due to the weight of ice, the deepest part of Sognafjorden would have been about 1,700 metres (5,600 feet) below the Pleistocene sea level. In order for a glacier to rest on the floor of the ocean in this depth of water it would have to have been about 1800 metres (5,900 feet) thick.

Fjord formation and ice thickness

The reasonableness of this analysis is sustained by geophysical study of the Skelton Glacier near McMurdo Sound in Antarctica at its confluence with the Ross Ice Shelf. The Skelton Glacier is in the process of eroding a deep fjord and the longitudinal profile of the fjord shows the typical basin and threshold configuration of other fjord profiles. The greatest known thickness of ice along the 62-kilometre (39-mile) length of the Skelton Glacier occurs at a point about 50 kilometres (30 miles) back of the floating snout. There the glacier ice is about 1,450 metres (4,760 feet) thick and is in contact with the sea floor, which at that point is over 1,240 metres (4,070 feet) below sea level. Because it is known that the glaciers in Antarctica were thicker at some time in the geologically recent past, the Skelton Glacier could have attained a thickness equal to or greater than the postulated thickness of the glacier that occupied Sognafjorden during the Pleistocene.

GLACIALLY ERODED ROCK BASINS IN NONMOUNTAINOUS REGIONS

The bedrock basins in fjords and other glaciated valleys tend to be elongated in the direction of glacier flow. This relationship has led geologists to assume that elongate bedrock basins in other regions of the glaciated part of the world were aligned parallel to the flow lines of the glacier. A detailed reconstruction of flow lines based on striae, grooves, and other glacier markings on bedrock in areas where rockbound lakes exist by the thousands shows that the orientation of the elongate basins bears no consistent relationship to the direction of glacier movement.

Bedrock lakes of glacial origin occur in very great numbers in areas of relatively low relief such as the Canadian Shield of North America, the nonmountainous parts of Sweden and Finland around the Gulf of Bothnia, and some parts of Scotland and Northern Ireland. These lake basins occur in many different shapes and sizes, both in the configurations of their shorelines and in their bottom topography. The depths of many of these lakes exceed 100 metres. Seneca Lake, one of the Finger Lakes of New York, for example, has a maximum depth of 192 metres (633 feet).

Prior to the growth of the continental ice sheets in North America, Europe, and northern Asia, the land surface was exposed to other agents of denudation. Exposed rock masses were weathered by chemical breakdown and physical disruption, and the comminuted rock fragments were carried by surface runoff to river channels that transported the sediment to the ocean basins. The length of time during which the forces of weathering and erosion were at work on the land area that was later covered by Pleistocene ice sheets varied from place to place. The geological record in all of the ice-covered areas of North America suggests that the period of downwearing of the earth's surface prior to the Pleistocene was about 60,000,000 years or longer. This was ample time for a well-developed stream-erosional landscape to be produced. The preglacial landscape was therefore one of valleys and divides with few, if any, lakes. The spread of glaciers over this kind of topography continued the wearing away of the land surface by exploiting the areas of deeply weathered bedrock. Most rockbound lakes within the Pleistocene glacier border, therefore, were probably formed by glacial excavation. As with the origin of fjords, however, not all geologists agree that glacial erosion has the efficacy to produce rock basins the size of the Great Lakes.

The Great Lakes of North America

The five Great Lakes of Superior, Michigan, Huron, Erie, and Ontario constitute the largest inland system of freshwater in the world. Their combined area is about 246,000 square kilometres (95,000 square miles), and all but Lake Erie have a maximum depth well below sea level. Geologically, the lake basins lie in bedrock although the shorelines are developed mainly in glacial drift. An exception is Lake Superior, the shoreline of which is almost entirely rockbound.

Since the late 19th century, geologists have proposed four processes to explain the origin of these lake basins: (1) stream erosion during the Tertiary Period (from 2,500,000 to 65,000,000 years ago); (2) glacial erosion during the Pleistocene Epoch; (3) damming by glacial deposits; and (4) large-scale crustal movements (tectonism). Over the years, some geologists have taken extreme views on one or more of these hypotheses. Some have argued that the excavation of such large basins by glacier ice is not possible, whereas others have held that glacier erosion has been the only agent in the creation of the basins.

The final answer is yet to be found, but any hypothesis must take into consideration the following facts: first, that glacier ice invaded all the Great Lakes is not questioned by any scientist familiar with the Pleistocene history of east central North America; second, the Great Lakes did not exist prior to the Pleistocene; and, third, because the long axes of the lakes are more or less parallel to the major structural trends of the bedrock in which they lie, the present sites of the basins were occupied in pre-Pleistocene times by a system of stream valleys that were adjusted to the regional structural elements of the bedrock trends.

Geophysical studies and bottom drilling by shipboard techniques in the 1960s in Lake Superior showed that glacial sediments many metres thick lie on the bedrock

floor of the lake. Glacial deposition rather than glacial erosion must have been operative at that time in the history of the basin. One geologist has even suggested that the Lake Superior Basin may have been an ancient rift valley, similar to those in Africa, which was later modified by the passage of glacier ice.

Still unanswered is the basic question of how much glacier erosion added to the size and depth of the basin. It cannot be proved beyond a reasonable doubt that Lake Superior or any of the other Great Lakes was excavated wholly by glacial erosion.

III. Landforms produced by glacial deposition

Materials that are picked up by eroding glaciers are transported by the moving ice and deposited at the base of the glacier or along its margin. Meltwater streams debouching from glacier margins carry sediment beyond the ice front and deposit it in a variety of forms (Figure 3). Ice-marginal lakes and even the oceans receive sedi-

From J. Zumberge, *Elements of Geology* (1963); John Wiley & Sons, Inc.

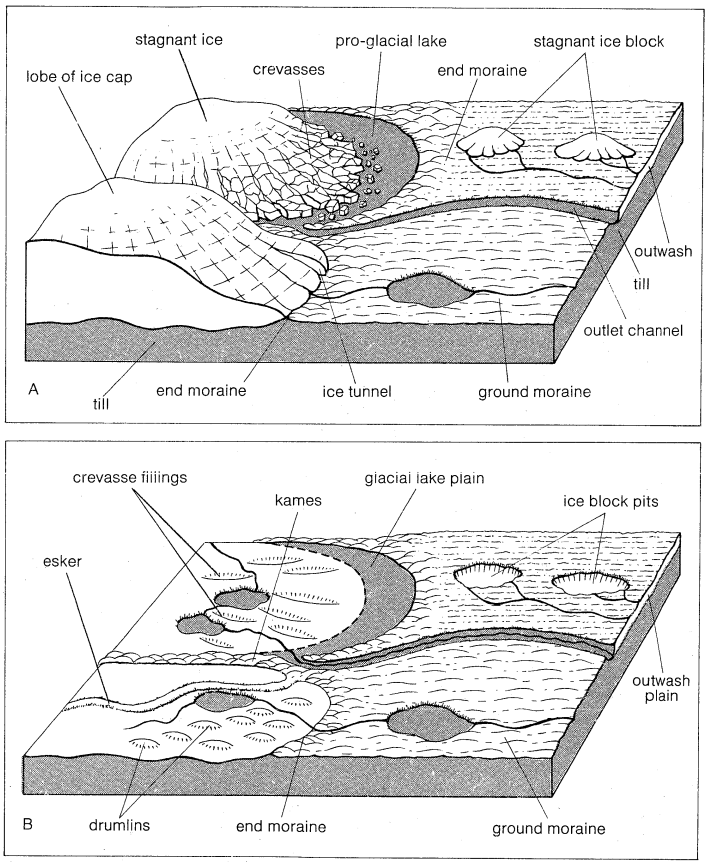


Figure 3: *Materials and landforms associated with glacial deposition.*
(A) Conditions during retreat of ice sheet. (B) Glaciated landscape after ice has disappeared.

Characteristics of glacial sediments

ments from melting icebergs as well as from glacial meltwater streams. All of these sediments, regardless of whether they are deposited directly by glacier ice, or whether they accumulated in meltwater river channels or ice-marginal lakes or seas, are collectively known as drift. This term is based on the mistaken assumption by geologists of the 19th century that the loose and unconsolidated surficial deposits of England and northern Europe, now known to be of glacial origin, were deposited in the universal seas of the Deluge related in the Biblical story of Noah. The term drift was used because the heterogeneous nature of some of the surficial materials—especially those that contained huge boulders—was explained as having been released from drifting icebergs that supposedly infested the higher latitudes of the worldwide oceans during the biblical Flood. This explanation is now known to be false, but the term has remained in usage to the present day.

Drift is classified into two major categories, nonstratified and stratified. The former implies the lack of any transporting agent capable of sorting a heterogeneous mixture of sediment into layers of gravel, sand, silt, or clay. The term stratified designates drift that was originally carried by glacier ice, but was subsequently deposited in layers of sorted sediments either along stream channels, in lake basins, or in parts of the ocean. Both nonstratified and stratified glacially derived sediments occur in geometric forms that are characteristic of a landscape produced by glacial-depositional processes. These processes are constructional in that landforms produced by them are "built-up" by an accumulation of materials, whereas glacial erosion is a destruction process and the landforms produced are destructional in origin.

NONSTRATIFIED DRIFT AND ASSOCIATED LANDFORMS

The main feature of nonstratified drift is the heterogeneity of its texture. A mixture of detrital particles ranging in size from clay, less than four microns (.004 millimetre) in diameter, to boulders, larger than 256 millimetres (ten inches) in diameter, characterizes many nonstratified drift deposits. The term till refers to the sedimentary materials of a nonstratified glacial deposit, and the term moraine is used with reference to its topographic expression. In Germany, Scandinavia, and some other parts of Europe, till and moraine are used synonymously, but in North America and the Soviet Union, till refers to the sedimentary material, and moraine refers to the surface morphology of a mass of till.

Glacial till. Widespread deposits of till occur where glaciers once covered the land. Tills laid down by Pleistocene glaciers are generally compact, clay-rich deposits that can be loosened with a pick and shovel, or excavated by earth-moving equipment. Tills deposited during older glaciations are indurated into hard rock. These ancient glacial sediments are called tillites and their distribution in space and time on the earth's surface provides a means for reconstructing the climatic conditions during past geologic time.

Unsorted deposits of nonglacial origin

Not all unsorted, heterogeneous sediments can be ascribed to glacial origins. The debris that accumulates as the result of a rock avalanche, a mudflow, or the poorly sorted sediments associated with flash floods or other fluvial activity are commonly mistaken for tills, or ancient tillites. Moreover, the absence of large boulders does not preclude an otherwise "layerless" sediment from being a nonstratified glacial deposit. Geologists who have a special interest in the study of tillites use the term diamictite to describe any nonstratified rock that has the characteristics of a till but the origin of which may not have had any association with glaciation.

Two kinds of till are generally recognized, lodgment till and ablation till. Lodgment till is the material released from the base of a moving glacier by pressure melting, whereas ablation till is a residual mass of debris released from the upper portions of the glacier ice in the ablation zone. Lodgment till is compact and dense and can occur in blanket thicknesses as great as 100 or more metres. Ablation till is loose, less compact than lodgment till, and generally much thinner than the lodgment till on which it lies.

Both lodgment and ablation tills contain rock fragments that differ in lithology from the local, underlying bedrock. All pebbles, cobbles, stones, and larger rock fragments contained in till are called clasts. Most clasts are derived from local bedrock sources, but some have been transported as far as 800 kilometres (500 miles) or more by glacial movement. A large clast that differs from the underlying bedrock is called a glacial erratic. An erratic whose place of origin can be identified can be used to indicate the general direction of ice movement. So-called indicator erratics normally consist of rock types that contain an unusual assemblage of minerals or possess other characteristics that make the geographic location of bedrock origin unassailable.

Some erratics are so large that they are mistaken for bedrock outcrops. The longest dimension of some of the largest ones exceeds 30 metres (100 feet), and they stand

above the general level of the glaciated landscape as conspicuous features.

Moraines. Many different types of moraines are recognized, each of which has some genetic implication in terms of its specific mode of glacial origin. The two main types of moraines are ground moraine and end moraine. Ground moraine is a blanket of till of varying thickness laid down beneath a moving glacier. Its chief morphologic characteristic is an undulating or gently rolling surface, commonly referred to as swell-and-swale topography. Surface drainage on ground moraine deposited by the most recent ice advances of Pleistocene glaciers in North America, Europe, and the Soviet Union is poorly integrated with many undrained swampy depressions and shallow lakes. The maximum relief of ground moraine is about six to eight metres (20 to 26 feet).

If the ground moraine is thicker than the maximum relief of the underlying bedrock surface, its topographic expression is generally independent of the topography on which it lies. Preglacial stream valleys may be buried to such an extent that their existence is not recognizable, except where occupied by chains of lakes formed from the melting of blocks of stagnant glacier ice.

Older ground moraine, deposited by earlier advances of Pleistocene continental glaciers, retains few of the original constructional features because of the action of surface runoff. The swamps and lakes have either been filled with muck, peat, and sediments washed in by streams, or drained by the downcutting of the outlet. The network of surface streams and tributaries is well integrated in the older ground moraine areas. These stream systems also erode the land surface, thereby destroying the original topographic expression of the ground moraine.

End moraines and their implications

End moraines are formed along the margins of a glacier and consist of belts of hummocky terrain that persist long after the glacier has retreated or disappeared. A glacier acts as a conveyor belt in which sediment is carried to the glacier margin where it is dumped. The formation of an end moraine requires an actively moving glacier with an adequate supply of sediment. Because of the meltwater discharge from a glacier margin, it is not uncommon for end moraines to contain considerable quantities of stratified drift, the topographic forms of which are described in the next section of this article.

The presence of an end moraine carries with it the implication that the ice margin was more or less stationary while the end moraine was constructed. A large end moraine implies a long period during which the glacier margin remained in one place, and a poorly developed end moraine of low relief implies a short period of ice-margin stability.

During general phases of retreat of glaciers, a number of successively younger, more or less parallel end moraines are built. Though these "recessional moraines" could represent a pause in the general shrinkage of the ice sheet or valley glacier, it is more likely that they reflect a series of readvances of the glacier of short duration. Composite end moraines are formed when a glacier retreats from a position marked by an end moraine, and then readvances at a later date to build another end moraine against the former one, or to completely override the older feature and deposit a new load of till on top.

End moraines plotted on a map of suitable scale provide the means of reconstructing the position of the glacier border as it existed when the moraines were built. The mapping of contemporaneous end moraines in areas formerly covered by the last advance of the Pleistocene ice sheets reveals that the glacier margin was lobate in form. In the Great Lakes region, several lobes were deployed along the axes of the modern Great Lakes. This relationship is clearly indicated by the festoon of end moraines around Lake Michigan, Lake Huron, and Lake Erie.

The area in which two lobes were in juxtaposition is now marked by adjacent end moraines built by each lobe. Such a moraine is called an interlobate moraine. One of the best known examples is the Kettle Moraine of Wisconsin, which was produced by the Michigan lobe and the adjoining Green Bay lobe.

A valley glacier builds an end moraine at its snout or terminus, and a lateral moraine along the glacier sides in contact with the valley walls. Two lateral moraines joining at the confluence of two valley glaciers combine to form a medial moraine. End moraines and lateral moraines are commonly preserved in formerly glaciated valleys. These old lateral moraines and their end moraine equivalents provide the basis for reconstructing the history of glacial advances and retreats in the Alps, the Rocky Mountains, and other mountainous regions of the world that contain shrunken remnants of extensive glaciers or have no modern glaciers at all.

Drumlins. A drumlin, ideally, is an elongate streamlined hill with a long axis that is parallel to the direction of ice movement. Commonly, drumlins occur in swarms or fields that give a definite "grain" to the topography. Individual drumlins are from less than a kilometre to several kilometres long, 300 to 600 metres (1,000 to 2,000 feet) wide, and five to more than 30 metres (ten to 100 feet) high. A single drumlin field may contain several thousand drumlins.

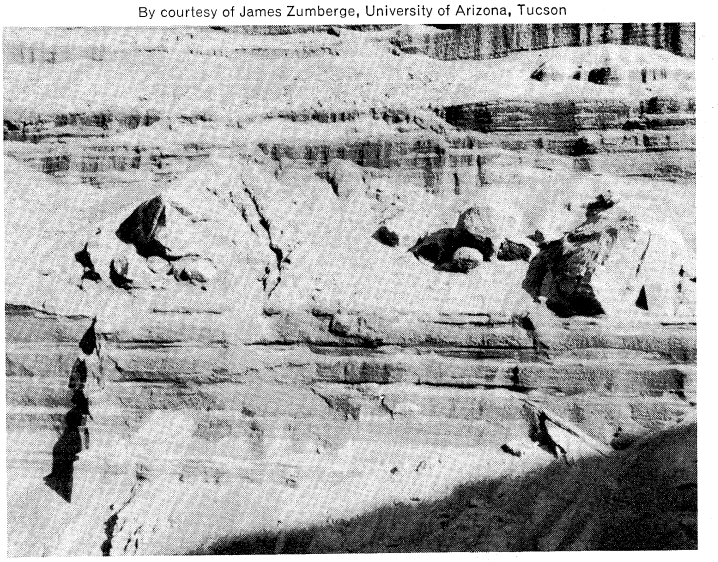
Drumlins are thought to be "ice-molded" forms and range in composition from 100 percent till to 100 percent bedrock. The latter are generally considered to be erosional in origin and are called rock drumlins. The till drumlins on ground moraine are shaped by the overriding ice in a manner not clearly understood. These are usually regarded as depositional or constructional in origin. Some drumlins contain bedrock cores with a till veneer.

STRATIFIED DRIFT AND ASSOCIATED LANDFORMS

The role of meltwater

Most glaciers, with the notable exception of the Antarctic Ice Sheet, lose most of their masses by melting in the ablation area. During the summer melt season, the glacier surface is characterized by meltwater channels in the ice. The water released by melting eventually reaches the glacier front through surface runoff, or discharge through subglacial tunnels. As the ice melts, the sedimentary load entrapped in the ice is released and redistributed by the meltwater streams in various ways. The general term, outwash, is applied to the sand and gravel fraction of glacially derived, water-laid sediments.

By courtesy of James Zumberge, University of Arizona, Tucson



Glaciofluvial sediments in southern Norway. Boulders are from an overlying till not pictured.

In valley glaciers the meltwater flows from the glacier snout down the nonglaciated part of the valley, carrying with it a heavy load of outwash, which is deposited along the many channels of the river. If the volume of debris released by the melting glacier is more than the river channel can transport, the outwash is spread over a large segment of the valley floor in a number of anastomosing channels through the outwash materials. These channels join, separate, and shift their positions as the volume of

water released from the melting glacier changes from time to time. The term braided stream is applied to a river with the aforementioned characteristics. Sediment deposited by a glacially-fed braided stream, called a valley train deposit, consists of stratified sands and gravels.

During the retreat of a continental glacier, several adjacent braided streams originating from the ice margin may deposit glacial outwash over an extensive area beyond the ice front. These flat plains of stratified materials are called outwash plains. Stagnant masses of glacier ice that become detached and isolated from the main glacier during the course of ablation and retreat may be buried or partially buried in outwash. Later, the melting of these ice blocks causes the outwash over and around them to collapse and slump into the resulting depression. The depressions are called ice-block pits and may persist for thousands of years after they are formed. An outwash plain pockmarked with ice-block pits is called a pitted outwash plain. Thousands of lakes in the glaciated regions of the world originated as ice-block pits.

Ice-contact deposits. Outwash that accumulates in contact with glacier ice is an ice-contact deposit. Some of these deposits occur as distinct morphological features, whereas others are so small or so intermingled with till that they have no recognizable surface expression.

Glacier margin conditions

The conditions around the margin of a glacier during the height of the ablation season are chaotic. Running water, mud slurries, slumping mounds of debris, stagnant blocks of ice, and quiet pools of water coexist within a very small area along the ice margin. Conditions are constantly changing, not only from week to week and day to day but from midday to midnight, as the volume of meltwater waxes and wanes with the diurnal changes in temperature, variability of incoming and outgoing solar radiation resulting from changes in cloud cover, and the precipitation of rain or snow. Glaciofluvial sediments that accumulate in the ice-marginal zone reflect the local environment of deposition. Fine silt and clay settle into pools, poorly sorted sand and gravel are almost ubiquitous, and ablation till and large boulders are widespread. Layers of glaciofluvial sediments that abut against an ice wall will be disturbed or destroyed when the adjoining ice wall melts.

Disjointed and distorted layering of glaciofluvial sediments is thus the chief depositional characteristic of ice-contact deposits. Ice-contact deposits are also identified by the abrupt vertical change in grain size of the different layers, a condition that reflects the rapid change in the depositional conditions.

The ablation zone of a valley glacier extends upvalley some distance from the glacier terminus. Meltwater streams issue not only from the toe of the glacier but also from along the sides of the glacier. These ice-marginal streams flow in a channel defined by the valley wall on one side and the sloping glacier surface on the other. Sand and gravel that accumulate in such an ice-marginal channel form a bench of outwash parallel to the valley wall. Retreat of the glacier results in the slumping of that part of the sand-and-gravel bench that was formerly supported by the ice. The geomorphic feature thus produced is a kame terrace, and if it contains ice-block pits, it is a pitted kame terrace. The collapsed side of a kame terrace has a characteristic slope equal to the angle of repose of the material of which it is composed, generally about 30°. This slope is called an ice-contact slope.

Kames and eskers. A kame is a mound or hill composed of ice-contact glaciofluvial sediment. A kame forms either as a mass of outwash that accumulates in a zone surrounded by stagnant ice, or as a cone of outwash banked against an ice slope or wall. Individual kames may be very irregular in shape, depending on the configuration of the ice walls that contained them. Kames occur in clusters or groups that stand above the general ground level. They are commonly associated with end moraines and are not easily distinguished from individual knobs or hillocks of till with which they occur. Ice-block pits may occur in close association with kames. A landscape dominated by many kames and ice-block depressions is called kame-and-kettle topography.

An esker is a long ridge composed of sand and gravel. Some eskers are only 100 metres (330 feet) in length, whereas others attain lengths of 150 kilometres (90 miles) or more. In plan view, an esker follows a sinuous course not unlike a stream channel; this fact has led to the generally accepted hypothesis that eskers are outwash-filled ice tunnels that formed in stagnant or nearly stagnant ice of the ablation zone. Eskers may attain a height of 100 metres, but most are seven to 15 metres (23 to 49 feet) high. They tend to occupy the lower ground of a glaciated landscape but do cross divides from one swale or valley to the next. The glacial meltwater flowing in an ice tunnel is analogous to water flowing through a pipe under pressure rather than water flowing in an open channel. Some of the longer eskers have tributaries, indicating that an extensive system of subglacial tunnels was present when they were formed.

The ice-tunnel hypothesis

The chief material in an esker is glacial outwash, mostly well-washed sand and gravel. Cross-bedding is visible where eskers have been breached by a road cut or where they have been exposed through exploitation of their gravel content. Some eskers are capped by a thin layer of ablation till, which is usually patchy and discontinuous. Others contain numerous boulders that presumably are residuals from the previous ice cover. The esker flanks are ice-contact slopes, a fact that accounts for their general steepness.

GLACIOLACUSTRINE SEDIMENTS AND ASSOCIATED LANDFORMS

Meltwater that accumulates along a glacier border is called a proglacial lake, one in which part of the shoreline is formed by glacier ice. When the glacier retreats, the proglacial lake may shrink in size or its waters may be drained completely. In either case, the area formerly submerged can be identified because features of the former lake basin are preserved. Among these are the shorelines composed of beach sands and gravels that fringed the old lake basins. These ridges may occur in a series of parallel ridges at different elevations, each of which reflects a different level of the proglacial lake during the deglacial history of the glacier that formed part of its shoreline. Old beach ridges parallel to the coastlines of the Great Lakes mark the shorelines of the precursors of these water bodies. These ancient beaches became natural trails for Indians, and even today, secondary roads and major highways follow their crests as a legacy of early transportation routes established by 19th-century settlers.

The floors of the former proglacial lake basins are now flat lake plains. In some places these plains have been trenched by streams so that the nature and composition of the old lake sediments can be observed directly. These glaciolacustrine deposits are layered and well-sorted, and reflect the different depths of water in which they were deposited. In the nearshore zone, sandy material predominates, and in the deeper parts of the old basins, silts and clays are the dominant grain sizes. Clasts of pebbles and even boulders are commonly found embedded in the beds of clay and silt. The most plausible explanation for their presence in fine-grained lake sediments is that they were rafted in by ice, either icebergs or drifting slabs of lake ice that were detached from the shore during the spring breakup.

Other glacial lacustrine features include deltas built by streams flowing directly off the ice or by the normal accumulation of detrital sediment deposited by other streams emptying into the lake. The back or proximal side of a delta built by the discharge from a superglacial channel or a subglacial tunnel may have an ice-contact slope similar in steepness to the ice-contact slopes on kames and eskers. The surface of the delta, however, tends to be flat, and the lakeward or distal edge of the delta is characterized by a gentle slope that merges with the lake plain.

IV. Landforms produced by periglacial processes

The periglacial environment is generally confined to areas of high latitudes or high altitudes in which the

freeze-thaw process, sometimes called frost action, is dominant. Most workers agree that most of the periglacial zone is coincident with the areas of the earth underlain by permafrost. More than three-fourths of Alaska, half of Canada, and most of Siberia are underlain by permafrost, and hence are characterized by periglacial conditions. Smaller periglacial areas exist in high mountainous regions where permafrost may be absent but where strong freeze-thaw action prevails nonetheless.

PERMAFROST

Characteristics and occurrence of permafrost

Permafrost (also known as perennially frozen ground) is a surficial layer or zone of naturally occurring material the temperature of which has been at or below 0° C (32° F) continuously for more than two years. Water may be present either in the frozen state or in a supercooled condition (*i.e.*, in liquid form at temperatures below freezing), or the permafrost may be dry. Most permafrost contains ice in one form or another, the most common occurrences being in the form of pore ice, segregated pure ice lenses, or ice wedges. Pore ice is ice that fills the pores and interstices of rock or soil, just as water occupies the pores of natural materials in nonpermafrost areas. Ice segregations are seams or irregular-shaped pure ice masses that form by the drawing of water into the permafrost layer from below, thereby increasing the total bulk moisture content of the permafrost zone. Ice wedges are ice masses that occur as wedge-shaped sheets with a vertical or inclined attitude. Their thickness ranges from less than 30 centimetres (about one foot) to more than three metres (ten feet) and they possess an internal foliation that is more or less parallel to the boundaries of the ice wedges. Ice wedges form in only the most vigorous periglacial climates, those in which the mean annual air temperature is −7° to −8° C (19° to 18° F) or colder.

The permafrost zone is divided into two broad geographical units, the continuous zone and the discontinuous zone. In the Northern Hemisphere the southern boundary of the discontinuous zone is roughly parallel to the 0° C (32° F) mean annual surface temperature isotherm. From this boundary northward, the permafrost layer increases in thickness and continuity until, at the boundary between the discontinuous and continuous zone, permafrost exists everywhere except beneath lakes and rivers that do not freeze to the bottom in winter.

The permafrost layer is bounded on the top by the permafrost table that is defined by the depth of summer thaw. The base of the permafrost is determined by the mean annual temperature of the region, seasonal snow cover, thermal conductivity of rock and soil, vegetational cover, and past climatic conditions. The base of the permafrost in northern Siberia is 1,500 metres (5,000 feet) below the surface, 600 metres (2,000 feet) in northern Alaska, and it becomes progressively shallower toward the southern boundary.

The layer between the permafrost table and the ground surface is called the active layer, and it is here that frost action is most vigorous. The active layer is less than 30 centimetres thick in the northern areas of continuous permafrost, and more than three metres thick near the southern boundary of permafrost.

In nonpermafrost regions, especially at high elevations, frost action results in the creation of rock rubble —accumulations of coarse, angular boulders and smaller particles that are loosened from a cliff or rock escarpment by the freeze-thaw process. Water that accumulates in cracks or joints in an exposed bedrock surface, exerts a strong wedging action during the freezing process. This frost-wedging process is responsible for the origin of many distinct landforms associated with the mountainous regions of the world, regardless of the presence or absence of permafrost.

TALUS, ROCK GLACIERS, AND BLOCKFIELDS

Debris formed by frost wedging along a steep rock wall, such as might exist in a glacially carved valley, accumulates at the base of the cliff as talus. If the cliff face is a uniformly straight rock wall without indentations or notches, the talus forms a uniform apron at the cliff base. The inclination of a talus apron or talus slope is determined by the angle of repose of the constituent talus fragments. The largest and most angular blocks produce talus aprons with the highest slope angles. These exceed 35° (measured from the horizontal) in some cases, but slope angles of about 30° are more common.

Along rocky cliffs that contain notches or indentations in the form of rock chutes, the talus blocks accumulate as talus cones. Several adjacent talus cones may coalesce to produce a compound talus apron, which unlike the simple talus apron, has an undulating surface because the talus originates primarily from the chutes instead of uniformly along the entire cliff face.

Physical properties and flow of rock glaciers

A talus cone that contains interstitial ice as a cementing agent will flow and become a rock glacier. Some rock glaciers are thought to be relicts of true valley glaciers in which the superglacial debris has become so concentrated because of thinning of the glaciers that no glacier ice is visible at the surface. Rock glaciers that originate in cirques may have originated in this way, but others appear to bear no relationship to previous valley glaciers. In plan view, a rock glacier is a tongue or lobe of angular rock debris that shows signs of movement similar to a true glacier. Surface ridges parallel to the edges of the rock glacier, and terminal hummocks and ridges, not unlike end moraines, are manifestations of glacier-like movement. Surface movement studies on rock glaciers indicate velocities of about one metre or less per year. These are less than velocities of true glaciers because of the higher proportion of rock debris in the rock glaciers.

Rock glaciers range in width to more than 300 metres (1,000 feet) in length to one kilometre, and in thickness to 30 metres (100 feet) or more. Excavations in several rock glaciers have revealed the presence of interstitial ice within three metres (ten feet) of the surface. Rock glaciers are active (moving) or inactive (not moving). The latter are clearly distinguishable by the presence of lichens and other vegetation on the surface and marginal slopes.

Frost rubble that accumulates on a level or gently undulating surface forms a blockfield, also known as a felsenmeer. The blocks are derived from frost heaving of the underlying bedrock. The size and shape of an individual block is determined by the spacing and orientation of joints in the bedrock. Most of the detritus in a blockfield is very angular, a criterion that distinguishes blockfields from ablation till. Blockfields are actively being produced today under the periglacial environment of high mountains or polar regions, but others are relicts of a past periglacial climate and are covered with lichens, moss, turf, or higher vegetational forms.

PATTERNED GROUND

Ice-wedge polygons. Areas underlain by permafrost exhibit a characteristic microrelief pattern consisting of a network of intersecting ridges or troughs that outline polygonal ground areas from about three metres (ten feet) to more than 30 metres (100 feet) in diameter. Polygonal ground is conspicuous in the tundra areas of Alaska, Siberia, northern Canada, and other polar areas.

Formation of polygons

The ridges or troughs that outline the polygons are caused by the presence of ice wedges, which underlie them. Ice wedges are formed when the ground contracts during the cold winter months of the extreme periglacial zones. Thermal contraction produces a polygonal network of cracks that fill with water from melting snow in the spring. Because the subsurface temperature is below freezing, the water in the cracks freezes to form an ice vein. Summer warming causes the ice vein to expand slightly against the enclosing sediments on either side of the vein. The following winter, contraction occurs again and the entire cycle is repeated. Over a period of several hundred years, the original ice vein becomes an ice wedge and the displaced sediments are thrust upward to form the ridges that enclose the polygons.

Ice-wedge polygons are low-centre polygons if they are outlined by a ridge, and high-centre polygons if they are outlined by a trough. The ridges surrounding the low-

centre polygons are indicative of actively growing ice wedges. The troughs surrounding the high-centre polygons indicate that the material upturned by ice-wedge expansion is being removed by surface erosion.

Relict ice-wedge polygons are those formed during the Pleistocene when periglacial conditions were much more widespread than today. These fossil polygonal networks are best seen in freshly plowed fields because native vegetation or field crops tend to obscure the outlines of polygons.

Stone nets, stripes, thermokarst, and pingos. Repeated frost action also produces sorting in coarse clastic materials. The sorting process is complex and results in stone nets, circles, or polygons on flat surfaces, and stone garlands and stone stripes on hillsides.

Stone nets contain coarse cobbles and boulders on their borders and finer material near the centre. The diameter of stone polygons ranges from a few centimetres to about six metres (20 feet). They do not necessarily require permafrost conditions for growth.

Thermokarst is the general name applied to an area in which the thawing of permafrost has created a variety of features including lakes, mounds, and a hummocky surface similar to limestone terrains. Lakes are formed by the sinking of the ground surface when ice-rich permafrost is melted. The melting of ice wedges leaves the centre of the polygon elevated as a mound or hummock.

Pingos are ice-cored elliptical or circular hills of frozen sediments ranging in height from three metres to more than 60 metres (200 feet) and from less than 30 metres (100 feet) to nearly 400 metres (1,300 feet) in diameter. They occur in permafrost regions and are very conspicuous because they tend to occur singly or in clusters of two or three that rise abruptly above the general level of the tundra.

Pingos form when free groundwater becomes trapped by permafrost that is advancing into an area in which it did not previously exist, such as beneath a lake that has become filled with sediments. The freezing of the groundwater causes the overlying sediments, also frozen, to heave upward. The updoming of the surface materials causes them to crack and exposes the ice core to melting. As melting progresses, a summit depression or small crater forms. If the periglacial climate is replaced by more temperate conditions, the crater is enlarged, and the pingo itself decays to a circular peat-filled basin surrounded by a low ridge or rampart.

EOLIAN DEPOSITS

Deposits produced by wind action require a source from which sand and silt can be deflated. These sources are generally of three types. The first includes extensive stretches of desert regions in which weathered bedrock or regolith, or alluvial deposits are exposed to attack by wind. The second is composed mainly of outwash areas and till plains associated with continental glaciation. The third is found along coastal regions in which waves provide a renewed supply of beach sand for redistribution by the wind.

Sand dunes and loess deposits

Eolian deposits associated with glaciation are of two types: sand dunes and loess (*q.v.*). Sand dunes occur in a variety of geometric shapes and patterns that depend on wind direction and velocity, supply of sand, and amount of vegetational cover. Many of the sand dunes formed during the Pleistocene were derived from outwash areas. Most of them are stabilized today because of the encroachment of vegetation.

Loess is a windblown silt that occurs as a "blanket" deposit over large areas of the Mississippi Valley and northern High Plains of central North America, in parts of Europe, and over large areas of the eastern Soviet Union. Loess deposits in these areas were derived from outwash areas during the Pleistocene, especially the floodplains of the major rivers that carried large volumes of glacial meltwaters heavily charged with silt and other clastics.

A loess terrain is characterized by a "billowy" topography that may reflect the hills and valleys on which it was deposited. The silty texture of loess, its high water holding capacity, and its rich content of minerals make it one of the best materials for the formation of fertile soils. The crop yields of corn, wheat, and other grains grown on loess soils testify to the enormous value of loess as a natural resource.

Loess deposits have an internal structure and cohesiveness that accounts for the vertical bluffs along river channels eroded into a loess-covered terrain. These escarpments reveal the generally nonstratified structure of loess and provide excellent exposure for the study of these remarkable windblown materials associated with Pleistocene glaciation.

V. Modern landforms and former glaciation

The Ice Age ended about 10,000 years ago, but more than half the earth's land surface retains landforms that are a direct or indirect result of the climatic conditions associated with this last major epoch of geological history. These include such features as modern lakes and the strandlines of their expanded ancestral basins, coastal features related to former sea levels, the modern courses of major rivers that came into existence as a result of the ice advances, and relict landscapes that bear the imprints of erosional or depositional processes connected with the climate of the Pleistocene.

MODERN LAKES AND PLEISTOCENE FORERUNNERS

Two regions will serve to illustrate the relationships of modern lakes and their Pleistocene forerunners. The Great Lakes and their ancestral water bodies had a long and complex history directly related to the general retreat of the glacier lobes of late Pleistocene time. On the other hand, the lakes of the western United States, especially those in Utah and Nevada, were never directly influenced by glacier ice but their histories reveal a close relationship to the glacial and interglacial climates of the Pleistocene.

The lowlands now occupied by the Great Lakes were filled with lobate tongues of glacier ice during the beginning of the final deglaciation of the Pleistocene. Water bodies that formed along the lobate ice margins increased in area as the lobes continued their irregular but persistent retreat. The elevation of the water surface of any one of these ice-marginal or proglacial lakes remained constant for a period of time to allow wave action to build beach ridges around the part of the lakeshore that was not ice-bordered. A lake level would rise, however, if the ice-lobe advanced temporarily, thus reducing the area of the lake basin. Or, the lake level would fall if the ice lobe retreated and uncovered more glacier-free terrain over which the lake waters could spread. During some episodes of retreat, much lower outlets became ice-free so that the lake level dropped markedly when the waters abandoned the old higher outlet for the new lower one.

The shorelines of the modern Great Lakes are roughly parallel to the older strandlines, the study of which has provided the details of the complex history of the ancestral Great Lakes.

Lake Bonneville and Great Salt Lake

The ancient beach ridges that occur in western Utah define one or more ancestral water bodies that existed during the Pleistocene. The largest of all these former lakes was Lake Bonneville the shrunken remnants of which are Great Salt Lake, Lake Provo, and Lake Sevier. It is generally agreed that the many fluctuations of the water level in Lake Bonneville and other contemporary lakes in the Basin and Range Province of the western United States were caused by increases or decreases in water volumes that coincided with the major advances and retreats of the Pleistocene ice sheet. Higher levels existed during glacial periods, and low levels, such as those today, are indicative of interglacial times. The implication is that the same climatic changes that caused ice caps to grow and expand also caused wetter conditions to prevail over the watersheds encompassing these basins. The wetter conditions that caused higher lake levels are called pluvial, and the greatly expanded water bodies in the Bonneville basin are termed pluvial lakes.

But not all the changes in level of Lake Bonneville are attributable to climatic changes from pluvial to nonplu-

vial climates. Detailed studies indicate quite clearly that downcutting of the outlet was a nonclimatic cause of at least one lowering of the water level. Another change in lake level was caused by a lava flow that diverted a stream into the basin, causing an increased volume of water and a subsequent rise in lake level. Whatever the climatic or nonclimatic causes of the ancient Bonneville shore features, they remain today as a link with geologic events of the Pleistocene.

COASTAL FEATURES AND SEA-LEVEL CHANGE

During the Pleistocene Epoch, sea level fell during the buildup of the continental ice sheets, and rose during their retreat. The maximum lowering during the Pleistocene was about 120 metres (400 feet) below the present ocean surface, and the maximum rise was about 90 metres (300 feet) above modern mean sea level. During the Wisconsin or last glaciation, sea level was about 90 metres below present sea level, and during the last interglacial (Sangamon), sea level was about 18 metres (60 feet) higher than at present. During the last 20,000 years, sea level has gradually risen to its present position.

The effects of these glacial-eustatic changes in sea level are preserved along the coastal regions of the world. Higher sea levels are recorded in raised beaches, wave-cut coastal terraces, and elevated coral reefs. The evidence for lower sea levels is found in submerged ridges that appear to be former beach ridges along Pleistocene coastlines that existed during periods of maximum glaciation.

Other effects of a lower sea level are the terraces along major river valleys such as the Mississippi and Rhine. These and other rivers eroded their channels deeper during the low sea levels of the glacial advances, and filled them with alluvium during the high sea levels of interglacial times. The repetition of the cut-and-fill cycle, coupled with a general absolute elevation of the coastal land surface, is the simplest explanation for these terraces.

The persistent rise in sea level during the last 200 centuries accounts for many of the estuaries and long coastal indentations that characterize many present-day coastal regions of the world. These long and relatively narrow bays and inlets are the drowned mouths of river valleys. A coastline dominated by drowned river valleys is called a ria coast. A ria differs from a fjord in that the latter was deepened by glacial erosion whereas the former was never occupied by glacier ice.

DRAINAGE PATTERNS AND ICE DISRUPTION

When the Pleistocene ice sheets began their first expansion over the North American, European, and Asian continents, the existing drainage lines were disrupted if not destroyed completely. The present system of river channels in the areas overrun by the continental glaciers thus differs from the preglacial system. In some cases, major changes enlarged the areas of some drainage basins at the expense of others. In other cases, the segments of some river channels were dislocated by the advancing ice and forced to flow in new channels parallel to the ice margin. In still other instances, the preglacial system of river channels was obliterated through filling by drift. In the glaciated regions of the world in which detailed borings and geophysical exploration have been accomplished, the preglacial topography reveals the configuration of the old drainage patterns. Some are drift-filled bedrock valleys that date from pre-Pleistocene time; others were formed during interglacial times.

One of the most thoroughly studied examples of major drainage changes brought about by the advance of the Pleistocene ice sheets occurs on the upper reaches of the Missouri River and its tributaries. Before the onset of glaciation, the preglacial Missouri system flowed northeastward toward the Hudson Bay lowland. Invasion of this area by ice from the northeast blocked the system and created a new glacier-margin channel that flowed southeastward and eventually became part of the modern Missouri River.

Elsewhere in the world it has been shown that segments of the Thames River in England and the Rhine and other rivers in northern Germany, plus several smaller stream systems in Scandinavia, Poland, and the Soviet Union, occupy channels that resulted from diversions caused by invading ice.

CLIMATIC IMPLICATIONS

Most of the landforms of the earth are not the products of a single geologic event or a constant climatic regime. Although there can be no doubt about the glacial origins of such landforms as drumlins, eskers, moraines, and the like, other landforms beyond the glacier borders may bear more subtle imprints of Pleistocene climatic conditions. These climatic conditions included not only the colder periods associated with the glacial stages but also the warmer periods of the interglacials. Paleontological evidence indicates that at least one of the interglacials was characterized by a climate warmer than today, a fact that is frequently overlooked in deciphering the origin of landforms beyond the borders of Pleistocene glaciation.

Warmer climates are more conducive to rock weathering than colder climates. Wetter climates produce more runoff but also more vegetation than drier conditions. Aridity enhances eolian activity and increases surface runoff rates, because of scarce vegetational cover and torrential thunderstorms that are commonplace in many arid or semi-arid regions of the world.

Scientists have access to evidence other than landforms, such as fossil plants and animals, from which climatic conditions can be inferred. If such means are absent, however, landforms may be the only means for deducing the nature and extent of past climatic changes. The relationships between landforms and climate can be understood only if the contemporary geomorphic processes are understood. In the quarter century following World War II, great impetus was given to the quantitative study of geomorphic processes. These include fluvial mechanics, glacier dynamics, eolian processes, weathering rates, and the many intricate and complex processes that are unique to periglacial conditions. Insofar as these studies have illuminated the geomorphic results of those processes, they are useful also in aiding the students of Quaternary history in reconstructing the succession of climatic fluctuations that have dominated the last 2,500,000 years of Earth history.

BIBLIOGRAPHY. R.J.E. BROWN, *Permafrost in Canada* (1970), includes a brief general discussion of permafrost conditions in Canada and a lengthy discourse on the relationship of permafrost to engineering and construction problems in the north (well written with a minimum of technical jargon); J.L. DAVIES, *Landforms of Cold Climates* (1969), a college-level treatment of geomorphic processes and landforms associated with glacial and periglacial environments, with most examples drawn from southeastern Australia (including Tasmania) and New Zealand; C. EMBLETON and C.A.M. KING, *Glacial and Periglacial Geomorphology* (1968), a scholarly, comprehensive treatment of the origin of glacial and periglacial landforms, written for the student with some previous knowledge of geomorphology or physical geography, and including many references to original works, organized on a chapter-by-chapter basis; R.W. FAIRBRIDGE (ed.), *The Encyclopedia of Geomorphology* (1968), a many-authored, international, one-volume encyclopaedia of geomorphic terms, processes, and studies, with precise definitions and relevant source literature; R.F. FLINT, *Glacial and Quaternary Geology* (1971), encyclopaedic and world coverage of all aspects of Pleistocene glaciation, Quaternary climatic changes, the fossil record, and geomorphic history of nonglaciated regions, including a bibliography of more than 1,300 entries covering the relevant literature of the world; K. RANKAMA (ed.), *The Quaternary*, vol. 1 (1965), comprehensive, authoritative summaries of the Quaternary history of Denmark, Norway, Sweden, and Finland, with excellent photographs, many diagrams, extensive bibliography, and a good index; and vol. 2 (1967), a companion volume with the same format, but covering the Quaternary of the British Isles, France, Germany, and The Netherlands.

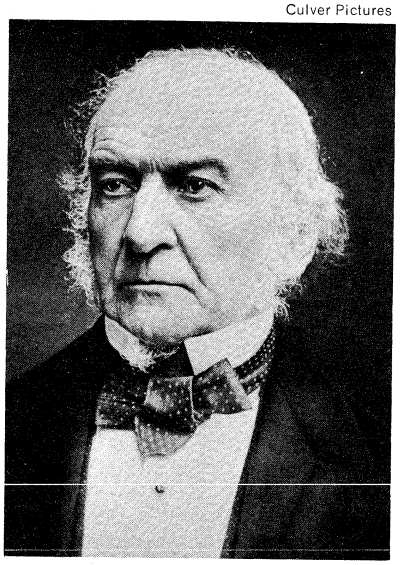
(J.H.Z.)

Gladstone, William Ewart

William Ewart Gladstone was beyond doubt the greatest British statesman of the 19th century. As leader of the Liberal Party and four times prime minister he, along

with his great opponent and rival Benjamin Disraeli, dominated British politics throughout the later Victorian age. His humanitarianism and his high-minded notions of political morality make him an outstanding figure in the history of European Liberalism. Gladstone was born in Liverpool on December 29, 1809, of purely Scottish descent. His mother, formerly Anne Robertson of Dingwall, was of a good Highland family; and his father, John, was descended from the Gledstanes of Lanarkshire. John Gladstone made himself a merchant prince by trading with the East and West Indies and became a leading citizen of the great slaving port of Liverpool. He was a member of Parliament from 1818–27 and died worth about £600,000. William Ewart Gladstone, named after a friend of his father's, and the fifth of six children, was sent to Eton, where he did not particularly distinguish himself. He first displayed his considerable powers of intellect at Christ Church, Oxford, where in 1831 he secured first classes in classics and mathematics.

Culver Pictures



Gladstone.

Though his father dissuaded him from his original intention of taking orders in the Church of England, he devoted his life, as he saw it, to serving the principles of the gospel in politics. As a disciple of the conservative philosopher Edmund Burke, he mistrusted parliamentary reform; his speech against it in May 1831 at the Oxford Union, of which he had been president, made a strong impression. One of his Christ Church friends, the son of the Duke of Newcastle, persuaded the Duke to support Gladstone as candidate for Parliament for Newark in the general election of December 1832; and the "Grand Old Man" of Liberalism thus began a parliamentary career of more than 60 years as a Tory member for what was little better than a pocket borough.

Election to Parliament

His abilities were noticed early; his maiden speech on June 3, 1833, defending the treatment of slaves on a plantation his father owned in Demerara, British Guiana, made a decided mark. He held minor office in Sir Robert Peel's short government of 1834–35, first at the treasury, then as undersecretary for the colonies. His first book, *The State in Its Relations with the Church*, was published in 1838; in it he argued narrow Anglican doctrines he soon outgrew.

After two other women had refused him, in July 1839 he married Catherine, the daughter of Sir Stephen Glynne of Hawarden, near Chester. A woman of lively wit, complete discretion, and exceptional charm, she was utterly devoted to her husband, to whom she bore eight children. This marriage gave him a secure base of personal happiness for the rest of his life. It also established him in the aristocratic governing class of the time; his wife's mother had been Pitt's first cousin.

The influence of Peel. Macaulay, in reviewing Gladstone's first book, had referred to him as the rising hope of stern and unbending Tories. His early parliamentary performances were indeed strongly Tory; but time after time actual contact with the effects of Tory policy was to force him to take a more liberal view. His conversion from the conservatism in which he had been brought up to the liberalism that gave him lasting fame took place in prolonged stages, over a generation. He took his first steps in a liberal direction during Peel's second ministry (1841–46). Peel rightly judged that the son of so successful a merchant would be useful at the board of trade. Gladstone, made vice president of the board under the ineffective Lord Ripon, complained privately that he was "set to govern packages"; yet the wisdom of Peel's choice was soon apparent. The vice president's powers of application astonished even his hardworking colleagues.

Under Peel's supervision he embarked on a major simplification of the tariff; while mastering the complexities of this subject, Gladstone became indeed a more thoroughgoing free trader than Peel himself. The Prime Minister felt that Gladstone was outstanding among all the promising young men in the government, and in May 1843 invited him into the Cabinet as president of the board of trade. The Railway Act of 1844 set up minimum requirements for railroad companies and provided for eventual state purchase of railway lines. Among other useful tasks, Gladstone much improved working conditions for London dock workers. Early in 1845, when the Cabinet proposed to increase a state grant to the Irish Roman Catholic college at Maynooth, Gladstone resigned—not because he did not approve of the increase but because it went against views he had published seven years before. At the end of the year he rejoined the Cabinet as secretary of state for the colonies, and so gave up his seat. For six months, till Peel's government fell in June 1846, he was in office, but not in Parliament —a position of doubtful constitutional propriety. While he was at the Colonial Office, he was led nearer to liberalism by being forced to consider the claims of English-speaking colonists to govern themselves.

Resignation over Maynooth

Private preoccupations. The Glynne family estates were deeply involved in the financial panic of 1847. For several years Gladstone was concerned with extricating them, devoting his customary energy to the intricacies of industrial investment and land tenure. In the course of these operations, he became the largest landlord in the county of Flint. At about the same time he began a habit of charitable work, which was open to a great deal of misinterpretation; he often tried, in the streets of London, to persuade prostitutes to enter a "rescue" home that he and his wife maintained or in some other way to take up a different way of life. He spent much time and money on these efforts till well past his 80th birthday.

Several of Gladstone's closest Oxford friends were among the many Anglicans who converted to Roman Catholicism under the impact of the Oxford Movement. Gladstone, though he had been brought up by an evangelical mother, had moved to a High Anglican position when in Italy, just after leaving Oxford; and once he had reached it he retained it. It pained him deeply when his younger sister became a Roman Catholic. Neither affection nor argumentative skill could ever persuade him to become a Catholic himself, but the suspicion that he was one dogged him and was used against him from time to time by his adversaries, of whom he had many in the University of Oxford, for which he was elected MP in August 1847. He scandalized many of his new constituents at once by voting for the admission of Jews to Parliament and many more by his tolerant opposition to Lord John Russell's Ecclesiastical Titles Act of 1851.

Gladstone made his first weighty speech on foreign affairs in June 1850, opposing foreign secretary Lord Palmerston's high-handed jingoism in the celebrated Don Pacifico debate over the rights of British nationals abroad. That autumn he visited Naples, where he was so appalled by the conditions that he found in the prisons that when he returned to London next February he could talk of little else. In July 1851 he published two trenchant letters to Lord Aberdeen that described what he

had seen and appealed to all Conservatives to set an iniquity right. The results were far from what he had desired. For the time, the Neapolitan prisoners were treated even worse than before, and most conservatives, all over Europe, were deaf to his appeal. But Palmerston circulated the letters to all the British missions on the Continent, and they delighted every liberal who heard of them.

Financial policy. For nine years after Peel's death in 1850, Gladstone's political position was seldom comfortable. As one of the most eminent of the dwindling band of distinguished Peelites, he was mistrusted by the leaders of both the main parties and distrusted some of them—particularly Palmerston and Disraeli—in his turn. He refused to join Lord Derby's government in 1852. At the end of that year, by a brilliant attack on Disraeli's budget, he brought the government down and took a long stride forward in public estimation as a result. He then joined Aberdeen's coalition as chancellor of the exchequer. His first budget speech, on April 18, 1853, adumbrated his great qualities of leadership. In a bold and comprehensive plan he made further large reductions in duties, proposed the eventual elimination of the income tax, and, with considerable political courage, carried a scheme (which every other Cabinet member had at first opposed) for the extension of the legacy duty to real property.

First budget speech

His budget provided the backbone of the coalition's success in 1853, a year in which he spent much time devising a scheme for a competitive civil service system. He was also busy preparing the Oxford University Bill that was passed in the following year; this local preoccupation kept him from taking any detailed interest in the events which led up to the Crimean War. He defended the war as necessary for the defense of the public law of Europe; but its outbreak disrupted his financial plans. Determined to pay for it as far as possible by taxation, he doubled the income tax in 1854. When Aberdeen fell in January 1855, Gladstone agreed to join Palmerston's Cabinet; but he resigned three weeks later, with two other Peelites, rather than embarrass his party by accepting a committee of inquiry into the conduct of the Crimean War. He was, as a result, unpopular in the country; and he made himself more unpopular still by speeches in Parliament in the summer of 1855, in which he held that the war was no longer justified, as its proper objects had already been attained. Meanwhile, still haunted by the horrors of Naples, he helped finance a quixotic project to send a steamer to rescue some Neapolitan prisoners confined on a Mediterranean island. The good or harm that might have resulted were both averted when the vessel sank on the way.

Gladstone always kept up his reading in classical studies, as well as in theology and Italian poetry; and he used his leisure while out of office to prepare a long book, *Studies on Homer and the Homeric Age* (3 vol., 1858), which suggested that ancient Greek life had been designed by providence to show men how they should behave to each other. He helped to defeat Palmerston in the Commons by a speech on China, in March 1857, and later that year opposed, on religious grounds, a bill putting divorce within the reach of the poor with such persistency that he was accused long afterward of having invented parliamentary obstruction himself. He twice refused to join Derby's government in 1858, in spite of a generous letter from Disraeli, but accepted an offer to visit the Ionian Islands Protectorate as lord high commissioner in the winter of 1858–59.

Joins Palmerston's Cabinet

In June 1859 Gladstone cast a silent and unavailing vote for Derby's Conservative government on a confidence motion and caused surprise by joining Palmerston's Whig Cabinet as chancellor of the exchequer a week later. His sole, but overwhelming, reason for joining a statesman he neither liked nor trusted was the critical state of the Italian question. The triumvirate of Palmerston, Russell, and Gladstone did indeed help, over the next 18 months, to secure the unification of almost all Italy, but on other matters the cabinet was much divided.

Gladstone was constantly at issue with his Prime Minister over defense spending. By prolonged efforts, he managed to get the service estimates down by 1866 to a lower figure than that for 1859. He took little other part in the government's foreign policy, except for a reference to the seceding Southern American states as "a nation" in October 1862. The national economy responded well to his policies at the exchequer, which included the further abolition of import duties by the celebrated budget of 1860. That year he supported an Anglo-French trade treaty, which shortly doubled the value of Anglo-French trade, and proposed to abolish the duties on paper, which to Palmerston's ill-concealed delight the House of Lords declined to do. Next year Gladstone repeated the proposal, and included it with all the other budget arrangements in a single finance bill that the Lords dared not amend, a procedure that has been followed ever since. Another particularly useful step was the creation of the post office savings bank. These measures brought him into increased contact and increased popularity with the leaders of working class opinion; journeys round the main centres of industry did the same.

In the general election of July 1865, Gladstone was defeated at Oxford but secured a seat in south Lancashire. When Palmerston died in October and Russell became prime minister, Gladstone took over the leadership of the House of Commons, while remaining at the exchequer. He was not well suited for the new post; he was too busy to ingratiate himself to backbenchers, and while Whigs distrusted him as a former Conservative, Conservatives distrusted him as a renegade, and radicals as a churchman.

By now quite convinced of the need for a further reform of Parliament, he introduced a bill for the moderate extension of the franchise in March 1866. Violently attacked by Whig dissidents led by Robert Lowe, it foundered in committee in June, and the whole government resigned. Next year Disraeli introduced a stronger Reform Bill that gave a vote to most householders in boroughs. Disraeli became prime minister early in 1868. Two months later, Gladstone carried against him in the Commons three resolutions calling for the disestablishment of the Protestant church in Ireland, and that autumn, in *A Chapter of Autobiography*, he explained the reasons that had led him so far from the conclusions of his first book. Russell had by now resigned from active politics, and it was to Gladstone that the Liberal whips looked for instructions during the general election at the end of the year. Though Gladstone lost his Lancashire seat, he was returned for Greenwich; and the Liberal Party won handsomely in the country as a whole. His abilities had made his its indispensable leader, and when Disraeli resigned, Queen Victoria called on him to form a government.

First administration. Gladstone's first Cabinet (1868–74) was perhaps the most capable of the century. Its Prime Minister tried to supervise the work of each department, devoting his main efforts to Irish and foreign policy. The Irish church was successfully disestablished in 1869, and a first attempt to grapple with oppressive landlordism in Ireland was made unsuccessfully in 1870; abroad, an attempt to promote disarmament in 1868 failed when Bismarck refused to consider it. The Franco-German War took the government completely by surprise, and the Cabinet would not allow Gladstone to propose to Prussia the neutralization of Alsace and Lorraine. The principal achievements of 1871 and 1872—a London declaration by the great powers that they would not in future abrogate treaties without the consent of all the signatories, and the settlement by arbitration of the "Alabama" claim of the United States—look well in retrospect but were thought pusillanimous at the time. The most useful reforms at home were administrative, except for the Education Act of 1870 and the Ballot Act of 1872; these measures supplemented the work of parliamentary reform but antagonized opposite wings of the Liberal Party. When an Irish University Bill failed by three votes to pass the Commons, in March 1873, Gladstone resigned but was forced back into office by Dis-

raeli's refusal to form a government. In August he had to reshuffle his Cabinet and again took on the chancellorship of the exchequer himself. Fascinated by another plan to abolish income tax altogether, he dissolved Parliament suddenly in January 1874, but his party was heavily defeated and his government resigned. Gladstone gave up the party leadership (though he remained MP for Greenwich) and retired to Hawarden to write pamphlets attacking papal infallibility and articles on Homer.

Retirement

Bulgarian atrocities. The indifference of Disraeli's government to the brutality of Turkish reprisals against risings in the Balkans, in 1875–76, brought Gladstone back to active politics. For years he had held that the only just solution to Balkan problems was self-government. He joined, rather late, an agitation begun by Liddon, E.A. Freeman, and W.T. Stead against the government's Eastern policy. In September 1876, he published a famous pamphlet on "Bulgarian Horrors and the Question of the East," which demanded that the Turkish irregulars should remove themselves, "one and all, bag and baggage," from the peninsula. London society and the London mob were both against him; the Queen, under the blandishments of Disraeli, strongly disapproved; the Whiggish elements in his own party were lukewarm or indifferent at first; and only some radicals, with whom he had little in common, really supported him. Yet in the end he triumphed. He gave up his Greenwich seat and stood for the Scottish county of Midlothian. In two tremendous outbursts of oratory, in November 1879 and March 1880, Gladstone secured his own return to Parliament, but he also did much more: he overthrew a government. The general election of March and April 1880 had more, and on the whole more ardent, contests than any before, and it was this one man's eloquence that decided the result and secured a large Liberal majority. The feat is unique. Gladstone was able, by his manifest sincerity and skill of argument, to convince a majority of the electorate that recent Conservative policy had been morally wrong, and the Conservative government had to resign.

Midlothian campaign

Second administration. In his second administration (1880–85), Gladstone foolishly combined again for two and a half years the duties of prime minister and chancellor of the exchequer. Party lines had not yet set firm, and his large apparent majority in the Commons was unruly. It was not until 1884 that he could introduce the measure for which most Liberals had been pressing—a third Reform Act that nearly doubled the electorate by giving votes to householders in country districts. This measure was passed only after a stiff quarrel with the House of Lords. Gladstone had hoped to settle the Eastern question quickly and then retire. As it was, he and Granville, the foreign secretary, did manage by a brusque naval threat to compel Turkey to cede Thessaly to Greece; yet Granville would not let him go on to give up Cyprus, as the government had already been weakened by other questions and had vital work still before it. There were still graver troubles in Ireland, in the throes of agricultural catastrophe. The exceedingly complicated Irish Land Act of 1881, largely Gladstone's own work, did in the long run promote the prosperity of the Irish peasant; but violent crime continued, culminating in the murder of Gladstone's close friend and nephew-in-law Lord Frederick Cavendish by Irish Nationalists. No alternatives to strong police powers were left in Ireland, and measures to restrict the freedom of Irish members to obstruct the work of the Commons had to be taken. Gladstone hated both, but had to sanction them.

A third imperial imbroglio arose in 1882 when a series of unavoidable decisions compelled the Cabinet to authorize the occupation of Egypt. Gladstone's settlement of the Egyptian debt question (1885) was honourable to his belief in the concert of Europe but had the unintended effect of tying British foreign policy to German. The worst mistakes he ever made were to allow Gen. C.G. Gordon ("Chinese Gordon"), whom he never met, to go to Khartoum in Sudan and then fail to rescue him; Gordon was killed in January 1885, and his death cost Gladstone much in popularity. Firm handling of a dispute with Russia over the border of Afghanistan did something to restore his prestige; but when the government was defeated on the budget in June 1885, he was glad to resign. He refused a gracefully worded offer of an earldom from the Queen.

Failure to rescue Gordon

Irish Home Rule. Though he had only spent three weeks in Ireland in his life (in 1877), Gladstone had imagination enough to appreciate the full force of Irish nationalism. As his close colleagues knew, he had for years looked favourably on the case for Irish Home Rule, that is, for a subordinate parliament in Dublin. In the autumn of 1885 he believed the time for it was ripe; but as a combination of Irish with Conservative votes had defeated him in June, he waited silently to see what an Irish–Conservative combination would produce. The general election of November–December 1885 returned a Parliament in which the Liberal members exactly equaled the total of Conservatives plus Irish. At this moment, Gladstone's youngest son revealed his father's conversion to Home Rule, and most Conservatives therefore turned against it. Lord Salisbury's government was defeated when Parliament met, and Gladstone formed his third Cabinet in February 1886. His Home Rule Bill was rejected in Parliament in June, by a large secession of Whigs, and in the country at a general election in July, and Gladstone resigned office.

He had kept his Midlothian seat, unopposed, and carried with him into the new Parliament a personal following 190 strong, supported by the National Liberal Federation, the most powerful political machine in the country. He devoted the next six years to an effort to convince the British electorate that to grant Home Rule to the Irish nation would be an act of justice and wisdom. This policy was abhorrent to the English upper classes, and for the first time a marked class division between the leading parties opened. At the jubilee of 1887, Gladstone was cheered from the pavements but hissed from the balconies. The act was symbolic of the position he had reached as the great popular hero of the age. His reputation stood higher with Scotsmen, Irishmen, and Welshmen than with Englishmen, except for the English Nonconformists, to whom his tendency to regard and describe great political questions as moral ones made a strong appeal. Portraits of him were far commoner in poor men's cottages than those of any other political leader. He spoke at many great meetings and cooperated with the Irish leader Charles Stewart Parnell. But in 1890, when he was 80, momentary excitement led him into a dangerous quarrel with Parnell about the political consequences of the O'Shea divorce. (Gladstone had not believed the rumours about Parnell's liaison, holding that Parnell would never "imperil the future of Ireland for an adulterous intrigue.") He never sought to correct the stories Parnell spread about him in Ireland. He sanctioned an extensive program of Liberal reforms drawn up at Newcastle in 1891, because it was headed by Home Rule, and on this platform the Liberals won in the general election of 1892 a majority of 40—if the Irish voted with them.

Quarrel with Parnell

Gladstone, an "old, wild, and incomprehensible man of 82½," as the Queen called him in a letter at the time, formed his fourth Cabinet in August 1892. Its members were held together only by awe of him. He piloted another Home Rule Bill through 85 sittings of the Commons in 1893; the Lords rejected it by the largest majority ever recorded there, 419–41, but the full discussion in the Commons brought Ireland eventual benefit. The Cabinet rejected Gladstone's proposal to dissolve.

He could not agree with his colleagues that a large increase in naval expenditure was necessary and finally resigned—ostensibly because sight and hearing were failing—on March 3, 1894. He was much mortified by the coolness of his last official interview with the Queen he had served loyally all her reign; by now she frankly detested him and had to struggle to conceal the fact in his presence. He retired to Hawarden and busied himself with a critical edition of the works of Bishop Joseph Butler (2 vol., 1896). Humanitarian to the end, in his last great speech, at Liverpool in September 1896, he de-

nounced Turkish atrocities in Armenia. After a painful illness, he died of cancer of the palate at Hawarden, on May 19, 1898. He was buried in Westminster Abbey.

Character. His truly extraordinary vigour far exceeded that of other men and was coupled with no less extraordinary powers of self-control and an iron devotion to duty. Training and natural ability led him to qualify statements and subdivide arguments; his thoughts indeed were often complicated, but his character was fundamentally simple. Thus, it was a simple sense of duty that took him into politics, a career in which he never felt really at home and for which he was in some ways unfitted, not least by his tendency to believe that other men's motives were invariably as disinterested as his own and by his excessive anxiety to maintain the consistency of his own conduct. Political courage and personal magnanimity he had in abundance, and he was the most efficient administrator of his age. His gift for concentration was remarkable; this helped people who did not know him well to think him hard or even hypocritical. But no one who knew him intimately doubted his entire sincerity or failed to be captivated by his delightful manners, his warmth, and his range of mind. His combative instinct, quickness of understanding, retentive memory, and inexhaustible fund of phrase made him a fearsome adversary in debate. Purely as an orator, he had two or three equals in his own day; as a statesman, only Peel came near him. A few British prime ministers —Walpole, Chatham, Pitt, Churchill—have been leaders as great; none has been more inspiring. Lord Acton, indeed, assessing for Gladstone's daughter, in 1879, her father's standing among the world's statesmen of the past two centuries, concluded that "in the three elements of greatness combined—the man, the power, and the result—character, genius, and success—none reached his level."

BIBLIOGRAPHY. Among nearly 40 lives, the best is still JOHN MORLEY, *The Life of William Ewart Gladstone*, 3 vol. (1903), although it emphasizes Gladstone's liberalism too much and his dependence on God too little. Many family points suppressed by Morley are in SIR PHILIP MAGNUS, *Gladstone* (1954); and in S.G. CHECKLAND, *The Gladstones, 1764–1851* (1971), the best modern studies. His political and literary papers, in 760 volumes in the British Museum, await full research; he has left his mark also in thousands of Public Record Office files. Of his published letters, AGATHA RAMM (ed.), *The Political Correspondence of Mr. Gladstone and Lord Granville, 1868–76*, 2 vol. (1952), and . . . *1876–86*, 2 vol. (1962), are important and revealing; as is D.C. LATHBURY, *Correspondence on Church and Religion of William Ewart Gladstone*, 2 vol. (1910). PHILIP GUEDALLA (ed.), *Gladstone and Palmerston* (1928) and *The Queen and Mr. Gladstone*, 2 vol. (1933), provide plenty of colour. Of Gladstone's own writings, the *Autobiographica*, vol. 1 (1972), though written in old age, is rewarding. His *Chapter of Autobiography* (1868) and *Midlothian Speeches . . . 1885* (1886, reprinted 1971) repay study. He collected many of his learned articles in *Gleanings of Past Years, 1843–78*, 7 vol. (1879), and *Later Gleanings* (1897). There is a bibliography of his work in A.T. BASSETT, *Gladstone's Speeches* (1916). M.R.D. FOOT (ed.), *The Gladstone Diaries*, 2 vol. (1968), so far only covers 1825–39.

(M.R.D.F.)

Glass, History of

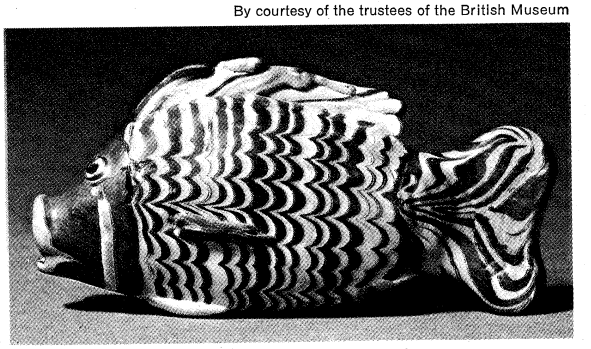
From very early times glass has been used for various kinds of vessels, and in all countries where the industry has been developed glass has been produced in a great variety of forms and kinds of decoration, much of it of great beauty. The aesthetic or artistic aspect of glass is the subject of this article. For the composition and properties of glass and the manufacture of various glass products such as glass containers, window glass, plate glass, optical glass, and glass fibres, see GLASS PRODUCTS AND PRODUCTION.

ANTIQUITY

Early glass. It is not certain in which of the civilizations of the ancient Near East glass was first made. The earliest wholly glass objects from Egypt are beads dating from some time after *c.* 2500 BC. A green glass rod found at Eshnunna in Babylonia may go back earlier, possibly to 2600 BC. A small piece of blue glass found at Eridu dates from before 2200 BC. The manufacture of glass vessels, which may have begun slightly earlier in Mesopotamia, was carried to a high point of excellence in Egypt during the 18th dynasty (*c.* 1490 onwards). These vessels

Egyptian technique

By courtesy of the trustees of the British Museum



Fish of core-made glass with "combed" decoration, Egyptian, New Kingdom, 18th dynasty (*c.* 1363–46 BC). In the British Museum. 0.141 m × .069 m.

are distinguished by a peculiar technique: the shape required was first formed of clay (probably mixed with dung) fixed to a metal rod. On this core the body of the vessel was built up, usually of opaque blue glass, on which, in turn, were coiled threads of glass of contrasting colour. The threads were pulled alternately up and down by a comb-like instrument to form feather, zigzag, or arcade patterns. The threads—usually yellow, white, or green in colour, and sometimes sealing-wax red—were rolled in (marvered) flush with the surface of the vessel. Finally, if desired, handles—often of translucent glass and sometimes of patterned "canes"—were added. The vessels were nearly always small, mainly for unguents and the like. Occasionally glass was decorated on the lapidary's wheel. Glass is known to have been made on the palace site of Tell el-Amarna, the residence of Akhenaton (reigned *c.* 1379–1362 BC), and the number of fragments found in and near the palace of Amenhotep III (reigned *c.* 1417–1379 BC) at Thebes suggests that it was made there also. This palace activity seems then to have died down and after the 21st dynasty (1085–945 BC) to have ceased altogether.

In Mesopotamia the Nineveh tablets of the reign of Ashurbanipal (668–*c.* 626 BC) and the remains of glass in various forms excavated at Nimrūd (ancient Calah, Assyria) indicate that glassmaking was carried on there during the 8th to the 6th centuries BC. It is probable that certain vessels of palish-green or deep blue glass, cut from a solid mass as if from stone, are Mesopotamian and date from as early as the 8th century BC (as a dish from controlled excavations in Phrygia proves). A vase of this type, contrasting completely with the core-wound glass of Egypt, bears the cartouche (panel enclosing the name) of the Assyrian king Sargon II (reigned 721–705 BC), and it is probable that glass treated in this way was manufactured over a long period in Mesopotamia.

Glass was made in Greece in Mycenaean times (*c.* 1400–1200 BC) usually in the form of small molded architectural details. A few pieces suggest, however, that perhaps some vessel glass also was made by the Egyptian technique, though not in Egyptian forms. Other Aegean-area glass of this period may have been imported from Egypt.

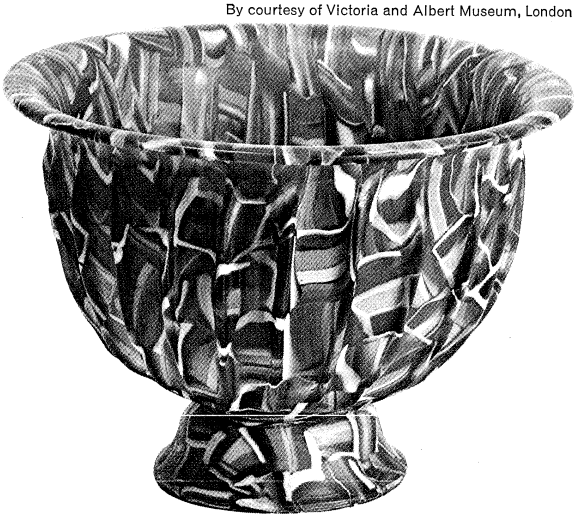
In general, glass of the earlier half of the 1st millennium BC is scarce and displays little homogeneity. From the 6th century BC, however, glass begins to appear in great quantities once again, particularly on the Greek-inhabited islands of the Aegean, in Greece itself, in Italy and Sicily, and even farther west. This contrasts with the meagre contemporary finds on Egyptian soil. The later glasses in the old Egyptian core-wound technique were probably made in Syria or some part of the Greek world. Such vessels were still small but differ in shape from the earlier Egyptian dynastic work. They were usually decorated

with light-coloured threads on a dark, usually blue, ground (familiar from the Egyptian 18th dynasty), but a notable variation was displayed in pieces decorated with dark purple threads on a white ground. In the Hellenistic period (roughly from the 4th century BC) the shapes of glass degenerated. The technique of decoration, however, remained the same; new colour combinations were used, and indeed these combinations continued into the era of blown glass.

Dominance of Alexandria in glassmaking

The Roman Empire. In Egypt during the Ptolemaic period (330–305 BC) Alexandria came to the fore in glassmaking. By about the 1st century BC, which saw the beginnings of glass as known today, it had become preeminent in certain glass techniques. Alexandria inherited and perfected the manipulation of coloured glass rods to make composite canes, which, when cut across, revealed a design (mosaic glass). Slices from such canes could be arranged side by side to produce repetitive patterns. When, as often happens, the cane slices show starry or flowerlike designs, the resultant glass is called millefiori ("thousand flowers"). An Alexandrian technical speciality more important for the future, however, was molding,

By courtesy of Victoria and Albert Museum, London



Bowl of pressed mosaic glass, probably Alexandrian, 1st century AD. In the Victoria and Albert Museum, London. Diameter 15.5 cm.

glass being pressed into, or powdered glass melted in, a mold. A combination of this process with the millefiori technique produced bowls with variegated designs in infinite variety. Sometimes glass of various colours was irregularly compounded to give the effect of a natural veined stone; occasionally enclosures of gold leaf in the glass simulated the glitter of natural pyrites (aventurine glass). Bowls were often finished around the rim with a cordon made of a clear glass thread twisted with one of opaque white. Sometimes such cable threads were themselves coiled round and round from a centre to make a bowl of lacy appearance, with the opaque white glass threads apparently set in a clear colourless matrix.

All these pieces might be finished with a fire polish by returning them to the furnace, but many mold-pressed glasses were, in fact, given a rotary polish, either by means of a spinning wheel fed with abrasives or by a process similar to lathe turning, in which the object spins and the tool is stationary. Similar equipment probably produced the numerous pieces that give every appearance of having been cut from a solid block of glass or at least from a thick, mold-pressed blank. Such pieces (usually flat dishes or two-handled cups) follow the contemporary forms of pottery and metalwork. Wheel engraving appears to have become an Alexandrian specialty around the 1st century BC and probably continued so throughout the two succeeding centuries. Alexandrian wheel engravers produced not only massive cut shapes, but also intaglio (incised) and relief surface decoration, the latter by laboriously grinding back the surface of the glass to form a background for the design. Simple motifs such as lotus buds or lotus flowers were produced in this way and occasionally more elaborate figural compositions were also done. Other specialties attributed to Alexandria were enamel painting (pigments mixed with a glassy flux were fused to the surface of the glass vessel by a separate firing) and an extraordinary technique of sandwiching a gold leaf etched with a design between two layers of clear glass.

Introduction of blown glass

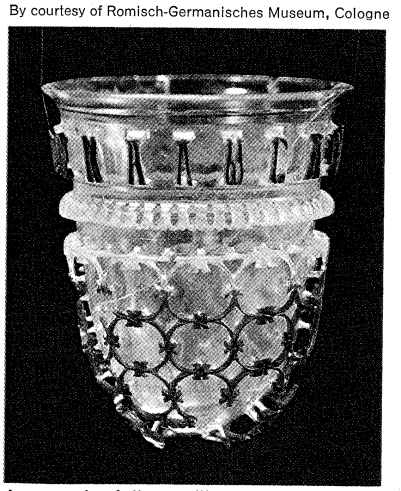
The most important innovation in the whole history of glass manufacture was blowing. Perhaps by a stroke of pure inventive genius it was perceived that glass on the end of a hollow metal tube could be blown into a mold as easily as it had theretofore been pressed in. The next stage was to use molds for forms, such as flasks, that could not be made by pressing. Finally, it was realized that the glass bulb on the end of the blowpipe could be shaped freehand to any form desired, and handles, feet, and decorative elements could be added at will. This liberating discovery, probably made during the 1st century BC, gave rise to the astonishing growth of the glass industry in Roman imperial times. In addition to the luxury vessels of types already described, which were produced with an elaboration of skill that astonishes and often baffles the modern technician, commercial containers in great variety were mass-produced in common greenish glass on a scale that was not matched until the 19th century.

The discovery of glass blowing may well be credited to the Syrian glassworkers, since the first mold-blown glasses bear the signatures of Syrian masters and since the readily ductile Syrian soda glass was especially apt for this purpose. Syrian glassworkers, however, seem to have migrated wherever demand promised a ready market, and some masters of mold blowing appear to have moved to Italy early in the 1st century AD; in the course of that century Italy became an important glass-producing area. Glass engraving especially seems to have flourished there and particularly one form of the art—grinding through an opaque white layer to a darker ground (cameo glass). The most famous example of this exacting technique is the Portland vase, in the British Museum, London. The capacity of the Italian glass craftsman to surpass all earlier masters in work of the most complex character is seen in the so-called cage cups (*diatreta*), on which the design—usually a mesh of circles that touch one another, with or without a convivial inscription—is so undercut that it stands completely free of the body of the vessel, except for an occasional supporting strut. These cups were made perhaps at Aquileia and date from the 3rd and 4th centuries.

Cameo glass engraving technique

Parallel to the pottery industry, glassmaking spread from Italy to northern Gaul, in particular to the valleys of the Rhône and the Rhine. In Britain the industry was probably not of great importance. The Rhineland, how-

By courtesy of Romisch-Germanisches Museum, Cologne



An example of *diatreta* ("cage cups"), perhaps made for a Greek living in the Rhineland, 3rd to 4th century AD. In the Römisch-Germanisches Museum, Cologne. Height 12.1 cm.

ever, became one of the great glassmaking areas of the Roman world (partly, it is thought, because of successive migrations of Near Eastern workers) and, although Rhenish glass is always recognizably Roman, several types of decorated glass were specialities of the district. Glasses decorated in serpentine patterns by threads trailed on and then pressed flat and notched are perhaps the most important and typical (*Schlangenfadengläser*). A considerable school of glass engraving also seems to have flourished, probably around Cologne. Although some engraving shows an impoverished linear style eked out by lines scratched with a hard stone point, some is executed by means of wheels sufficiently thick to permit rounded cuts corresponding to the modelling of the human figure, and simulating it when the piece is seen against the light. Both types of decoration flourished in the 3rd and 4th centuries.

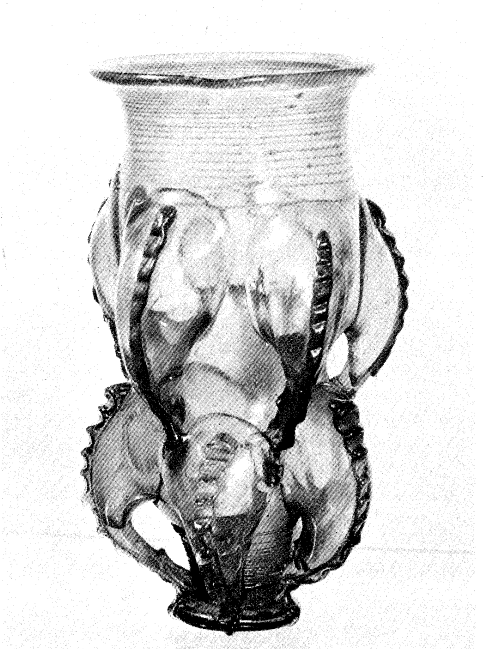
In Egypt in the later centuries of the Roman epoch glass was in frequent use for tableware, but artistic standards were not high. Plain dishes, cups, bowls, and lamps are characteristic; the glass of such tablewares ranges from an almost colourless "metal" (basic glass) of good quality to a greenish brownish substance full of bubbles and impurities. Decoration in this late period is mainly restricted to a few rough-cut lines, an occasional group of coloured glass blobs on the lamps, or a zigzag trail of glass thread running between the lip and the shoulder of a vase. In Syria during the same period, however, this trailing technique, which was particularly suitable to the ductile Syrian material, was carried to extreme lengths—threads circling the body or neck of a vessel, a profusion of zigzags, and fantastically worked handles.

Syrian use of trailing technique

MIDDLE AGES

With the breakdown of the Roman Empire, glassmaking fared differently in different parts of the world. In the East, urban life continued relatively undisturbed, and glassmaking evolved in an unbroken progress into Islāmic times. In the northern provinces, however, glassmaking became an affair of small, often isolated, glasshouses working in the forests that supplied them with fuel. Relatively simple shapes were made of an impure greenish or yellowish material, and decoration was restricted to simple trails of thread. Considerable virtuosity, however, was displayed from *c.* 500 onward in the manufacture of elaborate and fantastic *Rüsselbecher* ("claw beakers") on which two superimposed rows of hollow, trunklike protrusions curve down to rejoin the wall of the vessel above a small button foot.

By courtesy of the trustees of the British Museum



Rüsselbechen ("claw-beaker") Frankish glass, probably 7th century AD. In the British Museum. Height 19.0 cm.

In the East, Syria appears to have continued its predilection for trailed and applied ornamentation. In Egypt the art of glass suffered a catastrophic decline; only small rough vessels of impure green or blue material were manufactured.

Byzantium. In Byzantium itself the position of glassmaking is obscure. A distinction was made between *vitrarii* ("glassmakers") and *diatretarii* ("glass cutters") in edicts of Constantine the Great, Theodosius, and Justinian—suggesting that cutting played an important part in Byzantine glass decoration. This is borne out by the fact that cut glass made up the greater part of the glass that was brought back from the sack of Constantinople by the crusaders and placed in the treasury of St. Mark's in Venice. Apart from a few pieces of obviously Roman glass, presumably kept as heirlooms in Byzantium, these glasses are decorated either with tessellated (mosaic) patterns of overlapping round or oval facets or with round bosses in relief. These same two forms of cutting are observable in glass of the 5th century excavated at Kish in Mesopotamia; it is a fair assumption that Byzantine taste in glass, as in some of the other arts, was strongly influenced by the East. It is probable, however, that some enamelled and gilt glass also was made in the Byzantine provinces (*e.g.*, in Corinth), if not in Byzantium itself.

Glass from the sack of Constantinople

Islām. In the 7th century the whole Near East was overrun by the Arabs, and a number of rival dynasties were established in different parts of the conquered territory. An Islāmic civilization developed comparable to the preceding area of Greco-Roman culture, and a distinctively Islāmic glass style evolved. Although often it is not possible to say where a particular glass was made, different parts of the Islāmic world seem to have shown predilections for one or another type of glassmaking. In Syria, pieces more or less heavily decorated with trailed threads or applied blobs and pieces blown in molds, patterned with ribs or other allover designs, were still made. In Mesopotamia, glassmaking and, in particular, engraving flourished, especially during the ʿAbbāsid dynasty (750–1258), and attracted many of the best artists in the Islāmic world. Not only were the earlier modes of facet and boss cutting continued, but (perhaps deriving from them) two splendid new styles were created, one of linear intaglio, the other of relief cutting (outlines were left in relief by cutting back the ground and were then enlivened by crosshatching). Bowls, bottles, and ewers of remarkable sumptuousness were decorated with forms of running animals and plant scrolls. The quantity of engraved glass of these types found in Persia suggests that such work was done there also.

In Egypt there was both innovation and, after the post-Roman period, a notable revival of earlier techniques. Among the innovations was the stamping of glass by means of tongs, one jaw of which was patterned. The technique also is found in other lands. One extension of it, by which a bottle's upper and lower halves, made separately in contrasting colours, were decorated by the tongs and then joined together, was probably a Syrian innovation. More important was the Egyptian invention of lustre painting. In its simplest form it consisted of painting with a pigment containing silver that when fired in a smoky atmosphere (*i.e.*, without oxygen) produced on the glass a thin, metallic film that varied in colour from pale yellow to brown. Intact bowls and a bottle decorated by this technique exist, but whole classes of much more elaborate lustre-painted glass are represented only by fragments. A very wide variety of sumptuous polychrome effects are represented, although many were probably not produced by lustre properly so-called. The technical processes by which these effects were achieved are not yet understood.

Egyptian tong stamping and lustre painting

Egyptian Islāmic revivals in glass included millefiori effects, mainly in plaques for wall decoration, and white fern and feather patterns produced on dark glass vessels by combed and imbedded glass threads. Glass cutting was

also practiced in Egypt, mainly for the production of deeply incised small perfume bottles of square section, the bases of which were often cut into four tapering feet ("molar tooth" bottles). It seems probable that Egypt also perfected the techniques of gilding, decisive for the next phase in Islāmic glassmaking. In gilding, gold leaf is applied to an object that is then fired to fix the glass.

Syrian enamelled and gilt glass

Glassworkers migrating from Egypt to Syria after the fall of the Egyptian Fāṭimid dynasty in 1171 may have laid the foundation of the Syrian art of enamelled and gilt glass. Although earlier phases of this art are incompletely understood, the first group of enamelled and gilt glasses seems to be one in which thick enamels are used (particularly white and turquoise blue), often in series of beadlike drops; this group is tentatively associated with the town of Raqqah in Syria. A similar doubt surrounds the origins of two broad families into which Syrian glass of the 13th century is divided. One, characterized by the use of thick, jewellike enamels, is connected with the town of Aleppo; the other, notable for its exquisitely painted small-scale figural decoration, is attributed to Damascus.

Both cities were famous for their glass at this time, but it is uncertain what each produced. Wherever made, these two types of glass represent one of the highlights in the history of the art, whether one considers the rich green, red, yellow, white, and turquoise-blue enamels of the "Aleppo" group or the masterly red outline drawing of the "Damascus" group.

Toward 1300, Chinese influence, infiltrating by way of the Mongols and Tatars, makes itself felt in the decoration of these glasses, as is apparent in the series of great mosque lamps that then began to be inscribed with the names of rulers and great officers of state in Egypt. From a peak of excellence at the beginning of the 14th century a decline set in, greatly precipitated by the Mongol conqueror Timur's sacking of the chief Syrian cities at the end of the century.

Damascus fell finally in 1400, and it is recorded that the glassworkers of that city were carried into captivity in Samarkand. Nevertheless, some enamelled glass of inferior quality continued to be made in the 15th century, perhaps in Egypt. By the end of that century, however, there is evidence that mosque lamps were being made in Venice for the oriental market and the great Near Eastern tradition of enamelled and gilt glass was clearly moribund.



Bottle of enamelled and gilt glass decorated with Chinese motifs and an inscription in Kufic lettering praising an unknown sultan, Syrian, Mamlūk period, *c.* 1300. In the Victoria and Albert Museum, London. Height 43.5 cm.
By courtesy of Victoria and Albert Museum, London

MID-15TH TO MID-19TH CENTURY

Venice and the façon de Venise. A glass industry was already established near Venice in the 7th century, and vessel glass was made there by the last quarter of the 10th century. In 1291 the glass furnaces were removed to the neighbouring island of Murano to remove the risk of fire from the city. Although Venice had constant contact with the East, there is no evidence that it was indebted to that source for its skill in glassmaking. Venetian enamelled glasses appear in the second half of the 15th century, and, although their technique is essentially similar to that of the Syrian glassmakers, it is likely that they are of independent development. Little is known of the vessels made before this period, but it is evident from representations in pictures that they were mainly footed flasks and low beakers. The Venetians attributed the introduction of enamelling to a member of the glassmaking family of Barovier. The earliest pieces known, commencing with a goblet dated to 1465, certainly show no signs of outside influence. These, like most Venetian glass of the period, were inspired by the artistic ideals of the Italian Renaissance. The decorations represent triumphs, allegories of love, grotesques (fanciful combinations of human and animal forms), and so forth, with borders of dots of enamel laid on a ground of gold etched in scale pattern. Many of these pieces were of richly coloured glass, blue, green, or purple.



Goblet, green glass enamelled and gilt, Venetian *c.* 1500. In the British Museum. Height 22.2 cm.
By courtesy of the trustees of the British Museum

The Venetians were keenly aware of Roman achievements in glassmaking as in the other arts; they reproduced mosaic, millefiori, and aventurine glass, and glass resembling natural layered stones (*calcedonio*, sometimes miscalled *Schmelzglas*), and they even copied a Roman form of bowl that had vertical, external ribs. All these types of glass were Venetian specialities, and they were probably developed as a part of the extensive local bead industry.

Clear, colourless *cristallo* glass

The greatest achievement of Venice, however, and that upon which its great export trade came to be based, was the manufacture of clear, colourless glass, which was apparently exclusive to Italy during the Middle Ages. From its resemblance to natural crystal, this material was called *cristallo*, although in fact it often has a not unpleasing brownish or grayish cast. Made with soda, it was very ductile and cooled quickly. It therefore de-

manded of the workmen great speed and dexterity, and this, in turn, affected the nature of the glasses made. In the first half of the 16th century the Venetian glassblowers produced glasses of an austere simplicity. As the century proceeded (and more markedly still in the 17th century), however, there was a tendency to produce elaborate and fantastic forms. Enamelling on glass went out of fashion in Venice (except on pieces for export) in the first half of the 16th century. Its place was taken to some extent by the use of opaque white glass threads for decorative purposes (latticinio). This form of decoration became progressively more complex; opaque threads were embedded in a matrix of clear glass and then twisted into cables, which were themselves used to build up the wall of a vessel. The height of complexity was reached when a bulb of glass decorated with cables or threads running obliquely in one direction was blown inside a second bulb with threads twisted in the other direction. The composite globe thus formed was then worked into the desired form. This resulted in a vessel completely covered with a lacy white pattern (*vetro di trina*). Other methods of decoration at this time were mold blowing and dipping a vessel while hot into water or rolling it on a bed of glass fragments to produce a crackled surface (ice glass). *Cristallo* was also found suitable for engraving with a diamond point, a technique which produced spidery opaque lines that were especially suitable for delicate designs. The technique seems to have come into use about 1530.

Spread of Venetian glass-making

The glassworkers of the island of Murano were forbidden to leave Venice or to teach their secrets to outsiders, under dire penalties both to themselves and their families. Such was the demand for Venetian glass in the rest of Europe, however, and such was the desire of kings and nobles to control and reap the profits of its manufacture, that many Venetian workmen in the course of the 16th century were tempted to abscond to other countries, where they helped to set up glassworks. Furthermore, at Altare, near Genoa, existed a second great centre of glassmaking. Its glass was so like the Venetian in style and material that it is nowadays impossible to distinguish between the two. The glassworkers at Altare, moreover, were governed by no such laws as the Venetians; rather, they made it their policy to supply their men and teach their methods wherever there was demand for them. Thus, the fugitive Venetians and the willing Altarists spread the Italian art of glass to the rest of Europe, and glasshouses were established in France, Spain, Portugal, Austria, and Germany, while in the North, Antwerp seems to have been a secondary source of diffusion.

By courtesy of the trustees of the British Museum



Venetian glass ewer in the form of a nef ("ship"), attributed to Ermonia Vivarini, *c.* 1520. In the British Museum. Height 34.3 cm.



Latticinio tazza, colourless glass with opaque white decoration, Venetian, 16th century. In the Corning Museum of Glass, New York. Height 13.7 cm., diameter of rim 15.2 cm.

By courtesy of the Corning Museum of Glass, New York

Italian glassworkers ranged as far north as England, Denmark, and Sweden. Their labour was necessarily diluted by that of native workmen to whom they were often required to teach their methods. Variations in locally available raw materials modified the quality of the glass, and local taste influenced the form and ornamentation of the objects they produced. Nevertheless, in the late 16th and the 17th centuries an international style in glass developed, wholly Italian in origin and inspiration (*façon de Venise*).

Although there was everywhere a family likeness among glasses of the *façon de Venise*, certain countries developed types peculiar to themselves that are worthy of mention. Thus in Spain not only were fantastic and even bizarre shapes evolved in green glass, but in Barcelona a characteristic kind of enamelled decoration was developed, the peculiarities of which include a light-leaf-green colour and a constantly recurring lily-of-the-valley motif (late 15th–16th century). Elsewhere, at Hall, in the Tirol, a characteristic decoration with the diamond point, often supplemented by cold painting (*i.e.*, unfired oil—or other paint applied to a finished object), was favoured in alternating broad and narrow upright panels containing symmetrical scrollwork or coats of arms and other devices. Almost equally stiff and formal diamond-point work is to be seen on glasses probably made at the London glasshouse of Jacopo Verzelini (examples dated between 1577 and 1590). A more promising development of diamond-point engraving occurred in the Netherlands. There too the work of the 16th century was relatively formal and stiff, linear and clear, with simple hatching only. In the succeeding century, however, diamond-point engraving became initially more supple and pleasing, only to degenerate eventually into overelaboration.

Diamond-point engraving in the Netherlands

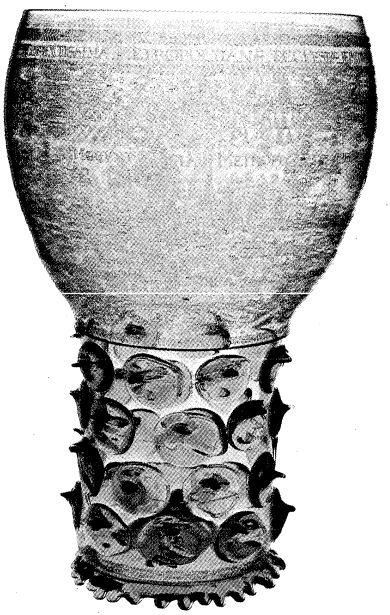
Diamond-point engraving was practiced there widely by talented amateurs in the 17th century, among them Humanists such as Maria Tesselschade Roemers Visscher, her even more famous sister Anna Roemers Visscher, and Anna Maria van Schurman. The latter two decorated their glasses with flowers and insects drawn with a gossamer touch, often accompanied by epigrams in Latin or Greek capitals scratched with severe precision or in the free scrolled style of the Italianate writing

masters of the time. A similar calligraphy was practiced later in the century by the amateur Willem Jacobsz van Heemskerk, with notably beautiful results.

Engraving in the first half of the 17th century gradually abandoned linear clarity in favour of crosshatched chiaroscuro (shading) effects, the highlights formed by sometimes completely opaque spots. Many artists worked in this manner; two are worthy of special mention. One was an accomplished engraver signing "C.J.M.," whose earliest dated glass is of 1644; the other was Willem Mooleyser, of Rotterdam, who worked in the last two decades of the 17th century with a scribbled freedom and vigour that raised his work above the average. By the end of the century this type of diamond-point work was superseded in popularity by wheel engraving.

Germany. In Germany toward the end of the 17th century a reaction to Venetian glass styles seems to have set in. In that country there had been a continuous survival, probably from late Roman times, of a local type of green glass, a product of forest glasshouses made with potash obtained by burning forest vegetation and called therefore Waldglas ("forest glass"). From this material, often of great beauty of colour, were made shapes peculiar to Germany, notably a cylindrical beer glass studded with projecting bosses, or prunts (*Krautstrunk*, or "cabbage stalk"), and a wineglass (*Römer*) with cup-shaped or ovoid bowl set on a similarly prunted hollow

The Waldglas tradition



"The Mainz Dean and Chapter *Römer*," with diamond-point engraving of the city of Mainz; Netherlandish, 1617. In the Bayerisches Nationalmuseum, Munich. Height 32.5 cm.
By courtesy of Bayerisches Nationalmuseum, Munich

stem. This became the classic German shape of wineglass, which survived into the 18th century and, with modifications, to the present day. Apart from these indigenous forms, German glass in Venetian-type *cristallo* developed local characteristics of its own in the latter part of the 17th century.

In Nürnberg, for instance, the tall-stemmed Italianate goblet underwent a transformation into a severe glass with stem composed of no more than a baluster-shaped element and a bulb, joined together by a number of disk-shaped elements, or mereses, and attached to foot and bowl by the same means. Such goblets display some of the most accomplished glass engraving that was ever practiced.

The leader and founder of the Nürnberg school of engravers was Georg Schwanhardt, a pupil of Caspar Lehmann. Lehmann had been gem cutter to the emperor Rudolf II in Prague and there had taken the decisive step of transferring the art of engraving from precious stones to glass. His first dated work is a beaker of 1605; in 1609 he obtained an exclusive privilege for engraving glass. Although he is the first great personality in glass engraving, he was not the first to practice the art in the German area. On Lehmann's death in 1622 Schwanhardt inherited his patent and moved to his own native city, Nürnberg, where a whole school of glass engraving grew up around him and his family. Schwanhardt's work is characterized by delicate, tiny landscapes, often accompanied by bold formal scrollwork. His son Heinrich excelled in minute landscapes but also engraved inscriptions of fine calligraphic quality. Other notable Nürnberg engravers of the late 17th century were Paulus Eder; Hermann Schwinger, a master calligrapher; and H.W. Schmidt and G.F. Killinger, both notable for the delicacy with which they rendered landscapes. Somewhat similar work was done at Frankfurt am Main by members of the Hess family.

In Bohemia, after Lehmann's death, little engraving of high quality was done. Just before 1700, however, with the perfection of a massive, crystal-clear, potash-lime glass that allowed cuts of considerable depth, the engravers of the Bohemian–Silesian area came into prominence. The harnessing of the mountain streams in the Riesengebirge for water power enabled engravers (those of the Hirschberger Valley in particular) to practice relief engraving, which demands immense energy for grinding down the background of the design. Massive covered goblets were decorated with powerful acanthus scrolls in the contemporary baroque taste. Relief engraving (*Hochschnitt*) was only occasionally used by itself in the Bohemian–Silesian area in the 18th century; more often it was employed in conjunction with intaglio (*Tiefschnitt*). By the turn of the 18th century the engravers of this area —anonymous workmen regarded as artisans rather than as artists—had acquired great technical skill; this enabled them to adapt to glass all the changing fashions of the 18th century in the decorative arts. Glass engraving, often of fine quality, was also practiced in many parts of Germany—notably Thuringia, Saxony, and Brunswick—but the most significant work of the late 17th and early 18th centuries was that done in Brandenburg. There, the glassworks at Potsdam (moved to Zechlin in 1736) produced massive goblets and beakers that were engraved—usually to order for the court—in Berlin, where a water-powered engraving shop had been installed in 1687. Both relief and intaglio engraving were practiced, the latter being favoured. This workshop, indeed, produced perhaps the greatest of the German intaglio engravers, Gottfried Spiller, whose deep cutting on the thick Potsdam glass has seldom, if ever, been surpassed. A notable, if lesser, engraver from the same shop was Heinrich Jäger; and later, in the 1730s and 1740s, work of high quality was done by Elias Rosbach.

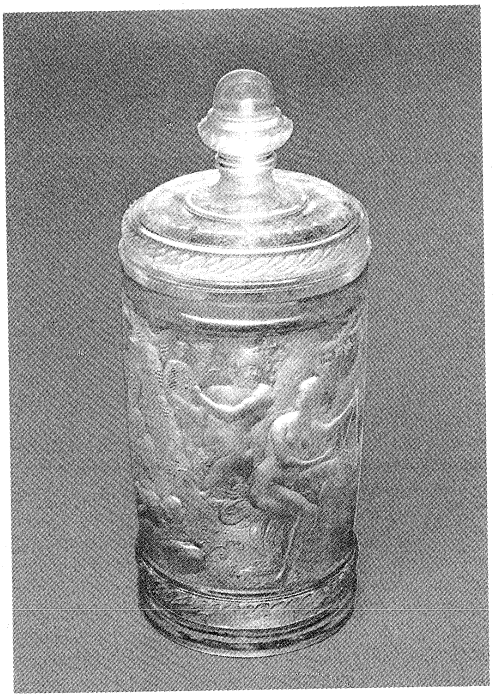
Bohemian relief and intaglio engraving

Another workshop of great significance was established toward the end of the 17th century at Kassel, in Hesse. There perhaps the greatest of all the relief engravers, Franz Gondelach, handled glass with a truly sculptural feeling.

In the second half of the 18th century, engraved glass declined in favour, although the technical skill required for its production never died out in the Bohemian–Silesian area. It experienced a great revival in the second quarter of the 19th century, when the taste of the newly prosperous bourgeoisie favoured elaborate decoration. The engraving of this period is often skillful in the extreme, although marred by excessive naturalism. Striking innovations of the period were the use of a casing (normally ruby red, blue or opaque white) through which the design was cut down to the colourless glass. A yellow coating (the silver stain of the stained-glass artist) was often used in the same way. Notable engravers of this epoch were Dominik Bimann, August Böhm, A.H. Pfeiffer, and members of the Pelikan and Simm families.

Second in importance only to engraving as a method of decorating glass in Germany was enamelling. Germany had proved a profitable market for enamelled Venetian glass during the 16th century, and, in the latter part of that century, glass enamelling began to be practiced in the Germanic lands themselves, most notably in Bohemia. This enamelling, in bright opaque colours, was much fa-

German enamelled glass

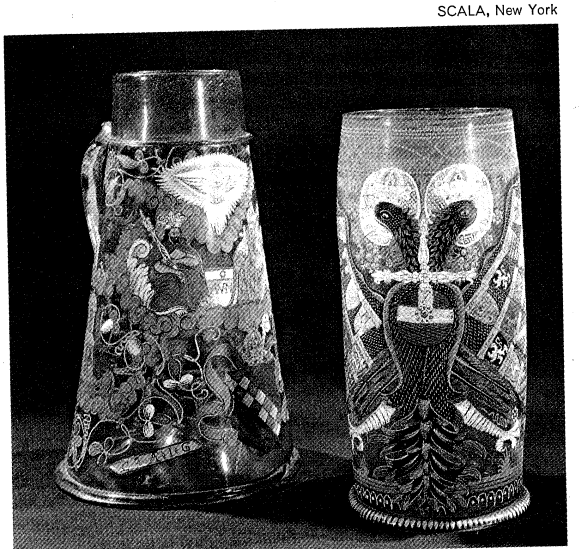


Beaker and cover, unpolished intaglio engraving with relief-cut laurel frieze by Gottfried Spiller, c. 1700. In the Kunstmuseum Düsseldorf, West Germany. Height 27 cm.

By courtesy of Kunstmuseum Dusseldorf,West Germany

voured throughout the 17th century, chiefly on the cylindrical drinking glasses, often of great size, known as *Humpen*. The glass they were made of was frequently impure and of a greenish or yellowish cast, while the painting itself was the simplified repetitive work of artisans rather than of original artists. Nevertheless, the gaiety of colour of these glasses and a certain naïveté in their painting give them an authentic unsophisticated charm. The most favoured types of decoration include a representation of the imperial double-headed eagle (*Reichsadlerhumpen*); representations of the emperor with his seven electors, either seated or mounted on horseback (*Kurfürstenhumpen*); subjects from the Old and New Testaments; and allegorical themes such as the Eight Virtues and the Ages of Man. These were painted between borders of multicoloured or white dots or intersecting ellipses, often on a gold ground. This general style continued into the 18th century; but in the course of that century the levels of artistic and technical competence sank and the tumblers and spirit bottles, which were the main types produced, can be regarded only as objects of peasant art.

SCALA, New York

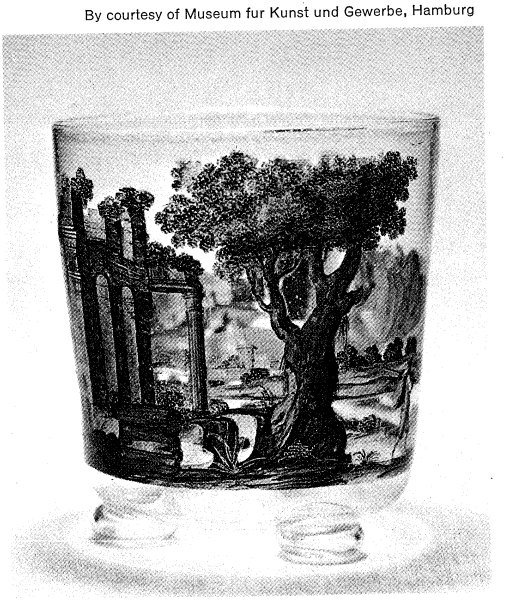


Humpen (enamelled drinking vessels), German, 17th century. (Left) Tankard decorated with a representation of the Trinity. Height 30 cm. (Right) *Reichsadlerhumpen*, decorated with the imperial double-headed eagle. In the Germanisches Nationalmuseum, Nürnberg. Height 28 cm.

A far more sophisticated type of enamel painting was carried on during the third quarter of the 17th century at Nürnberg. There, painting in black or sepia (*Schwarzlotmalerei*)—a technique borrowed from the stained-glass artist—was used to decorate the small cylindrical beakers (often resting on three hollow ball feet), which were a locally favoured shape. Other colours, notably red used in touches with the black, were occasionally employed. The greatest and most original artist of this school was Johann Schaper, who painted delicate architectural and landscape compositions in which a fine point was used to etch in details. The best of Schaper's followers were J.L. Faber, Hermann Bencherlt, Johann Keyll, and Abraham Helmhack, but none of them equalled him in artistic competence. Comparable work appears to have been done, although on a more restricted scale, in the Rhineland, notably by Johann Anton Carli of Andernach. At the beginning of the 18th century *Schwarzlot* painting, often with touches of gold, was practiced in Bohemia and Silesia and reflected the changing fashions in the decorative arts. Daniel Preissler and his son Ignaz are known to have done this work.

The *Schwarzlotmalerei* technique

By courtesy of Museum fur Kunst und Gewerbe, Hamburg



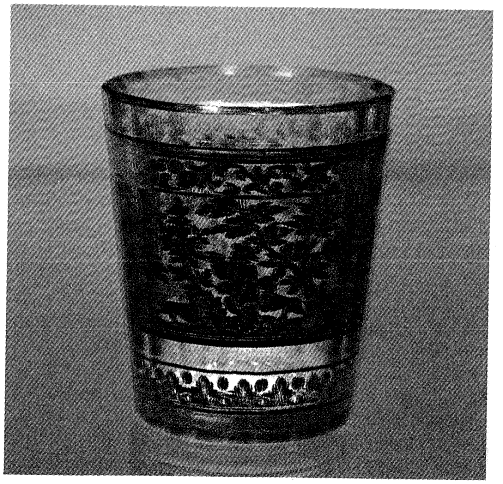
Beaker painted with black and sepia enamel (*Schwarzlotmalerei*) by Johann Schaper. Nürnberg, 1664. In the Museum für Kunst und Gewerbe, Hamburg. Height 8.9 cm.

In the first half of the 19th century the decorators of vessel glass once again borrowed from the stained-glass artist. Samuel Mohn, his son Gottlob Samuel Mohn, and Anton Kothgasser painted the beakers typical of this "Biedermeier" period in transparent enamels and yellow stain.

A technique peculiar to Bohemia in the 18th century was that of the "gold sandwich glasses" (*Zwischengoldgläser*). These were beakers or less often goblets made of two layers of glass, exactly fitting one over the other, between which was sandwiched a gold leaf previously etched with a steel point to the desired design. The earliest work in this technique was anonymous, but late in the century J.J. Mildner employed it with notable success, making gift tumblers decorated with medallions of etched gold or silver leaf (often backed with red pigment) and sometimes also engraved on the wheel or with the diamond point.

The *Zwischengoldgläser* technique

England. Glass was certainly made in England during the later Middle Ages, but most of it was used for church windows (see STAINED GLASS). The vessel glass of the peri-



Zwischengoldglas ("gold sandwich glass"), double-walled beaker decorated with a bear hunt, Bohemian, *c.* 1730. In the Kestner-Museum, Hannover, West Germany. Height 8.9 cm.
By courtesy of Kestner-Museum, Hannover, West Germany

od has not been much studied and is only imperfectly understood. Only by the second half of the 16th century does the picture become clearer. Two lines of development may be traced in this period. One is the glass of German waldglas type, made in the woods that supplied the furnaces with fuel and a source of potash. These glasses were made by workers whose traditions were those of Lorraine and northern France. Much of their production was of window glass, but they also made vessels in a modest variety of shapes and modes of decoration. Chief among them was a tumbler-like drinking glass with a low, double foot-rim produced by pushing in the bottom of the bulb from which the glass was made; this might be decorated either by mold-blown diaper (overall repeat) patterns, by swirled ribbing imparted by mold blowing and subsequent twisting, or by a zone of trailed threading below the rim. Applied notched ribbons or small circular motifs also were used. Small bottles of mold-blown hexagonal section or of flattened ovate form with diagonal ribbing also were made. The second line of development was that of the international Venetian style brought by immigrant Italians; this, however, in time acquired an English idiom. The work was done mainly in London.

17th-century English glass

In the 17th century these two traditions were welded into one, spurred by the proclamation of 1615 that forbade the use of wood in glass furnaces, as well as in certain other industries, in an effort to prevent the deforestation of the country. Thereafter, with coal as the sole means of fusing glass, glassworks tended to be located where coal deposits (and the frequently concomitant fire clays for making glass pots) were abundant. Since such areas for the most part were those that have been continuously occupied by industry (*e.g.*, the Stourbridge area and Tyneside), exploration of the early glass factory sites has seldom been practicable. Little, therefore, is known of provincial glassmaking in England in the 17th century, but it is clear that Venetian influences gradually replaced the earlier Waldglas tradition, which had depended on supplies of wood. Some idea of the new style may be gained from the fragments of glasses often excavated in London and other cities. It is frequently difficult to distinguish between an English glass and an imported European one, although a certain coarseness may be taken as symptomatic of English make.

During the first half of the 17th century, glassmaking was among the English industries for which monopoly rights were granted by the crown; the greatest of a series of monopoly holders was Sir Robert Mansell, who effectively controlled the industry from 1623 until his death in 1656. After the Restoration, although some monopolies were granted for certain categories of glasswares, an increasingly important role in the English industry was played by the Worshipful Company of Glass Sellers (reincorporated in 1664), which was able to keep closely in touch with the needs of the English market. Its members seem to have laid stress on simplicity of shape and durability of material, as appears from the correspondence of one of them, John Greene, with his suppliers in Venice. Dissatisfied with the quality of glass supplied to them and no doubt also anxious to make England independent of foreign sources of both finished glass and raw materials, they commissioned George Ravenscroft to make experiments with native materials in the hope of evolving a more solid glass than the Venetian and one that more closely resembled rock crystal.

Discovery of lead glass

Ravenscroft was completely successful; his crucial discovery was the value of adding lead oxide. His "glass of lead," evolved about 1675, was perfected toward the end of the century and set a standard for the rest of Europe. It was solid and heavy and more durable than the Venetian-type glass, which it progressively displaced. It was also characterized by brilliance and dark shadow paradoxically combined. It was slower to work than the Venetian glass and gradually the Venetian idioms were dropped from English glassmaking in favour of a genuine native style. This style is best exhibited in the drinking glasses that, by the end of the 17th century and the beginning of the 18th, constituted the chief glory of the English industry. These often massive baluster-stem glasses were composed of a usually funnel-shaped bowl and a stem compiled of any of a large variety of pear-shaped and bulbous knops (ornamental knobs). In their simplicity and the harmony of their proportions they rank among the classics of the Queen Anne style.

Toward the middle of the 18th century, taste in the arts generally inclined to lighter forms, and in glass this tendency was given additional impetus by an excise (1745–46) levied on glass by weight. Drinking glasses became slighter, the bowls smaller, and the stems taller and more slender. The loss in architectonic values was often offset by extraneous decoration. At first this tended to be concentrated in the stem. Bubbles of air had sometimes previously been enclosed in a knop forming part of the stem of a wineglass, and these bubbles were now drawn out and twisted so that they formed a cable of air ribbons inside a cylindrical stem. Stems of this type were popular about the middle of the century. Just before 1750 a stem decorated with threads of opaque white glass instead of air twists came into favour. These stems were made by much the same techniques as the Venetian latticinio glass. They remained in fashion until about the time of the second Glass Excise Act in 1777, which imposed a tax on the opaque white "enamel" glass, previously exempt.

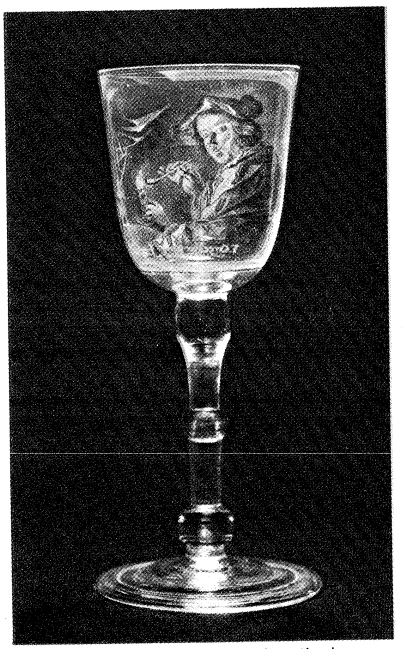
These forms of ornament had been restricted to the stems of glasses, but other methods of decoration were simultaneously evolved to embellish the whole glass. First of these was engraving, which had been sporadically practiced in England as early as the end of the 17th century. This work and the inscriptions, coats of arms, and arabesque borders in German style that were engraved during the first 20 years or so of the 18th century were undoubtedly the work of immigrant (probably German) artisans. By 1735, however, at least one English engraver was capable of executing such commissions and from about this time engraving on glass began to take on a more English character. An artless use of floral motifs, chinoiseries (Chinese themes), and scenes from country life is typical of the engraving of the third quarter of the 18th century, as were the frequent representations on glasses of Jacobite themes—portraits of the Old and Young Pretenders (James III and Charles Edward), the rose with buds, the honeysuckle, and the other flowers used in the symbology of the Stuart cause, together with the mottoes of such "loyal" societies as the Cycle Club.

Development of engraving and enamelling in England

Engraving never reached great heights in England, but English glasses were in demand by engravers in Europe, particularly in the Netherlands, where the work of at least one notable artist—Jacob Sang, of Amsterdam—was almost exclusively done on imported English drinking glasses. English lead glass also seems to have been

particularly favoured by the Dutch diamond-point engravers, whose work in this period was executed almost exclusively in stipple (*i.e.*, dotted engraving). The chief masters of this delicate art, in which the design seems no more than a bloom on the surface of the glass, were Frans Greenwood of Dordrecht, the originator of the style, and David Wolff of The Hague, whose work, if uninspired, is of high technical accomplishment.

By courtesy of Museum fur Kunst und Gewerbe, Hamburg



Glass goblet with diamond-point stipple engraving; signed "F. Greenwood fecit 1764." Holland. In the Museum für Kunst und Gewerbe, Hamburg. Height 28 cm.

Enamelling, the second decorative technique of foreign inspiration, began to be used on English glass in the mid-18th century. It embellished opaque white glass in imitation of china—a type of work usually associated with the name of Michael Edkins, a Bristol artist, but in fact done in many parts of the country. Perhaps the most original work in this medium was done on clear glass by members of the Beilby family of Newcastle upon Tyne during the 1760s and 1770s. Their rendering in usually blue-toned white enamel of ruins, trophies of arms, and rural pastimes, often framed in scrollwork of the utmost delicacy, is one of the best things in English Rococo glass. Gilding was also used at this time to decorate glasses, usually with simple designs of vines and grapes.

Importance of cutting as a decorative technique

These ornamental techniques, however, were of ephemeral growth in England. Far more significant than any of them, because more firmly rooted in the very nature of English glass, was the art of cutting. Although literary references to cut glass occur before 1720, the earliest known pieces can hardly be dated much before 1725. On them the cutting is mainly confined to brims and feet, which are scalloped or notched; or, on wineglasses, to the thicker parts of the glass, such as the stem, which might be fluted or cut in an allover pattern of flat diamonds. Throughout the period from about 1745 to 1770, shallow cutting was the norm. Diamonds, hexagons, flutes, and scale pattern were combined with segmental lunate cuts (produced by holding the glass at an angle to the cutting wheel) and with triangular and diamond motifs in very low relief. All of these elements could be combined to produce designs of great complexity and richness. This period marked the golden age of English cutting.

About 1770 a plainer style, employing mainly flutes, responded to the rising Neoclassical fashion in the other arts. The flutes were sometimes combined with diamonds in relief. When further taxes were imposed on glass in 1777 and 1781 and when in 1780 trade between England and Ireland was freed, it was this relief-diamond style that was taken up in Ireland by the glasshouses founded there. The Irish glassworkers could afford to be more lavish with their material and on this thicker glass increasingly deeply cut diamonds and other relief motifs could be produced. About the turn of the century the diamonds began to be reduced in size and to be incorporated into a diaper pattern covering whole areas, often alternating with fields of larger truncated diamonds, the surfaces of which were themselves diversified with cut crosshatching. Such designs were often combined with deeply cut horizontal grooves. These styles, which were subsequently followed in England as well as in Ireland, finally led to a complete breaking up of the face of the glass into points and ridges, with increased prismatic effect but with a disastrous loss of surface quality, which is one of the peculiar beauties of glass. The prismatic brilliance was enhanced by the progressively greater purity and whiteness of the glass made during the second quarter of the 19th century. The temptation to cut ever more deeply and with greater complexity finally seduced the glassmakers into producing the "prickly monstrosities" of the Great Exhibition of 1851.

Throughout the 18th century there had been great admiration in Europe for English lead "crystal," and in the second half of it some of the European glasshouses were using lead oxide and had contrived to produce a comparable material. English cut glass was admired and exported, and the styles of cutting of the late 18th and early 19th centuries were much imitated abroad.

(R.J.Ch.)

United States. Glassmaking was apparently the first industry to be transplanted from Europe in the wake of the Spanish conquerors. As early as 1535 glass was being made at Puebla in Mexico, and in 1592 a glasshouse was located in the territory of the Río de la Plata in the town of Córdoba del Tucumán, Argentina. Broken glass, undoubtedly of European origin, was remelted at Córdoba and fashioned into various objects including thick, semi-transparent flat glass.

The London Company of Virginia set up a glasshouse in Jamestown in 1608 for the manufacture of "glasses" and beads. A "tryal of glasse" was sent off to England before the winter of 1609, the "starving time" during which 440 of the colony's 500 inhabitants died. In 1621 the company tried again and, although the second attempt was more carefully planned, it too failed. Excavation of the site has revealed that glass was melted in considerable quantities though no evidence of glass bead manufacture has been found.

South Jersey-type glass. For more than a century after Jamestown, there was little American glass. The earliest successful glasshouse was begun in 1739 by Caspar Wistar in Salem County, New Jersey. The fact that his works produced only humble utilitarian vessels and windowpanes saved him from extermination by the "lords of trade." Wistar died in 1752, after which the factory was operated by his son Richard. It was offered for sale in 1780. Although few, if any, objects exist that can be assigned to the Wistar Glass Works with certainty, it is important as the cradle of the American glass known today as South Jersey type. That glass is the work of individual glassblowers using ordinary bottle or window glass to make objects of their own design. Applied glass and, occasionally, pattern molding were the only feasible means of decoration, and the resultant loopings and threadings are typical of European traditions. One decorative device, the lily pad, is of particular importance, as no European prototype is known. A hot mass of glass applied to the base of the bowl is pulled up around the sides in a series of projections in which the bowl appears to rest.

Wistar's glasshouse

The second great name in early American glass is Henry William Stiegel. In 1763, 13 years after his arrival in America and after several years in the iron business, he built his first glasshouse in Lancaster County, Pennsylvania. Like Caspar Wistar, Stiegel at first was concerned with the manufacture of bottles and windowpanes. With the founding of his second house at Manheim, Pennsyl-

Stiegel glass

vania, in 1765, however, he ventured into the table-glass business. No longer beneath the notice of the "lords of trade," he reported to them in 1767 that the glass he made was both inconsiderable in quantity and ordinary in quality.

This report is in sharp contrast to the many advertisements in which he favourably compares his wares with English imports. Encouraged by the patriotic adoption of the non-importation agreement, Stiegel built a third glasshouse, the American Flint Glass Works, also located at Manheim and completed in 1769. Adverse economic conditions, caused by both the approaching war and the colonial preference for imported tablewares, brought final failure in 1774.

Few pieces can be attributed with confidence to the Stiegel factories, and, like that of Wistar, his name survives as the founder of a tradition. Stiegel-type glass is characterized by the use of clear and artificially coloured glasses; by extrinsic decoration such as engraving, enamelling, and pattern molding; and, in general, by two distinct styles, one employing English and the other German techniques and decorative devices. Certain mold-blown patterns, such as the diamond daisy and daisy in hexagon, are believed to have been originated at the Stiegel houses, no European prototypes having been identified.

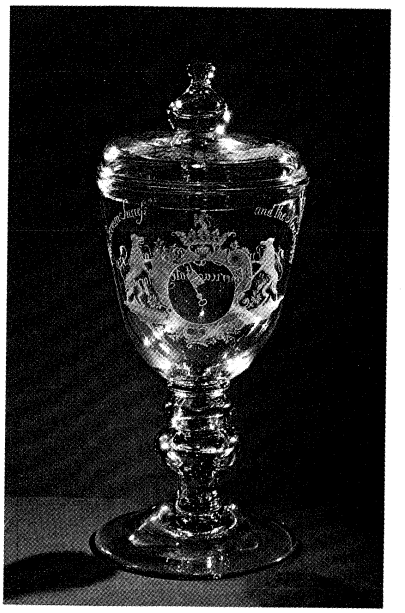


Sugar bowl with cover, pattern molded, attributed to the glassworks of Henry William Stiegel, Manheim, Pennsylvania, c. 1765–74. In the Corning Museum of Glass, New York. Height 15.6 cm.

By courtesy of The Corning Museum of Glass, New York

Post-Revolutionary glassworks. Before the turn of the century, several other glassworks were founded, but few survived the Revolution. These houses were devoted largely to the manufacture of bottles and window glasses and, with the notable exception of the New Bremen Glassmanufactory, most of the offhand (*i.e.*, shaped by hand) pieces that can be tentatively assigned to them are of the South Jersey tradition. Three of these enterprises are of particular importance. First, the New Bremen (Maryland) Glassmanufactory, founded by John Frederick Amelung and Company, is of special interest as many of its presentation pieces are both signed and dated as well as being among the finest produced in the United States before 1800. Originally from Bremen, Germany, Amelung was persuaded to go to America for the express purpose of founding what he believed to be a much-needed industry. By 1785 his works offered green and white hollow ware for sale; by 1795 the glassworks themselves were offered for sale. One of the most famous pieces in the history of American glass is the Bremen pokal (the German word for goblet), blown and engraved in 1788 and sent back to Amelung's financiers in Bremen (now in the Metropolitan Museum of Art, New York City), probably the only return they ever received on their investment.

The second factory of importance, later known as the



Bremen pokal, presentation piece engraved at the Amelung works near Frederick, Maryland; inscribed "Old Bremen Success and the New Progress and New Bremen Glassmanufactory—1788—North America, State of Maryland." In the Metropolitan Museum of Art, New York. Height with cover 28.6 cm.

By courtesy of the Metropolitan Museum of Art, New York, Rogers Fund, 1928

Olive Glass Works, Gloucester County, New Jersey, was completed in 1781 by former employees of the Wistar Glass Works, the Stanger brothers. In addition to the many fine South Jersey pieces attributed to this house, it is of interest because of its long history, eventually becoming part of the Owens Bottle Company, a forerunner of the Libbey-Owens-Ford Glass Company.

The Pitkin glassworks

The third notable venture begun before 1800 is the well-known works associated with the name Pitkin. Erected at East Hartford, Connecticut, near the Connecticut River in 1783, it was intended for the manufacture of window glass, but in 1788 it was converted to the manufacture of bottles and flasks. The factory thrived until 1830 and is best known for the half-post (*i.e.*, dipped twice up to the neck) ribbed flasks in natural browns, ambers, and greens. Today the word Pitkin denotes a type of flask and not a specific glassworks.

After the War of 1812. The few houses that survived the 1790s and the depression after the War of 1812 had multiplied to more than 90 by 1830. For convenience, the glassworks are divided into three geographical groups: New England, the Middle Atlantic, and the Midwest. Until that time, they had produced little more than simple imitations of European glasses, at best interesting and often very handsome combinations of various decorative devices and traditions. The big change occurred between 1830 and 1840 with the production of fine lead glass, the use of the full-size incised mold, and, finally, the pressing machine.

Bakewell's fine glass

The glasshouse known as Bakewell's was synonymous with the finest achievements of the revived industry. Originally established in 1808 in Pittsburgh, the first city to use coal for fuel in glassmaking, the company survived under several different firms until 1882. Glass cutting, introduced to Pittsburgh by William Peter Eichbaum, glass cutter to Louis XVI, was an important part of Bakewell's operation. In addition to being the first American company to supply the White House, serving President James Monroe in 1817, Bakewell's produced such specialties as lead-glass tumblers with "sulphides" (cameo insertions of white fireproof material in an envelope of glass) in the bases portraying the Marquis de Lafayette, Andrew Jackson, New York governor George Clinton, Benjamin Franklin, and George Washington.

The company also held the first patent on mechanical pressing, granted in 1825 for a device to make knobs.

Fine lead glass in the New England area was first successfully made in the South Boston works of the Boston Crown Glass Company. Thomas Cains was making flint glass there in 1813. He left the firm in 1824 to found the Phoenix Glass Works in South Boston, which survived until 1870. One particular device usually associated with the Boston manufactories of this period is the guilloche, or chain, employed in the decoration of a large variety of tableware.

The New England Glass Company, founded in 1818 in Cambridge, Massachusetts, maintained the same high standards as Bakewell's, even to the point of making glass for President Monroe. This factory held the second patent on a device for mechanical pressing, granted in 1826, and produced quantities of pressed glass of all types before it was moved to Toledo, Ohio, in 1888. The New England Glass Company was also famous for its very fine free-blown and engraved glass. In addition, vessels were made there in the so-called blown three-mold technique, in which decorative designs adapted from cut-glass patterns of the period were impressed in the glass by blowing in molds hinged in two, three, or more sections. More than 400 different molds have been determined and grouped according to pattern under three primary headings: geometric, arch, and baroque. By 1830 this type of production was being replaced by the much more efficient pressing machine.

The three-mold technique

Deming Jarves, one of the founders of the New England Glass Company, founded the Boston and Sandwich Glass Company in 1825. Because of his *Reminiscences of Glassmaking*, extensive advertisements, and thorough excavations of the factory site in Sandwich, Massachusetts, more is known about this particular factory than any other of the period. Consequently, "Sandwich" has become a generic term for pressed glass even though many other factories used identical machinery and, in some cases, identical molds. Jarves's first patent on a pressing device, the fifth to be granted, was received in 1828 after the Boston mold maker Hiram Dillaway entered his employ. Jarves founded the Mount Washington Glass Works in 1837 in New Bedford, Massachusetts, and the Cape Cod Glass Works in 1857.

Among the outstanding makers of fine lead glass in the middle Atlantic states were the Brooklyn Flint Glass Works of John L. Gilliland and Company and the Dorflinger Glass Works. Gilliland, a partner in the Bloomingdale Flint Glass Works, sold out in 1823 and founded his own works in Brooklyn, New York. In 1864 two members of the Houghton family acquired controlling interest, and in 1868 the works was moved by barge to Corning, New York, to form part of the now famous Corning Glass Works.

Historical flasks. Perhaps the most fascinating aspect of American glass is a series of pictorially molded bottles known as historical flasks, produced between 1815 and 1870. Three hundred ninety-eight different examples have been divided into the following groups: (1) Masonic; (2) emblems and designs related to economic life; (3) portraits of national heroes and designs associated with them and their deeds; and (4) portraits of presidential candidates, emblems and slogans of political campaigns. In the second group are a number of interesting designs encouraging the United States system of better internal transportation and high protective tariffs. Among the 16 celebrities portrayed in the third and fourth groups are Jenny Lind, the Swedish singer; Lajos Kossuth, the Hungarian patriot; Marquis de Lafayette, the French hero of the American Revolution; and the notorious Thomas W. Dyott, a patent-medicine vendor and bottle manufacturer. These containers were used also as propaganda during political campaigns. William Henry Harrison is pictured in this connection with other impedimenta relative to the "Log Cabin and Hard Cider" campaign of 1840.

The first 25 years of pressed glass, 1825 to 1850, are referred to by collectors as the "lacy period." A milestone within this brief span occurred in 1830 with the development of the cap ring, a device that ensured uniform thickness at the edge of each piece regardless of the amount of glass forced into the mold. Before this date most impressed designs were inspired by Anglo-Irish cut glass, often coupled with popular American devices such as a sheaf of wheat. Between 1830 and 1840 the objects were thinner and more lavishly decorated, often including elaborate motifs based on the classic and Gothic revivals. Because of the unpleasant surface left by the mold and in an effort to imitate the brilliance of cut glass, unstippled areas were filled in with overall lacelike patterns; hence the term "lacy." About 1840 economic conditions forced glassmakers to revert to cheaper molds and simpler geometric forms and to abandon the stippled patterns.

The "lacy period" of pressed glass



"Portrait of Kossuth" flask, Bridgeton, New Jersey, 1840–55. In the Corning Museum of Glass, New York. Height 17.5 cm.
By courtesy of the Corning Museum of Glass, New York

During this period the mechanical press became firmly established, and by mid-century glassmaking had become one of the United States' new mass-production industries.

(T.S.Bu.)

MID-19TH TO 20TH CENTURY

The modern history of glass can be said to begin in the middle of the 19th century with the great exhibitions and with the new self-consciousness in the decorative arts that they expressed. Glassware was being publicly discussed in art journals and collected in museums, and this new spirit of awareness led to a greatly increased exchange of ideas among the leading glass centres and to the borrowing of ideas from the past.

In some degree the established glass-producing centres were still concerned in the modern period with the styles of glassware for which they had achieved an earlier reputation. The English glasshouses continued their production of deeply cut crystal; engraved glass and to a lesser extent coloured and painted glass were given the greatest attention in central Europe; the Venetian glasshouses at Murano were the leading exponents of furnace-manipulated glass. But alongside these traditional methods of using and decorating glassware can be discerned the development of a renewed interest in the beauty of the material itself. Expressed in various ways, in the use of thick masses and in internal figuring and patterning, this interest has been the keynote of the most significant modern contributions to the art of glass.

Pressed glassware, which had been first made with great promise in the first half of the 19th century, was being widely made in the middle of the century, and later, as a cheap imitation of cut crystal. The decorative possibilities of the process continued, however, to be exploited in a variety of popular wares; and in the 20th century a series of new simple forms of pressed glassware appeared that had been expressly designed in relation to the characteristics of its manufacture.

Great Britain. The Great Exhibition of 1851 was the culmination of a period of intense activity in the British

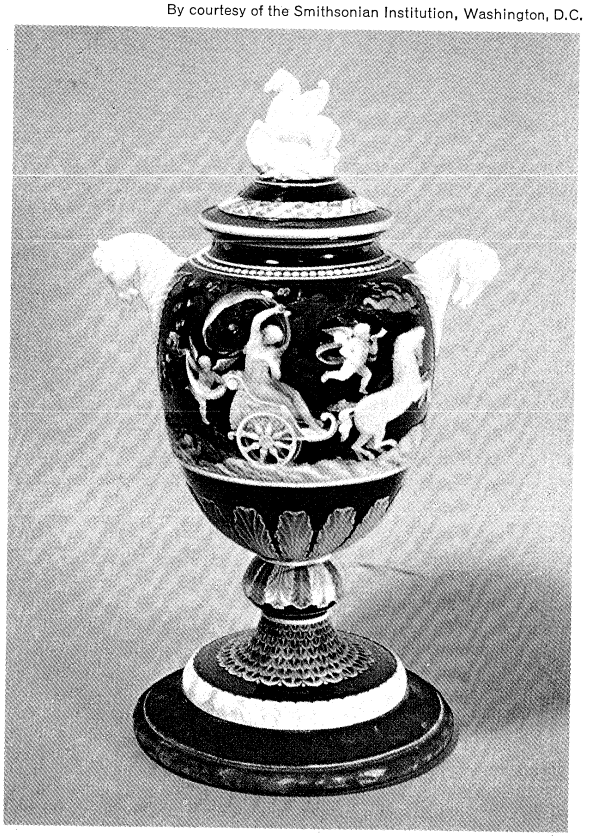
glasshouses. The excise duty on glass had been removed in 1845, and the British glassmakers were determined not only to excel in their traditional deeply cut crystal but also to rival the Bohemians and the French in coloured, layered, and enamel-painted wares. Probably the most enterprising of the English glassmakers of the period was Benjamin Richardson, of Wordsley near Stourbridge; surviving pieces of this period from the Richardson firm include some admirable painted and engraved pieces as well as crystal wares deeply cut in elaborate patterns.

Probably in reaction against the banality of pressed-glass imitations of cutting, the most sophisticated work in crystal during the later 1850s through the 1870s was decorated by engraving, often carried out by immigrant Bohemian craftsmen.

The Venetian style of furnace-manipulated glass was also exerting a strong influence. It can be seen, for instance, in the development of the elaborate Victorian centrepieces in the 1860s and 1870s. In some degree the Venetian style was also an influence, alongside that of the Far East, in the fashioning of the fancy wares that were made in Great Britain—as it was in the United States and elsewhere—during the 1880s and 1890s. These wares were often given specific trade names and were mostly made in the English Midlands by firms such as Thomas Webb & Sons of Stourbridge and John Walsh Walsh of Birmingham.

A striking form of mid-Victorian virtuosity was the cameo glass produced by Stourbridge glassworkers. This work, inspired by the Portland vase, required a lengthy process of etching and carving, normally through an opaque-white-glass layer to leave a white carved design in relief on a dark-coloured glass body. The first important pieces, such as the "Pegasus vase" (Smithsonian Institution, Washington, D.C.), were produced in the 1870s by John Northwood, and in the later part of the century the most distinguished cameo work was carried out by George Woodall.

By courtesy of the Smithsonian Institution, Washington, D.C.



The "Pegasus vase," carved in cameo relief by John Northwood of Stourbridge, England. In the Smithsonian Institution, Washington, D.C. Height 54.6 cm.

The influence of the Arts and Crafts Movement was toward the use of plastic forms and furnace decoration, which the English art critic John Ruskin had advocated in *The Stones of Venice*. In 1859 Philip Webb designed for William Morris some simply formed tableware that was made at the London glassworks of James Powell & Sons. From about 1880 this glassworks was under the control of Harry J. Powell who, working until World War I, developed a simple, dignified style of handmade blown glass, which was subsequently continued in designs by Barnaby Powell, James Hogan, and others.

Influence of the Arts and Crafts Movement

During the 1930s and after World War II other firms produced work in which a restrained and distinctively modern approach was made to the cutting of faultless crystal glass. Notable designs were produced by Keith Murray for Stevens & Williams shortly before World War II and by David Queensberry (12th marquess of Queensberry) for Webb Corbett in the 1960s. Among the more distinguished glass engraving may be mentioned the diamond-point fantasies of Laurence Whistler and the work of John Hutton, made by a movable wheel held in the hand, such as his great screen in the new Coventry Cathedral. The appearance of new factories in the 1960s, concerned primarily with form and colour, widened the scope of British glass design; and at this time the glass-teaching schools were especially significant as centres for original work by individual artists.

United States. By the middle of the 19th century, American pressed glass was already a disturbing influence on the design of the finer wares. Its decoration was by that time mostly designed in imitation of cut glass, and the process of fire polishing was being used to give a surface almost as smooth as that of blown glass. During the succeeding decades pressed-glass designs became increasingly complicated. This tendency was accentuated in the soda-lime glass that William Leighton began to use for pressed work at Wheeling, West Virginia, in the 1860s, and that was later widely used in the western glasshouses for the cheapest coloured wares.

Importance of pressed glass

In general the finer wares of the early part of the period were similar to those of the Biedermeier and later styles of Europe. The New England Glass Company at Cambridge, Massachusetts, was employing many European craftsmen and was producing a wide variety of richly decorated layered and engraved wares. At the Boston and Sandwich Glass Company layered glass was extensively used for large kerosene lamps. The effect of the competition of pressed glass on cut-crystal work can be seen in the appearance of fine-line cuttings, and, during the period up to the Philadelphia Centennial Exposition of 1876, the most significant crystal work was decorated by engraving. Louis Vaupel and Henry S. Fillebrown were two notable engravers employed by the New England Glass Company from 1856 and 1860, respectively.

At the time of the Centennial Exposition, cut-crystal work began to revive, and by 1880 a considerable boom in its production had developed—a boom that was to continue throughout the 1880s and 1890s. New industrial methods contributed to the production of crystal glass of flawless quality and to its deep cutting with mathematical accuracy in elaborate designs. Among many others, a noteworthy producer of this type of glass in the 1890s and later was the Libbey Glass Company, the successor to the New England Glass Company. Later, in the early years of the 20th century, intaglio cutting in crystal became popular, and work in this expensive process was carried out in a number of cut-glass factories such as the T.G. Hawkes Glass Company at Corning, New York.

As in Great Britain and elsewhere, a great amount of glass was made in fancy forms and colours in the 1880s and 1890s. Although undisciplined and often tasteless, such glass nevertheless preserves perhaps more than any other the flavour of the period. These wares, often bearing specific names such as Pomona, Burmese, and Peachblow, were made by such firms as the New England Glass Company, the Mount Washington Glass Company at New Bedford, and the Hobbs, Brockunier Company at Wheeling, West Virginia.

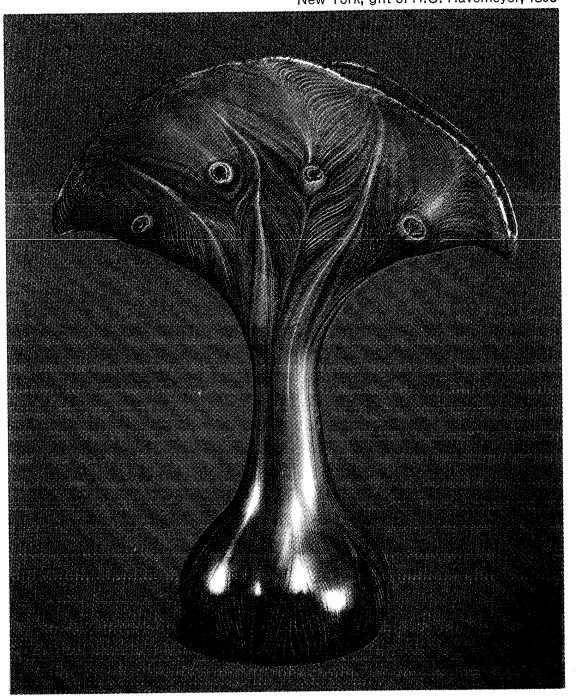
Although belonging essentially to the category of the

Tiffany's Favrile glass

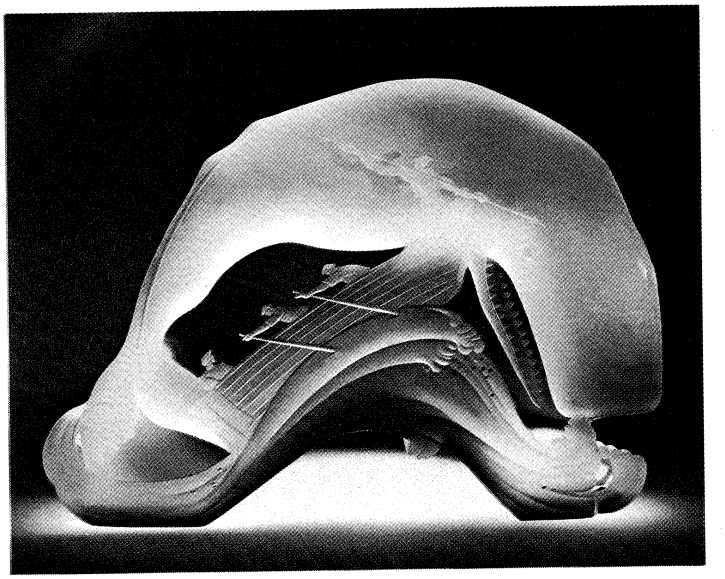
fancy glasses, the Favrile glass of Louis Comfort Tiffany represented an altogether higher level of achievement both in its shapes and in the colouring and figuring of the glass. It was first shown to the public in 1893, and in pieces that were produced a few years later Tiffany achieved an outstanding expression in glassware of the Art Nouveau style. Much of his work was in a heavily lustred glass that was considerably admired abroad, especially in central Europe where it created a new fashion.

From the period of World War I onward, new forms of pressed glassware appeared in simple, satisfying designs appropriate to their purpose and the process of manufacture, such as the Pyrex ovenware shapes of the Corning Glass Works. The Steuben Glass Company of Corning was known for fancy glasses designed by Frederick Carder, until in 1933 the company was given a change of direction by Arthur Amory Houghton, Jr., who, with the help of John Monteith Gates and the sculptor and designer Sidney Waugh, aimed to produce glass with engraved decoration that would rank as fine art. Other noteworthy modern American work included simple designs in blown glass by the Blenko Glass Company of Milton, West Virginia, and enamel patterned bowls by the independent artist Maurice Heaton. The appearance in the United States of studio blown glass, produced by individual artists, was a development of international significance. It was initiated in the 1960s notably by Harvey Littleton and Dominick Labino and included work such as that produced personally by Joel P. Myers at the Blenko Glass Company.

By courtesy of The Metropolitan Museum of Art, New York, gift of H.O. Havemeyer, 1896



Peacock vase, Favrile glass by Louis Comfort Tiffany. In The Metropolitan Museum of Art, New York. 35.9 × 29.2 cm.



"Moby Dick," Steuben crystal sculpture designed by Donald Pollard and Sidney Waugh, first piece made in 1959. Length 28.6 cm.
By courtesy of the Corning Museum of Glass, New York; photograph, Steuben Glass

Czechoslovakia, Austria, and Germany. In the middle of the 19th century the glasshouses of central Europe were producing a great variety of the layered and coloured wares that had become particularly associated with Bohemia in the preceding Biedermeier period. They were also producing a great amount of cut crystal glass in the deeply cut English style, and indeed work of this nature continued with little change throughout the modern period.

Revival of engraving

A revival of the indigenous art of engraving was initiated by Ludwig Lobmeyr, who from 1864 was in control of the Viennese firm of J. and L. Lobmeyr. His first opportunity came at the Paris International Exhibition of 1867, and his reputation was firmly established at the Vienna International Exhibition of 1873. He commissioned designs for his glasses from the leading Viennese architects and painters of the time, and his work was carried out by the finest craftsmen in Bohemia and Austria.

The Art Nouveau style, which went under the name of Jugendstil in central Europe, made a deep impression on central European glassware. The work made around the turn of the century abounds in slender shapes and flowing organic motifs. Glasses designed by Karl Köpping in Berlin, with long, waving stems and tulip-like bowls, were perhaps the extreme instance of Art Nouveau style applied to glassware. In 1897 an exhibition of glass by Tiffany was shown at several of the museums in the area. Not only the forms of the Tiffany glasses but also their figured and heavily lustred material attracted great interest. Several factories started making a similar heavily lustred glass, including the firm of J. Lötz' Witwe of Klášterský Mlýn (Klostermühle), which won a *grand*

By courtesy of the Victoria and Albert Museum, London



Bohemian layered-glass vase, painted and gilt by Wilhelm Hoffmann, Prague and Vienna, c. 1850–60. In the Victoria and Albert Museum, London. Height 42 cm.

prix at the Paris Exhibition of 1900 with this type of glassware.

From around 1900 onward a movement toward a modern purist approach to glass was largely fostered by the work of designers connected with the Vienna Kunstgewerbeschule (School of Industrial Art). Men such as Kolo Moser and Josef Hoffmann, who were also closely associated with the Vienna Werkstätte (Workshop), were designing glasses in simple rational forms. Much initiative in this movement was shown by the firms of E. Bakalowits Söhne of Vienna and J. Lötz' Witwe. The Czech architect Jan Kotěra was influential in the modern design of glass, and in the early years of World War I the Czech Artěl organization of artists and architects was concerned with the design of glass in a forward-looking Cubist manner.

Post-World War I glass

After World War I the outstanding figure in Czech glass art was Josef Drahoňovský, who was professor at the Prague School of Industrial Art. He was essentially a sculptor, and most of his glass designs were for sumptuously engraved glass of a monumental quality. His colleague in Prague, Jaroslav Horejc, designed for engraved work of a broadly similar character, some of it for the Lobmeyr firm of Vienna. The decades after World War II saw considerable activity in glass design. Notable artists in the 1960s were Stanislav Libenský, René Roubíček, Pavel Hlava, and Václav Cígler.

In Austria after World War I the Lobmeyr firm under the control of Stefan Rath produced many engraved and relief-carved pieces designed by artists such as Ena Rottenberg, Lotte Fink, and Vally Wieselthier. Lobmeyr also produced some of the best designs of Michael Powolny, who had his own workshop and had designed for the firm of J. Lötz' Witwe.

In Germany the outstanding engraver and glass carver of the period after World War I was Wilhelm von Eiff, who was professor at the Stuttgart Kunstgewerbeschule, while Bruno Mauder of the glass-teaching school at Zwiesel in Bavaria was a strong influence for the use of natural and appropriate glass forms. Some fine tablewares were produced, especially after mid-century, by designers such as Wilhelm Wagenfeld, Richard Süssmuth and Heinrich Löffelhardt. An interesting development was that of the Rosenthal firm, which used the Finnish designer Tapio Wirkkala and the Dane Bjørn Wiinblad to effect in each case matching glass and porcelain suites of the firm's own manufacture.

France. In France, as in central Europe and in England, the production of fine glassware in the middle of the 19th century was mainly divided between cut crystal and coloured wares. The "opalines," the semi-opaque white and coloured wares, often with elaborately painted and gilt decoration, were especially popular; and it was during these years that the French paperweights, containing coloured patterns, became internationally known and admired. The larger factories, particularly Baccarat and Saint-Louis, continued to participate in the international fashions of the rest of the century and beyond. But in France inventive genius manifested itself mainly in the work of individual artists and thereby a new spirit was introduced into the modern conception of glass.

In the late 1860s and 1870s three individual artists were experimenting in glasswork, and all of them were represented in the International Exhibition of 1878 in Paris. The first was Joseph Brocard, who was studying the enamelling of glass and whose main ambition was to reproduce medieval Syrian glass. The second was Eugène Rousseau, a commissioning dealer in ceramics who had turned to glasswork at the end of the 1860s and was at the height of his achievement in the years *c.* 1880. Typically his glasses were thick walled and translucid, often with interior crackling and shot with random streaks of colour. In 1885 he associated with E. Léveillé, who continued to work in a similar style after Rousseau's death in 1891. The third of the individual artists at the

Gallé glass

1878 exhibition and the best known of them was Émile Gallé of Nancy, who had been experimenting in glasswork since about 1867. His earliest work was in clear glass, lightly tinted and decorated with enamel and engraving. But he soon developed the use of deeply coloured, almost opaque glasses in heavy masses, often layered in several thicknesses and carved or etched to form plant motifs. His work reflected the prevailing interest in Japanese art and with its frequently asymmetrical form contributed largely to the Art Nouveau of the end of the century. In this period much of Gallé's manner was reflected in the glassware produced on a more commercial basis by the firm of Daum Frères of Nancy.

A number of French artists successfully explored the use of *pâte de verre* (powdered glass fired in a mold). The pioneer in its use was Henri Cros, who was working near the end of the 19th century. It was later the medium for important work by Albert Dammouse and François Décorchemont.

Among the later leaders of French glass art was René Lalique, who around the 1920s was producing his most typical work, which is characterized by relief decoration produced by blowing into molds or by pressing. He was a leading advocate of the use of glass in architecture and much of his work was in the form of lighting equipment and in details of interior decoration. The work of his contemporary, Maurice Marinot, was more in the tradition of Rousseau, with heavy, thick-walled vessels in strong forms often with boldly cut-away abstract decoration; and Henri Navarre in the 1930s was producing work of a similar monumental nature.

The most significant work of Jean Luce and Marcel Goupy, designers of glass and ceramics, was in the production of elegant tablewares. For a long period André Thuret made glasses in thick plastic forms; and Jean Sala worked in bubbled glass. The firm of Daum was distinguished, after World War II, by its thick clear glass vessels manipulated into flowing shapes to designs by Michel Daum.

The Scandinavian countries. Up to the time of World War I the Swedish glass industry produced little original work. The sudden development of modern Swedish glass in the 1920s was attributable mainly to the initiative of the Swedish Arts and Crafts Society that resulted in the employment of the painters Simon Gate and Edward Hald by Orrefors glassworks and Edvin Ollers by Kosta glassworks, both in the glass-producing area of Småland in southern Sweden. The first results were exhibited in Stockholm in 1917 and consisted of handblown, undecorated tablewares, together with the luxury "Graal" glass with internal stained decoration, which had been rapidly developed under Gate's inspiration at Orrefors. It was, however, engraved glasswork, chiefly that designed by Gate and Hald at Orrefors, on which the reputation of Swedish glass was established in the 1920s and particularly at the Paris International Exhibition of Decorative Arts in 1925.

Development of modern Swedish glass

In the 1930s came a change of direction. The Swedish factories began to take less interest in engraving and followed the initiative of the French artists in making thick tinted and figured glasses. In this mode they found their greatest success—attributed largely to their having achieved a system of intimate association between the artists and the glassmaker craftsmen.

At Orrefors additional artists were added to the establishment from 1929 onward, including Vicke Lindstrand, Sven Palmqvist, Nils Landberg, Edvin Öhrström, John Selbing, and Ingeborg Lundin. Each of them worked in an individual style, and in addition to decorative pieces many of them designed tablewares for the subsidiary Sandvik factory. At Kosta important work was produced by Elis Bergh and later by Lindstrand. Gerda Strömberg designed for both Eda glassworks and for Strömbergshyttan. In the 1960s many new methods of forming and decorating glass were explored by young designers; and an element of the current Pop art was discernible, such as in the work of Gunnar Cyrén at Orrefors.

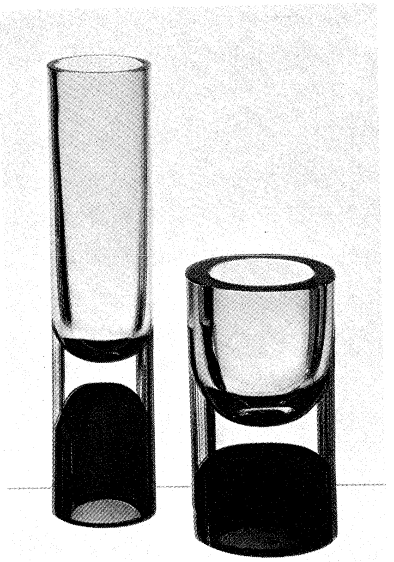
In Denmark the Holmegaard glassworks and in Norway the Hadeland glassworks both followed in some respects the example of Swedish glass. At Holmegaard the movement began in the late 1920s with the appointment as art director of Jacob E. Bang, whose designs included an amount of striking engraved work, and was continued in

the clean forms of his successor, Per Lütken. At Hadeland some distinctive glass was designed by a number of artists including Sverre Pettersen, Willy Johansson, and Arne Jon Jutrem.

Modern Finnish design

In Finland original modern work of great significance has been carried out. Following the example of the Swedish factories, the artist Henry Ericsson was appointed designer at the Riihimäki glassworks in the late 1920s, and Göran Hongell was employed in a similar capacity at the Karhula glassworks in the 1930s. At this time the well-known Finnish artists Arttu Brummer and Alvar Aalto were also concerned in glass design. Shortly after World War I the influential designer Gunnel Nyman

By courtesy of Die Neue Sammlung, Munich



Double-cased glass vases designed by Timo Sarpaneva, Iittala glassworks, Finland, 1957. In Die Neue Sammlung, Munich. Height (left) 30 cm., (right) 17.5 cm.

was producing glasses freely blown in thick masses to form asymmetrical shapes. Other important designers were Tapio Wirkkala and Timo Sarpaneva working for the Iittala glassworks, Kaj Franck for the Nnutajärvi glassworks (trading as Wärtsilä-Notsjö), and Helena Tynell and Nanny Still for Riihimäki. In the 1960s Timo Sarpaneva struck a new note with his sculptures formed from the charred inner surface of wooden molds, while Oiva Toikka designed for Wärtsilä-Notsjö objects of a markedly Pop art nature.

Belgium and The Netherlands. In Belgium the Val-Saint-Lambert factory was an important producer of heavily cut crystal throughout the period. It is also associated with layered work and was particularly prominent with original work of this nature around 1900. Later Charles Graffart designed for it wares made in a variety of techniques, some of them with engraved decoration.

The Dutch glassworks at Leerdam played an important part in the modern movement and followed a line of development distinct from that of the Scandinavian factories. In 1915 the decision was made to invite designs from artists, and by the early 1920s excellent simple tablewares were being made to designs by the architects K.P.C. de Bazel and H.P. Berlage and by the decorative artist C. de Lorm. From the early 1920s onward individually designed pieces called Unica were made; some of the earlier examples were by Chris Lebeau, but most were produced by Andries D. Copier. Later decorative work included designs by Floris Meydam and Willem Heesen.

Variations of traditional Venetian glass

Italy. By the middle of the 19th century, Italian glassmaking had partly revived. In the 1860s the Museo Vetrario was founded at Murano (Venice), and Antonio Salviati began to produce the glasses that attracted much attention at the Paris Exhibition of 1867. These were variations of the traditional Venetian style with elaborate furnace decoration, and the production of glasses of this nature continued at Murano throughout the remainder of the 19th century and beyond.

The 1920s saw the development of a more conscious spirit of artistry in Italian glasswork. Paolo Venini was concerned in producing simple elegant glasses designed by the decorative artist Vittorio Zecchin, and G. Balsamo Stella and his Swedish wife Anna were producing engraved work. In later years, both before and after World War II, much research was done in new methods of colouring and figuring glass; the results were seen in the glasses designed by Ercole Barovier for the firm of Barovier & Toso and in those designed by Giulio Radi for the firm Arte Vetraria Muranese.

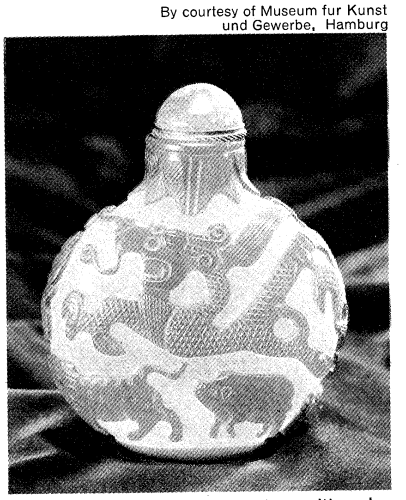
From the Venini firm, presided over by Paolo Venini until his death in 1959, came many interesting innovations, such as the colourful glasses designed by Carlo Scarpa and by Fulvio Bianconi and an interesting series by the Finn Tapio Wirkkala. For the firm of Vistosi some striking modern glasses were designed by artists such as Peter Pelzel and Alessandro Pianon. Some of the work, such as a series of vases designed by Flavio Poli for Seguso Vetri d'Arte, showed some influence from the thick-glass techniques of the north, but the modern Italian glass mostly retained a distinctly Venetian, volatile character. An experiment of interest was the production of a series of glass sculptures from sketches and models commissioned by the dealer Egidio Constantini from internationally prominent painters and other artists.

(Hu.Wa.)

CHINESE GLASS

Glass has never been truly at home in China. Records suggest that it was brought there from the West as early as the 3rd century, but finds of small glass objects of typical Chinese shapes dating from as early as the Han dynasty (206 BC–AD 220) suggest that, even if the material was brought from the West, it could be worked on the spot to conform to Chinese usage. It was no doubt regarded as a cheap substitute for jade. The Chinese themselves do not claim to have made glass before the 5th century, and even then it is doubtful if they knew more than how to make beads and other similar small objects. The vessels of glass occasionally found in burials of the T'ang (618–907) and later dynasties, although perhaps locally made, are more likely imports. Of the extant glass vessels typically Chinese in form, none can be shown to be of a date earlier than the reign of the K'ang-hsi emperor (1662–1722), and there is every likelihood that glassmaking was in fact introduced in this period when, through the Jesuits, China became vividly aware of Western culture. To this period probably belongs a series of bowls and vases of which the blown character is manifest. They are often of a deteriorated material that appears to suffer from the same defects as European glass of the same

By courtesy of Museum fur Kunst und Gewerbe, Hamburg



Snuff bottle, opaque whitish glass with red cut overlay, Chinese, 18th century. In the Museum für Kunst und Gewerbe, Hamburg. Height 10.5 cm.

epoch. During the reigns of the Yung-cheng (1723–35) and Ch'ien-lung (1736–96) emperors, the emphasis on blown forms is subordinated to the desire to make glass a surrogate for natural stones. Although the colours used are often not such as are found in nature, the glass is handled as though it were jade, the foot in particular being fashioned as though cut from stone. This lapidary treatment is further emphasized in the cased glass bottles cut on the wheel in such a way that the design stands in one or more colours on a ground of a contrasting tone.

(R.J.Ch.)

BIBLIOGRAPHY. There is ample literature on the history of glass. The following selection of titles includes basic reference works and handbooks as well as some specialized studies many of which contain bibliographical references. In addition, the *Journal of Glass Studies*, issued annually by The Corning Museum of Glass, includes extensive bibliographies.

The basic sources for medieval glassmanufacture are HERACLIUS, *Von den Farben und Künsten der Römer*, ed. by ALBERT ILG (1873); and THEOPHILUS PRESBYTER, *Schedula diversarum artium*, ed. by ALBERT ILG (1874; Eng. trans., *On Divers Arts: The Treatise of Theophilus*, 1963). GEORG AGRICOLA, *De re metallica* (1556; Eng. trans., 1912, reprinted 1950); and particularly ANTONIO NERI, *L'arte vetraria* (1612; Eng. trans. by CHRISTOPHER MERRET, *The Art of Glass . . .*, 1662), describe in detail glassmaking in the 16th and 17th centuries. See also JOHANN KUNCKEL, *Ars vetraria experimentalis*, 2 pt. (1679). Other technological studies are APSLEY PELLATT, *Curiosities of Glass Making* (1849); and ALFRED LUCAS, *Ancient Egyptian Materials and Industries*, 4th ed. rev. (1962).

EDWARD DILLON, *Glass* (1907); ROBERT SCHMIDT, *Das Glas*, 2nd ed. (1922); and W.B. HONEY, *Glass: A Handbook . . . Victoria and Albert Museum* (1946), are among the best and most comprehensive general surveys of the history of glass. *Masterpieces of Glass* (1968), a catalog of some of the holdings in the British Museum, is the most recent scholarly publication on the subject in general, accompanied by a large bibliography. CHARLES G. JANNEAU, *Modern Glass* (1931), is a review of world glass at the beginning of the 1930s. For a general study of the international development of art glass, see ADA BUCH POLAK, *Modern Glass* (1962). GEOFFREY W. BEARD, *Modern Glass* (1968), provides a brief account of modern glasswork from various countries.

Comprehensive illustrative material on glass of the ancient world is found in GUSTAVUS A. EISEN and FAHIM KOUCHAKJI, *Glass*, 2 vol. (1927); the most scholarly survey is that of THE CORNING MUSEUM, *Glass from the Ancient World: The Ray Winfield Smith Collection* (1957). Roman glass in particular was treated exhaustively by ANTON KISA in *Das Glas im Altertume*, 3 vol. (1908). Basic treatises on pre-Roman glass include H.C. BECK, "Glass Before 1500 B.C.," in *Ancient Egypt and the East*, pt. 1, pp. 7–21 (June 1934); POUL FOSSING, *Glass Vessels Before Glass-Blowing* (1940); and BIRGIT NOLTE, *Die Glasgefässe im alten Ägypten* (1968). In addition to KISA (*op. cit.*), general books on Roman glass, such as MORIN-JEAN, *La Verrerie en Gaule sous l'Empire romain* (1913); CLASINA ISINGS, *Roman Glass from Dated Finds* (1957); and DONALD B. HARDEN, *Roman Glass from Karanis Found by the University of Michigan Archaeological Expedition in Egypt 1924–29* (1936), are important for the understanding of this period. The latter has become the standard reference work for describing and cataloging ancient glass in general. Western glass of the 5th–8th centuries is treated in detail by D.B. HARDEN, "Glass Vessels in Britain and Ireland, A.D. 400–1000," in *Dark Age Britain* (1956). The standard handbooks on Islāmic and Western medieval glass are still CARL J. LAMM, *Mittelalterliche Gläser und Steinschnittarbeiten aus dem Nahen Osten*, 2 vol. (1929–30); and FRANZ RADEMACHER, *Die deutschen Gläser des Mittelalters* (1933). Byzantine glass is described in JOSEPH PHILIPPE, *Le Monde byzantin dans l'histoire de la verrerie, V^e–XVIe siècle* (1970). The basic handbooks on French and Belgian glass are JAMES BARRELET, *La Verrerie en France de l'époque gallo-romaine à nos jours* (1953); and RAYMOND CHAMBON, *L'Histoire de la verrerie en Belgique du IIe siècle à nos jours* (1955), the latter including ample bibliographic references on literary sources. WILLIAM A. THORPE, *A History of English and Irish Glass*, 2 vol. (1929), is still the standard reference work on English glass while HUGH WAKEFIELD covers *Nineteenth Century British Glass* (1961). German, Bohemian, and Austrian glass is treated exhaustively in ROBERT SCHMIDT, *Die Gläser der Sammler Mühsam*, 2 vol. (1914–27). For polychrome painting on vessels, see AXEL VON SALDERN, *German Enameled Glass* (1965). The handbooks on glass from *c.* 1800 to *c.* 1900 are GUSTAV PAZAUREK, *Gläser der Empire und Biedermeierzeit* (1923) and *Moderne Gläser* (1901). ASTONE GASPARETTO, *Il vetro di Murano dalle origini ad oggi* (1958), is the basic reference work on Venetian glass. The best surveys on Spanish glass are JOSEP GUDIOL Y RICART, *Los vidrios catalanes* (1941); and ALICE WILSON FROTHINGHAM, *Spanish Glass* (1964). For Scandinavian material, ADA BUCH POLAK, *Gammelt Norsk Glass* (1953); and HERIBERT SEITZ, *Äldre Svenska Glas . . .* (1936), should be consulted—both contain an English summary. Glass in the United States has been dealt with in great detail by GEORGE S. and HELEN MCKEARIN in *American Glass* (1948) and *Two Hundred Years of American Blown Glass*, rev. ed. (1966). LURA W. WATKINS, *American Glass and Glassmaking* (1950, reprinted 1970), presents a useful outline of 19th- and 20th-century American glass. RAY and LEE GROVER, *Art Glass Nouveau* (1967), is valuable for its colour illustrations of 19th- and 20th-century fancy glasses in American collections.

On Chinese glass, see W.C. WHITE, *Tombs of Old Lo-Yang* (1934); FRIEDRICH HIRTH, *China and the Roman Orient*, pp. 228–234 (1885); W.B. HONEY, "Early Chinese Glass," *Burlington Magazine*, 71:211–223 (1937); and H.C. BECK, "Far Eastern Glass: Some Western Origins," *Bulletin of the Museum of Far Eastern Antiquities*, 10:1–64 (1938).

Glass Products and Production

The common article of commercial manufacture known as glass is normally a transparent, hard, brittle substance formed from certain liquids that have the property of cooling below their freezing point without crystallizing, thus becoming liquids of increasingly high viscosity (stiffness) until eventually they are so stiff that by all ordinary definitions these liquids have the properties of a solid.

Of all the many substances that can form glass, the most widely used is silica (SiO_2), usually in combination with lime (calcium oxide, CaO) and soda (sodium oxide, Na_2O) in varying proportions, to make the soda-lime glass used in enormous quantities for windows, lamp bulbs, bottles, and low-cost tableware. Though glass could be made from silica alone, it would have an impractically high softening temperature, making it difficult to work. Nevertheless, it is the basic chemical nature of silica that determines both the structure of ordinary glass and the way in which it changes from a free-flowing liquid at high temperature to a virtually rigid condition as it cools.

The silica molecule

The silica molecule can be visualized as a single atom of silicon, to the four bonds of which oxygen atoms are attached. The bond angles are fixed at 108° to each other, in three dimensions, at the centre. The oxygen atoms thus occupy the four points of a tetrahedron surrounding the silicon. Each oxygen atom plays a role in two such tetrahedra, but the bonds that unite it with its silicon partners do not lie in a fixed plane or at a precise angle. There is thus an almost infinite range of possibilities in the relative dispositions of the many tetrahedra that comprise even a minute speck of silica.

The silicon–oxygen bond is very powerful; considerable heat energy must be supplied if such material is to flow freely, because the fluid condition arises from the continued breaking, remaking, and rebreaking of this bond. Though heating and cooling crystalline silica can result in resumption of the crystalline form (and in materials that are not glass formers will do so), if cooling proceeds rapidly enough, the temperature at which transformation could occur is passed too quickly for the necessary rearrangement of atoms to take place; thereafter, the material is too stiff. The continuous random changes in the disordered structure are brought to a halt, and the structure remains as it was in its last fluid moment; in effect, a liquid whose molecules are immobilized.

Such a random liquid-type structure is a good deal less compact than one in which a state of maximum order prevails. It is full of cavities and holes that are of very considerable importance in practical glassmaking. If a metallic oxide is incorporated into the glass, many of these spaces will be occupied by metal ions, each of which will employ one of the bonds of an oxygen atom, so removing one of the props of the firm three-dimensional silica network. This, and the fact that the oxygen-ion bond is nondirectional, results in a much lower vis-

cosity in the two materials at a given temperature. Ordinary glass contains about 15 percent soda to render it workable at moderate temperatures. Lime is also necessary to prevent water solubility.

This article has been divided as follows:

The general properties of glasses
- Chemical properties
- Physical properties
- Thermal properties
- Electrical properties
- Optical properties

History of glass production
- From ancient times to the 19th century
  - The blowing iron
  - The Middle Ages
  - Window glass
- Development of the modern industry

Production of conventional forms of glass
- Raw materials
- Furnaces
- Production processes
  - Flat glass
  - Bottles and containers
  - Light bulbs and tubing
  - Glass fibres
  - Toughened and laminated glass
  - Glass in electricity and electronics
  - Heat-resistant glass
  - Tableware
  - Decoration of glass
  - Vacuum flasks and jars

Production of special glass products
- Foam glass
- Glass bricks
- Sulfated glass
- Photosensitive glass
- Photochromic glass
- Opal glass
- Chemically toughened glass
- Titanized glass
- Optical fibres
- Electrically conducting glass
- Optical windows
- Pure silica glass

THE GENERAL PROPERTIES OF GLASSES

Chemical properties. Although glass is used for many purposes for which other materials are less suitable because of their chemical instability, there is nothing about the glassy state as such that ensures its stability. Water glass, for example, composed of silica and soda alone, is useful only as a temporary glaze because it dissolves readily in water.

Though pure silica is insoluble in water, it has poor resistance to attack by alkaline solutions. Thus, the familiar soda-lime glass contains the agent of its eventual destruction. After soda-lime glass has been in use for a time, moisture in the atmosphere begins to leach sodium ions from the surface; these form sodium hydroxide, an alkaline substance that, in turn, attacks the silica in the glass. Though this etching process causes no significant contamination of most substances likely to be stored in the glass and produces no readily apparent effect, the submicroscopic fissures it creates ultimately have a profound effect on the mechanical strength of the article. Carbon dioxide can also cause deterioration in glass surfaces when moisture is present.

Natural etching of glass

In spite of such deterioration, ordinary glass generally has a long, useful life. Furthermore, the number of substances that will form glass is sufficiently great that special glasses can be formulated that will resist attack by specific substances at normal or elevated temperatures. Boric and aluminum oxides, for example, can be mixed with silica to produce a glass highly resistant to attack by hot alkaline solutions; and glass containing large amounts of alumina and alkaline-earth oxides is resistant to hot sodium vapour that would destroy common soda-lime glass.

In some applications, all of the desired characteristics cannot be combined in a single glass. In such cases, ways can often be found of modifying the surface of the glass itself by some form of chemical treatment that gives the glass surface chemical properties not shared by the substance as a whole. When this is not practicable, a coating of one kind of glass may be applied to the surface of another kind, and the resulting material is, in effect, two-ply glass.

Physical properties. Although everyday experience would seem to suggest that glass is an exceptionally rigid material, its behaviour at room temperature is actually that of an elastic solid. It can be bent or otherwise deformed and, so long as its breaking strain is not exceeded, will return to its original undistorted shape when the strain is removed. Indeed, many of its present-day applications would be impossible unless glass were possessed of some degree of flexibility.

Though everyday experience also suggests that glass is brittle and of low strength, it is actually quite strong and will not fail as a result of simple compressional loading. Glass breakage, even when it results from impact, commences as tensile failure (*i.e.*, the glass is pulled apart, as described below).

When free from surface flaws, glass also possesses immense tensile strength. Generally arising as a result of chemical activity, surface flaws can be so slight as to defy detection, even with optical aid. However, their weakening effect can be measured by comparing the tensile strength of glass that has been in use awhile with the strength of glass in an undamaged state; *i.e.*, with its fire-finished surface unmarred by contact with other substances. Nearest to perfect glass are fine, newly produced glass fibres, which have been found to support loads of up to 500 tons per square inch (about 70,000 kilograms per square centimetre), a tensile strength about five times as great as that of the best practically available steel and about double the theoretical maximum for steel.

The way in which surface defects reduce tensile strength is demonstrated when a diamond cutter is employed to produce a controlled fracture in a sheet of glass. Once the sheet has been lightly scored, pressure applied on the scored side, over a suitable support, will cause separation. The relative rigidity of glass itself provides the lever that concentrates the available force on the very few interatomic bonds at the base of a tiny crack (microcrack). The ultramicroscopic structures fail under an assault by what is, relatively, truly colossal stress, which may amount to a tensile load of thousands of tons per square inch or centimetre. Effective leverage increases as the crack becomes deeper. The liquid structure of the glass presents no crystal boundaries or discontinuities that might arrest the progress of the crack, and it rapidly travels through the glass.

That glass is subject to fatigue fracture can be demonstrated by subjecting it to prolonged static loading. It will eventually fail under stresses far lower than those that a precisely similar sheet of glass can initially withstand. It is believed that such fatigue fractures are assisted by chemical assault. Once a microcrack exists with an orientation that allows it to be opened by the effect of the applied load, silicon ions in the exposed base of the crack can be dissolved out. Though deepening of the crack by such means will initially be very slow, purely mechanical forces will take over in time and the crack will spread very rapidly, eventually proceeding through the medium faster than the speed of sound.

Chemical action and cracking

The belief that chemical action does indeed assist in the initial deepening of a crack is supported by the fact that water, as has been seen, readily removes sodium ions from the glass. This property has long been exploited by glassworkers, who commonly wet the cracks made by a cutter before breaking sheets of glass.

Spontaneous fracture of badly annealed glass, in which internal strains are frozen in, can be seen from the preceding paragraphs to be inevitable when the surface has accumulated defects of the kind that are the natural consequence of normal use. Careful annealing to remove such strains is an important part of glassmaking technique. No less important are the methods by which glass surfaces are made more resistant to minor damage or protected against it.

Molten glass acts like a liquid in contrast to ductile metals, which quickly develop weak spots causing them

to narrow (neck) and eventually break when drawn by straightforward methods. Glass does not neck when drawn; tubing and rod can be stretched steadily, and almost indefinitely, as long as the viscosity of the material is carefully controlled.

Thermal properties. Glass expands when heated, shrinks when cooled, and is a poor conductor of heat. Though the surface of a slab of glass can lose heat very quickly by radiation into the atmosphere or through contact with any efficient heat-conducting medium, the interior loses heat by a slow process of conduction. In consequence, abrupt changes of temperature in rigid glass result in temperature differences between the interior and surface (thermal shock), creating stress that often leads to breakage. When glass is suddenly cooled, for example, the expanded interior puts the surface in a state of tensile stress. Careful annealing of ordinary glass eliminates the freezing in of strains that might aggravate stress resulting from mild thermal shock. Some kinds of glass expand and contract considerably as a result of temperature changes. Even well-annealed objects made from these are likely to fail as a result of thermally induced tensile stress acting on minor surface damage. Common soda-lime glass, with its high coefficient of thermal expansion, is unsuitable for any applications in which thermal shock may be encountered. Borosilicate glass (Pyrex type) is far more resistant to tensile stress; while both the so-called Vycor glass, containing 96 percent silica, and pure silica glass are virtually immune to thermal shock.

Thermal shock

The reaction of a glass to changes in temperature is very important in optics, in which use a dimensional change amounting to a single wavelength of light can very seriously affect performance. For critical applications, Pyrex and Vycor types are increasingly used.

Because no real stresses exist in molten glass, sudden extreme chilling of its surface has a contrary effect to that described above. Such chilling immediately creates a very tough outer skin that has a far higher viscosity than the interior. As the glass cools to room temperature and temperature differences diminish between interior and surface, the interior is denied the dimensional shrinkage normal to its temperature condition. In consequence, the interior is left under severe and permanent tensile stress; it exerts a powerful compressional force on the surface. Glass in this condition is very resistant to impact and more resistant than untreated glass to thermal shock.

Electrical properties. Although most glass contains metallic ions (charged atoms or molecules) capable of carrying a current, the high viscosity of glass impedes their movements and electrical activity. Thus, glass is an efficient electrical insulator, though this quality varies with viscosity, which is a function of temperature. Indeed, if glass is first heated sufficiently by normal methods, it can be treated as a conductor and melted electrically.

Though the use of glass as an insulating material in electrical equipment might seem to be ruled out by its conductivity at high temperatures, special varieties have been developed that retain desirable insulating properties at several hundred degrees Fahrenheit. Furthermore, the conductivity of small components can be greatly reduced by prolonged annealing, which ensures maximum compactness of the glass and considerably decreases movement of ions.

Optical properties. Unlike those in most solid substances, electrons in glass molecules are confined to particular energy levels, and hence they cannot absorb and re-emit light energy in the form of photons (units of light energy) by skipping from one energy level to another and back again. Light energy, in consequence, travels through glass instead of being reflected, so that glass is transparent. Furthermore, the molecular units in glass are so small in comparison to light waves of ordinary frequency that their absorption effect is negligible. Light waves in some frequencies, such as the infrared or ultraviolet, can cause glass molecules to vibrate, however, making the glass opaque to light of these frequencies. Furthermore, glass to which certain metallic oxides have been added will absorb wavelengths of certain colours, thus appearing tinted to the eye.

Because of its amorphous liquid structure, glass is transparent in a manner associated with ordinary liquids rather than with transparent crystals. Light is not polarized (constrained to vibrate in a single plane) in passing through glass, nor is it much reflected, as it is in such materials as mica, which has been strained at internal optical boundaries. This optical neutrality can, however, be destroyed locally in glass as a result of internal structural strains such as arise from inadequate annealing.

Light rays are deviated on passing from one transparent medium to another of different density unless the boundary between the two is at a right angle to the direction in which the rays are travelling. The extent to which light is bent, or refracted, on passing into a medium such as glass depends on the angle at which the light meets the surface (angle of incidence) and on the relative densities of the glass and the other medium involved, usually air. Refraction can be expressed as a constant, the index of refraction, derived mathematically from the ratio of the sines of the angles of incidence and refraction. The constant has a different value for different wavelengths and for different types of glass.

Refraction of light

On emerging from glass into air, a light ray will be similarly deviated, but in the opposite sense. That is to say, if the faces of the glass sheet are flat and parallel, the final path will be parallel to the original path of the ray, but if the faces are not parallel (as in a lens or triangular prism), the ray will emerge in an entirely new direction.

Since the index of refraction is different for different wavelengths of light, a prism can be used to separate the constituent wavelengths of "white" light. A lens can be considered as a series of infinitely small prisms the face angles of which are such as to deviate parallel rays of light toward a single point. Though no single lens can do this for all colours, a combination of lenses of different types of glass can do so for several different wavelengths, thus producing a focussed image free from any noticeable false spread of colour.

If the angle at which light rays inside glass arrive on the second surface is large enough, the deviation on emergence will give the light rays a direction parallel and close to the surface on emergence. This critical angle cannot be exceeded without the ray being reflected on the glass–air boundary. Called internal reflection, this phenomenon is widely exploited in making prismatic binoculars, single-lens reflex cameras, and range finders. It also has important new applications in light piping and fibre optics, which are described below.

HISTORY OF GLASS PRODUCTION

From ancient times to the 19th century. The oldest specimens of manufactured glass, beads from Mesopotamia and ancient Egypt, date from about 2500 BC. These beads and very much later specimens are of particular interest because their dates of origin reveal how long a period of time elapsed before glassmakers became aware of the many ways in which hot plastic glass could be manipulated.

Mesopotamian glass

Small glass vessels, nearly 4,000 years old and of presumed Mesopotamian origin, were neither blown nor molded but painstakingly sculpted from solid blocks. Only after a thousand years or more of experience did glass vessels begin to appear in Egypt; these were produced not by exploiting the ductile nature of hot glass but by an extension of the already established method of glazing pottery. Molten glass was poured over a shaped sand core in successive layers until the vessel was sufficiently sturdy. The surface was decorated with molten drops of coloured glass poured over the outer layer and pressed into it before the vessel had cooled. This method was employed for more than a thousand years until the introduction of a simple tool that revolutionized glassworking.

The blowing iron. Probably dating from about 200 BC, the blowing iron is believed to have been used first in

Babylon and adopted with enthusiasm by the Romans. Along with the blowing iron (an iron tube about 5 feet, or 1.5 metres, long with a mouthpiece at one end and a knob for holding the soft glass at the other), the craftsman employed a marver, or polished block of iron, and a pontil, a solid iron rod. A blob of soft glass on the knob end of the blowing iron was first rolled into a suitable shape on the marver. Next, it could be blown either into a constraining mold or freely in air. The pontil was employed to roll, squeeze, twirl, or pinch a similar blob of glass into shapes of almost any desired complexity and to weld them to the globe of hot glass on the blowing iron, making, for example, the stem and foot of a goblet. The still soft globe could be cut from the blowing iron with shears; the completed vessel, when cooled, was made to part company with the pontil by means of a sharp rap on the iron (cracking off).

Millefiori glass

Another development in glass manipulation adopted by the Romans led to what became known eventually as millefiori glass. Threads of coloured glass, made by stretching when hot, were laid in carefully arranged bundles so that they could be fused together to make a "cane," which could in turn be cut into slices, all with exactly the same surface pattern. The pattern units were then laid side by side on a suitable sand core and heated until they fused together to form a continuous self-supporting shell.

Both the Romans and the Egyptians showed astonishing skill in the way they used metallic oxides as colorizers. Very small differences in oxide content can drastically affect the final colour of a glass; yet colours and tints were reproduced time and again with remarkable consistency. Copper was used to make green and ruby-red glass; iron produced black, brown, and green; antimony, yellow; manganese was employed to make purple and amethyst glass. An opaque white glass, made by using tin, was important in glass cameo work, of which the famous Portland vase, made in 1st-century Rome, is an outstanding example. To make this vase, a layer of white glass was superimposed on a darker material and afterward sculpted, pierced, and cut away to leave the white figures in relief against the darker background.

Roman attempts to make flat glass by pouring slabs about one-half inch (12 millimetres) thick were unrewarding. Proper transparency cannot be achieved by such means without grinding and polishing the cast material; the lack of transparency and the difficulty encountered in making any but small panes by this method led to the introduction of stained-glass windows, first used in the Eastern Roman Empire in the early 12th century.

The Middle Ages. Glassmaking skills in Europe declined after AD 200; for about a thousand years, standards remained far below those of the Romans. The range of articles, as well as the quality of the material, was poor; the glass was of inferior colour and marred by streaks and bubbles. The stained-glass windows that began to appear in the new Gothic churches in Europe in the 12th century reached their full splendour in the 13th and 14th centuries. Many of the colours were produced by the method of fusing stains to the surface of the glass. Clear, colourless glass proved extremely difficult to achieve, however.

Glassmaking in Venice

The real revival of glassmaking skills in Europe came by way of Venice through contact with the Eastern Roman Empire (Byzantium). The Venetians made discoveries and innovations of their own, learning, for example, to eliminate all accidental colorizers from a glass melt by adding countercolorizers. The natural result was a gray glass the overall transparency of which was even less than that of the otherwise slightly tinted glass. So long as the amounts of the original colorizer and of its antidote were small and the thickness of the finished article was slight, however, the loss of transparency was less noticeable than the unwanted colour would have been.

The Venetians eventually redeveloped all the skills of the Romans. Their products and their secrets were so much sought that the glassworks, said to be a mile long, was moved in its entirety to the island of Murano in 1291 to protect its secrets and to make emigration of its workers more difficult. By the 16th century, however, most of the secrets and all of the manual skills of Venice had spread. Glass of improving quality was being produced over most of Europe in ever-growing amounts.

Although ingredients varied from place to place, most glass produced in the past, as today, was soda-lime glass. Satisfactory for most purposes because it is very stable chemically and of reasonable hardness, soda-lime glass is also easily made, and its moderate softening temperature makes it very workable, and it can be readily resoftened a number of times if necessary to complete an article. Above all, the materials needed for its manufacture were plentiful; sand and limestone exist almost all over Europe, and the soda ash was readily obtainable from the hardwood forests that also provided fuel for the furnaces. Desirable but scarcer materials included potash, produced by burning seaweed and favoured by the Venetians above soda ash, and calcined and crushed river pebbles (selected for their whiteness), which most Italian glassmakers preferred to common sand.

Glass furnace

The 15th-century German authority Georgius Agricola recommended a glass furnace nine to ten yards (eight to nine metres) wide and five times as long, large enough to keep 100 or more craftsmen busy. Agricola's specifications would have made glassworks in the mid-16th century very large compared to other industrial operations at the time. It was a standard practice to roast all the ingredients of a glass batch in order to remove as much moisture as possible before use. Mixing was carried out in a separate furnace; and after being stirred and heated for 24 hours, the ingredients were ready to be shovelled into fireclay pots for melting.

The furnace was roofed over and fitted with flues; its floor, called a siege floor, stood somewhat above ground level. The fuel chamber, which lay underneath this siege floor, had a series of holes along its length. These allowed the fuel to be ignited at many points and provided adequate access of air. The hot gases from the burning fuel entered the melting chamber through a series of vents confined to about three-quarters of the length of the siege floor. This arrangement provided the melting chamber with a relatively cool zone into which pots of molten and refined glass could be pushed to achieve working viscosity as they became needed by the workers.

Window glass. Glassblowing was indispensable to nearly all glassmaking operations until the 20th century. Not only tableware and bottles but flat glass, too, depended on the use of the blowing iron. Until well into the 19th century, the only method of making window glass was, first, to blow a large globe of glass; this was cut from the blowing iron after welding the pontil to a point diametrically opposite. The open side of the glass balloon, 20 inches (50 centimetres) or so in diameter, was then flattened on the marver; then, with the pontil being continuously rotated, the glass was reheated until it became very soft. Immediately after it was removed from the heat, the rate of spin of the pontil was increased; suddenly, the centrifugal stress on the soft glass caused it to flow rapidly outward from the centre and to take the form of a large disk, which was kept spinning until the glass was rigid.

Such glass of course was not truly flat. The disk was very uneven, being thickest near the centre and marked by concentric circular waves; at the very middle was the fractured nub, or crown, marking the point of former attachment to the pontil. Disks more than about five feet (1.5 metres) in diameter were hardly practical. The manufacturing process and the need for great economy in cutting the material, rather than aesthetic taste, account for the sector-shaped fanlights and small window panes in English Georgian houses of the 18th and early 19th century. Polished cast-plate glass, first produced in France toward the end of the 17th century, was little used, even by the wealthy, because of its prohibitive cost.

Flint glass

First made in the 17th century, using a large proportion of calcined flints with potash substituted for soda ash, flint glass was initially a failure. It spontaneously developed a maze of fine cracks that made it useless. The addition of lead oxide, however, eliminated this defect

and produced a fine lustrous glass, soft enough to be cut and engraved easily and of greater refractive power than the common crown glass. It was also more dispersive of light, but not to the same degree. Telescope makers took advantage of these properties: they placed a flint-glass lens and a crown-glass lens one behind the other to eliminate the worst of the coloured haze surrounding telescopic images of bright objects.

Development of the modern industry. During most of the 18th century, glass, apart from a few small windows, was a rarity in the homes of ordinary citizens. Containers and tableware were commonly made of pewter or earthenware. By the end of the 19th century, however, glass was a common material, pouring from the furnaces in a seemingly endless flood to be etched, scratched, engraved, and plated; rolled, ground, and polished; blown, cast, and molded into every conceivable form and for almost every conceivable purpose.

The first important mechanical manufacturing innovation was the hand-operated split mold, patented in 1821, which was made of two hollow blocks of iron. This device made it possible to blow the whole bottle, instead of the body only, in the mold. Sixty years later, semi-automatic bottle-making machines were coming into use.

The production of flat glass by breaking and spinning a blown globe gave way to the glass cylinder, blown by using compressed air, which could be slit lengthwise, reheated, and allowed to flatten on an iron table under its own weight. Although the natural fire finish was destroyed on one surface, the final product, still far from being truly flat, was flatter than crown glass. Output increased dramatically, so that the glassmaker and the engineer were able to astonish the world by midcentury with the Crystal Palace in London, made from 300,000 sheets of cylinder-blown glass resting on a lightweight iron framework.

Fundamental to the great increase in output was the introduction of the regenerative furnace, originally developed for the metallurgical industry, in which hot gases were recycled. This furnace provided far more usable heat from a given amount of fuel than had been possible before, and enormously speeded up the melting process. Of even greater importance was conversion of the old siege floor into a vast tank in which the materials were melted directly. Fed in at one end, the batch melted and flowed steadily toward the far end, at which it was taken up by the working teams, or later by machines, that converted it into finished glass products.

The greatly increased flow of glass from the new furnaces found steadily improving machines waiting to receive it, and the increasing amounts of mechanical power available began to make plate-glass polishing, though still costly, a practical business proposition that soon started to change the appearance of the main thoroughfares in large European and American cities.

Optical glass

Glass of optically reliable quality was a rarity at the beginning of the 19th century, except perhaps in France, where a means was found of producing truly homogeneous flint. Optical glass, formulated for the lens and instrument designer, however, did not exist. Its development called for the application of an altogether more scientific and less empirical approach to the problems of the effects of ingredients on the properties of glass. This progress came with the remarkable collaboration of the German scientists Carl Zeiss, Ernst Abbe, and Otto Schott in Jena, with the result that many new elements came into use in glassmaking; the effect on the quality of microscopes was immediate.

The new scientific approach to glassmaking was paralleled by a more scientific attitude toward the properties and possibilities of glass, so that new methods of using and handling glass continue to multiply, as do the applications of glass. If the beginning of the modern industry is dated from the beginning of the 20th century, then the number of important innovations and changes in glassmaking technique during its short history can be seen to outnumber very easily all those of the preceding 4,500 years. So numerous are they that a list would be tediously long and a description impossible. For example, there are at least four major changes in the methods used to produce flat glass; twin grinders that can produce about 400 square yards per hour of polished plate glass; machines that provide 2,000 or more lamp-bulb glasses per minute, and others that produce astronomical lengths of glass filament. There are also many newer glass types: heat-resistant, low-expansion glass; thermally toughened and laminated glasses; photosensitive and photochromic glass; foam glass, glass filters, glass fabrics; glass that is transparent to ultraviolet and to infrared rays; and glass that provides protection against X-rays and gamma radiation. Some of these processes and new glass types are described more fully in the later text.

PRODUCTION OF CONVENTIONAL FORMS OF GLASS

Raw materials. The greater part of glass produced today consists of variations of the silica–lime–soda mixture. Enormous quantities are used for flat glass of all kinds and for containers, including light bulbs. Optical glass, even in technologically advanced countries, accounts for only about 4 percent of the total value of the output; in terms of tonnage, it is very much less.

Modern continuous methods of production involve a high degree of organization. One of the first requirements is a continuous supply of raw materials, consistent in quality, that lend themselves to standardized cleaning treatment. Thus, the glass manufacturer first looks for conveniently situated sources of adequate size that he can own or control. An idea of the standards required of the ingredients at firing time is given by the fact that the iron oxide content of even common domestic ware is unlikely to be more than about one part in 2,000. For higher quality materials, this figure is frequently cut to one part in 10,000 and sometimes very much less.

Soda is normally provided in the form of sodium carbonate (soda ash), and lime as calcium carbonate. Magnesia is added to the batch when dolomitic lime is used. For more specialized glasses, boric acid and borax are employed to provide boron oxide; zinc is directly oxidized to provide zinc oxide; litharge (red lead) is ordinarily used to provide the lead oxide for crystal ware and for dense flint optical glass, in which it may amount to 50 percent by weight of the total batch. Many minor ingredients are used as colorizers, decolorizers, and to assist in refining the glass, or to give it special working properties or special characteristics when finished. Oxides of copper, chromium, cobalt, iron, and nickel are used as colorizers, as are colloidal gold and copper. Antimony and arsenic are useful refining agents: phosphates and fluorides are employed in opal glass. Table 1 lists the principal ingredients of important types of glass.

Use of cullet

One particularly important ingredient of glass has not been mentioned because it does not, or should not, affect the balance of the constituents of the final product. Cullet, or broken waste glass of the same type as that being manufactured, almost always forms from 50 to 75 percent of the batch. Cullet acts as a solvent. Because any glass becomes fluid at a lower temperature than any of its constituents and because glassmaking depends on the intersolution of its ingredients rather than melting, the entire process is enormously speeded up by the quick provision of an adequate amount of fluid glass.

Uniformity of sand-grain size is also important in melting. Sand grains of different sizes tend to separate as a result of the vibrations suffered during transit and handling. During the melting process, their behaviour would be unpredictable and the time allowed for homogeneous intersolution of the glass ingredients would sometimes be inadequate. Streaks of glass of differing viscosity would result. This, in turn, would cause undesirable optical effects and possibly provide a mechanically weak final product.

Furnaces. The regenerative furnaces used in large-scale glass production are invariably tank furnaces, normally designed for continuous production throughout the whole useful life of their refractory (heat-proof) linings. These linings eventually corrode to a point at which they become deeply scored, and pieces are likely to break

Table 1: Chemical Contents of Various Kinds of Glass

	composition (percentage)
Soda-lime silica	silica (SiO_2), 70; soda (Na_2O), 15; lime (CaO), 10; magnesia (MgO), 2.5; alumina (Al_2O_3), 2.5
Borosilicate	silica (SiO_2), 60–80; boric oxide (B_2O_3), 10–15; alumina (Al_2O_3), 1–4
High silica (shrunken glass)	silica (SiO_2), 96; boric oxide (B_2O_3), 3–4
Aluminosilicate	silica (SiO_2), 5–60; alumina (Al_2O_3), 20–40; lime (CaO), 5–50; boric oxide (B_2O_3), 0–10
Light barium crown	silica (SiO_2), 45–50; boric oxide (B_2O_3), 3–5; soda (Na_2O), 1; potassium (K_2O), 7; barium oxide (BaO), 20–30; zinc oxide, (ZnO), 10–15
Dense barium crown	silica (SiO_2), 30–40; boric oxide (B_2O_3), 10–15; zinc oxide (ZnO), 0–10; alumina (Al_2O_3), 0–10
Extra dense flint	silica (SiO_2), 20–40; potassium oxide (K_2O), 0–10; lead oxide (PbO), 50–80
Soft solder glass	silica (SiO_2), 5; boric oxide (B_2O_3), 15; lead oxide (PbO), 64; zinc oxide (ZnO), 16
Antiradiation	silica (SiO_2), 20; lead oxide (PbO), 80
Infrared (semi-conducting)	arsenic (As), 44; tellurium (Te), 24; iodine (I), 32
Photosensitive	silica (SiO_2), 72; soda (Na_2O), 17; lime (CaO), 11; gold (Au), 0.02; selenium (Se), 0.04
Photochromic	silica (SiO_2), 60; soda (Na_2O), 10; boric oxide (B_2O_3), 20; alumina (Al_2O_3), 10; iodine (I), 0.9; silver (Ag), 0.6; chlorine (Cl), 0.3

off. A furnace lining may last up to several years; its life depends not only on the material used in making it but also on the type of glass being made, for some glasses are much more corrosive than others.

A tank furnace may be nearly 30 feet (nine metres) wide and five times as long, with a capacity of more than 1,000 metric tons. The frit (glass ingredients), well dried and thoroughly mixed, is mechanically conveyed to overhead hoppers located at the end of the furnace, through which it is allowed to trickle down across the whole width of the furnace for introduction into the melt by mechanical pushers. The rate of feed is controlled automatically by the withdrawal of glass at the far end of the furnace, so that the level in the tank remains constant.

The furnace may be oil or gas fired; large fuel reserves are necessary in order to ensure continuity of firing. Electric furnaces, employing resistance heating, although used for special glasses produced in relatively small quantities, are too costly to operate for production of common glass. Electric boosters in regenerative furnaces, however, are used; these not only help to increase output but also provide a means of quickly and accurately controlling the heat distribution along the tank.

Heat distribution

Heat distribution is most important if a continuous supply of properly refined glass is to be delivered to the forehearth (point of extraction) at the required viscosity. During the early stages of the melting process, the newly introduced batch is converted into a stiff, doughlike mass in which the ingredients participate in the chemical reactions that will eventually yield glass. Even when odd lumps of broken glass (cullet) cannot be seen, as they can for a foot or two within the tank, the mass looks uncompromisingly gritty and full of bubbles. Steady addition of new material forces the batch further into the tank, where the temperature is maintained at about 1,500°–1,600° C (2,700°–2,900° F). At this temperature common glass is as fluid as water; bubbles formed by trapped air and gases liberated from chemical agents escape. The further end of the tank, toward which the material continuously flows, is much cooler (less than 1,000° C, or 1,800° F), so that working viscosity is assured at the delivery point.

Production processes. *Flat glass.* All previous methods of obtaining flat glass by flattening blown globes and cylinders and by drawing cylinders are now obsolete. Even the Colburn process for direct continuous production of flat sheet has been superseded. This was a horizontal process, which, though continuous, marred the fire-finished surface of the material. The principal methods used today are vertical drawing of flat sheet and roller casting, flat casting, and floating on molten metal.

Vertical drawing of flat sheet is basically an extension of the principle applied by the hand craftsman who relies on the ductility of very hot glass to draw a bar attached to his pontil into a long, thin rod of uniform cross section. Glass is delivered from the furnace by way of a throat, or bridge, which holds back any overly viscous patches and controls the flow into an adaptation of the forehearth called a drawing trough, or drawing kiln.

Drawing operations commence by dropping a long bait, made of openwork iron, into the hot molten glass passing through the drawing slot of a form, or drawing boat, in the trough. The glass quickly welds to the metal, which is then slowly raised, with glass running from it much as treacle (molasses) runs from a knife blade. Like treacle on a knife blade, too, it would quickly run in toward the middle and form a narrow stream, but for the fact that the edges of the rising sheet are cooled immediately as it rises from the drawing boat, by passing them between, but not touching, a pair of water-cooled steel boxes. The edges are then gripped by cooled, electrically-powered knurled rollers, which drive the hot, growing sheet upward to the cutting loft. The iron bait is discarded; its only role is to start the process by raising the glass until it can be gripped at the edges, after which the rising sheet will be self-sustaining so long as the gripping rollers continue to turn. The thickness of the sheet can be closely controlled by the speed at which it is drawn off from the trough.

The sheet is forced to rise through an annealing lehr (oven) in which close temperature control ensures that it is free from strains when it reaches the cutting loft about 30 feet (nine metres) above the furnace.

In the cutting loft, the glass sheet comes at last under some degree of direct human control. As it rises through the floor it is seized by large suction pads, suspended from overhead gear, and worked hydraulically by a human operator. As the top of the sheet reaches a predetermined height, it automatically triggers a mechanically operated cutter, which sweeps across the glass at an angle that allows for the continued rise of the glass during the cutting operation. A slight movement of the levers attached to the suction pads separates the standard sized sheet, which is then placed on an easel. There, the edges, marred by the gripping rollers when still soft, are cut away. These and the scraps that gradually accumulate as the result of accidental breakage are shovelled into a chute to be returned to the melt as cullet.

Rolled glass

At one time plate glass was produced by pouring it onto large tables with raised edges. Heavy rollers were then employed to squeeze the glass to an even thickness determined by the depth of the table edges. The modern equivalent of this process involves passing hot glass as a continuous ribbon between a pair of rollers so spaced as to control the thickness of the sheet. Though the glass is still hot and plastic when it issues from the forming rollers and passes over other rollers, it has attained a reasonable flatness by the time it reaches the annealing lehr. Though general surface quality suffers through contact with the metal, the final product, while lacking in transparency, is completely adequate for a number of uses. It is often given a texture, usually on one side but occasionally on both, by specially patterned rollers, so that it emerges finally as single- or double-reeded glass, so-called hammered glass, arctic glass, or Georgian glass; all these effects are intended to reduce further, rather than enhance, transparency. The process is also used to make wired glass by resoftening two sheets so that they can be united by rolling pressure over an intervening wire mesh.

The huge quantities of plate glass for the building industry are produced, like the bulk of other modern glass, by regenerative tank furnaces of the continuous type. Modern furnaces can easily deliver continuous ribbons up to eight or nine feet (2.4 or 2.7 metres) wide; these can be processed by twin grinders so that they eventually emerge, ground and polished on both faces, as high-

quality plate glass produced at up to about 1,000 feet (300 metres) per hour.

The degree of flatness and parallelism achieved by these machines is of a high standard, which can be improved on only by optical methods of grinding and polishing. Indeed, so good is it that selected small pieces of up to several square inches can often be brought up to a high optical standard without further grinding by polishing according to traditional handworking methods.

Flame-finished surfaces have already been seen to be impossible to preserve when plastic glass comes into contact with solid materials. The finish can only be restored by reheating, or replaced by grinding with abrasives and then polishing with agents such as rouge or cerium oxide.

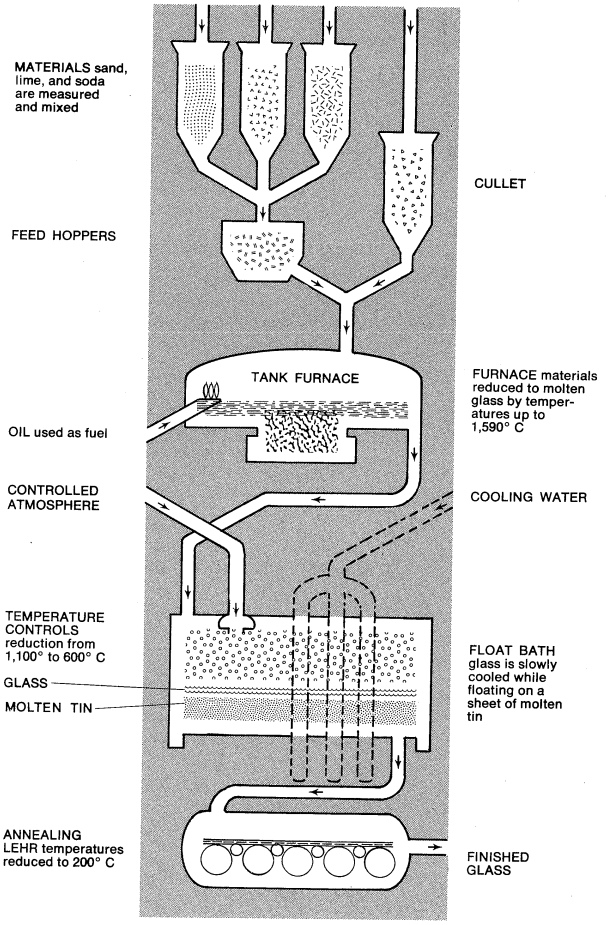
Sheet glass of admirable flatness for many common purposes, unmarred by glass-to-metal contact, is produced by the continuous vertical draw processes described above. Its virtues are most apparent, however, in thicknesses of about 1/16–1/8 inch (1.6–3.2 millimetres). For thicknesses not much greater than these, plate glass was, until the late 1960s, always necessary for purposes in which quality was important. Grinding and polishing costs are disproportionately high relative to weight cost in such material. Yet, there is a great demand for glass of good optical quality in these very gauges in which polishing costs are relatively so high. Glass of the required thickness and quality is now being produced in ever-growing quantities without expensive mechanical reworking by means of the float-glass process evolved in Great Britain.

Float glass

In the float-glass process illustrated in the Figure, the frit is melted in an oil-fired continuous tank furnace and, as in making plate glass, emerges as a continuous band. In the new process, it issues through a slot onto the surface of a pool of molten tin contained in a closed chamber, in which the atmosphere is carefully controlled to avoid oxidation of the metal, which would disastrously affect the transparency. Inside the chamber, the glass is heated from above to a temperature that renders it fluid, so that it takes up the form of a plane-parallel film of the required thickness on its bath of molten tin. Once past the high-temperature zone, the glass flows on through the chamber, still supported by the tin, until it is quite hard; then it passes into the annealing lehr. The end result is a glass that retains the fire polish of crown glass and vertically drawn sheet and that also has a degree of flatness and parallelism rivalling glass produced by the more costly grinding and polishing process. In the early 1970s, four float-glass tanks in Britain annually produced a 10,000-mile-long ribbon of the material, 11 feet wide.

Bottles and containers. The continuous regenerative tank furnaces used for the production of flat glass serve equally well with minor modifications to supply glass for the high-speed production methods of the container industry. Container-production methods consist of molding and of extensions of the ages-old glassblowing process, frequently in combination. By their nature, these are intermittent rather than continuous processes. They could, however, be given a virtual continuity if all the necessary stages in the production of a glass container could be passed through at sufficient speed. There is, however, a limit to the extent at which glass can be hurried through the complex operations necessary to make it into a bottle. The answer to the problem is the multi-station machine, which can be likened to a carousel, at every arm of which a bottle is at some stage of production. The machine revolves not continuously but in a series of stops and starts, as each arm receives its gobbet of hot glass, which has to be shaped and then removed to leave its place vacant at the end of the circuit so that a new measure of hot glass can be processed.

Glass molds are generally fabricated of fine gray cast iron worked to a polished finish. The term iron mold as used by glass makers refers to a type of mold, made in one piece, for forming shallow, easily removable items. Other molds, even though made of iron, are called press molds and paste molds.



Basic steps in the float-glass process.
Drawing by D. Meighan

Blow molding

Blow molding demands that the glass be provided in precisely measured amounts of the right consistency when introduced into the blowing mold. Furthermore, unless the container is to be very small, it must be of the right shape if correct distribution of the glass thickness is to be possible. This requirement calls for a preliminary molding operation to make the partially hollowed blank, or parison. At this stage the rim, or collar, of the bottle attains its final form in a removable part of the mold called the neck ring. The glass is introduced into the parison mold either by suction or, downward, by gravity, and it is hollowed either by a puff of compressed air, which also ensures proper overall contact with the mold interior, or by pressure of a suitably shaped die. The parison is then transferred by means of the neck ring to the blow mold in which it is finally shaped.

Press molding is little different in its essentials from the making of a parison by pressing, except that the pressing can be permitted to lose a good deal more heat during processing because it is the end product. Solid molds of low cost can be used instead of expensive opening molds, because they rarely need to be of great dimensional precision. A measured amount of glass is dropped into the mold and pressed to shape by a cooled plunger so that it fills the controlled space between the two units. Ashtrays, dishes, and ovenware are typical items so produced. Quality products are finally polished by high-temperature gas flames on the inside and around their rims while still in the mold.

As in other glass-shaping processes, molding involves careful temperature control. If the glass is too fluid when it is blown, a proper distribution thickness will be impossible; it may even be torn by sticking to the iron. If it is too cool, it will not flow as it should; the surface of the completed item will be of poor quality and any fine detail, such as a trademark or lettering, is apt to be poorly reproduced. Good container processing, espe-

cially in large sizes and complex forms, frequently requires that much ingenuity be exercised in distribution of the mold's own mass as part of the means of localized temperature control.

For practical reasons, many items have to be blown into molds despite the fact that their finish must be such as cannot be obtained when glass-to-metal contact is involved. Ordinary table glasses of tumbler shape are an example. The so-called paste mold is an iron mold lined with material, such as soap or beeswax, to which wood flour or sawdust can be made to adhere. Once the flour or sawdust has been converted into an absorbent carbon coating, it can be dampened before the glass is blown into it. The glass is kept rotating while being blown; the steam pressure generated by the heat prevents the glass coming into contact with the mold much as air pressure keeps a hovercraft from contact with the ground. Carboys and similar large containers can be blown into paste molds by skilled workers; at the other extreme are small electric bulbs, which are blown into paste molds at rates of many hundreds per minute.

Until 1821 only the body of a container could be blown into a mold; shape and general finish remained poor for many decades afterward. Even today most mold-blown articles need some finishing before they can be considered complete. Molds and neck rings must be made so they can open and release the product, and however well-fitting they may be, a seam mark is made on the glass. Nearly all modern closures require that a perfectly even and smooth finish be given to the rim so that a perfect seal can result. This finish is assured by a short application to the rim of a very hot flame, which causes the surface to flow and obliterate mold marks.

Finishing operations

Articles made on the blowing iron need more extended finishing treatment, as do paste-molded items such as table glasses. Waste material is removed by lightly scoring the surface at the appropriate position. The object is then rotated while being quickly heated by gas jets. This action cracks off the material above the scribed line. A well-directed needle of flame, sometimes preceded by a brief grinding of the edge to square it off, quickly melts the surface around the rim and rids it of the sharp corners that would soon cause it to become chipped if they were allowed to remain.

Light bulbs and tubing. Lamp bulbs can well be regarded as a specialized variety of bottle. Their extreme thinness, however, calls both for special handling methods and for special techniques to manufacture them economically and in large quantities. A machine can produce light-bulb glasses at rates of up to 33 per second. It is supplied with a continuous ribbon of hot glass, travelling between rollers that preform it with a series of regularly spaced depressions. These glass forms are blown by puffs of compressed air to their final shape and size inside paste molds, which rise around them as the ribbon passes along the conveyor.

Tubing, once produced by a combination of blowing and stretching, is almost all mechanically made today. A hollow mandrel (metal bar), through which air is forced at low pressure, takes the place of the blowing iron. Molten glass is allowed to flow over the downward-pointing mandrel at a rate that maintains an evenly thick coating during the whole process of drawing off. Air is passed through the mandrel so that as the glass is drawn off a tube is formed. A steady pull is maintained by the grip of a pair of moving asbestos-faced belts placed at such a distance from the mandrel that the glass is quite firm and nearly cold by the time it reaches them. Large-diameter tubing and cylinders, which cannot be dealt with in this fashion, are produced by a very similar process. They are drawn off vertically upward over a mandrel consisting of a refractory ring that is positioned on the surface of the molten glass, through which air is passed.

Glass fibres. Once a novelty, glass fibre is now produced in enormous quantities for its thermal and electrical insulating properties and to provide reinforcement of lightweight synthetic resins (plastics) to form boat hulls and car bodies. Filaments as fine as a few thousandths of a millimetre in diameter can be produced. The continuous-filament process is used when the fibres are intended for the making of glass yarn. Molten glass is fed through bushings (removable metal linings) that terminate in very accurately sized spinarets, or forming tips. About 400 such tips may be required to supply the very fine filaments to be united as a single strand of yarn.

The fine streams of glass issuing from the forming dies are drawn off at speeds measured in miles per minute. As they are drawn together, they are thinly coated with an adhesive, which ensures their cohesion as they pass on to the high-speed winders as a united strand. Previously, glass fibres were formed from remelted, uniformly sized glass marbles, but the simplified direct-melt method described is now commonplace and is used to produce many million miles of even quality filaments.

Glass wool

The production of glass wool demands very different treatment. Here, the problem is not to produce an orderly assembly of fibres but a mass of randomly distributed filaments. The system most generally used is the crown process. A steady stream of molten glass from the forehearth is made to fall into a rapidly spinning dish that has hundreds of minute holes in its wall. Fine streams of glass are flung out much as water is forced through the vents of the rotating drum of a spin dryer by centrifugal force. As they issue from the dish, they first meet a blast of air from an air ring that controls their direction downward toward a continuous fine spray of a binding agent. They are then subjected to a further air-jet treatment, the turbulence of which tumbles them into a state of maximum disorder as they fall onto a moving conveyor, on which they form a continuous mat.

Toughened and laminated glass. The rapidity with which cracks travel through glass when it fails under tensile stress was described above. In many applications, such as doors or automobile windshields, failures of this kind would have disastrous results; and neither wired glass nor very thick plate glass can be used. Fortunately, there are two ways—toughening and lamination—of ensuring that glass will not shatter into a mass of lethal fragments on fracture.

It has been known for centuries that a large blob of molten glass dropped into cold water will congeal into a rigid lump that can survive repeated blows with a heavy hammer. It was also known that such blobs (called Prince Rupert's drops after their 17th-century popularizer) immediately collapsed into a heap of granules when the tail of the drop was fractured by applying sufficient pressure with a vise or pliers. This condition can be simply explained.

The sudden chilling of the exterior of the drop renders it rigid before the interior has undergone the configurational shrinkage appropriate to its falling temperature. The forces acting to attain greater density place the surface under powerful compressive stress. This compression must be overcome before the surface can be subjected to any tensile stress whatever. It is easiest to do so at the tail end of the drop, at which little material and only relatively minor stress is involved; fracture there immediately releases a considerable amount of stored energy that sends shock waves through the interior, already overstressed in tension, so that it cracks in all directions, causing fragmentation of the entire drop.

Heat toughening of plate glass is carried out by heating the material to just below its softening temperature and suddenly chilling it with jets of air or cold gas. This process produces a condition that allows it to be bent farther, and to be hit about five times harder, than ordinary glass of the same thickness before failure results. An ordinary amount of bending merely goes part way to relieving the compressive stress. When it does break, the sheet disintegrates into small, relatively harmless cubelike fragments that can be seen to have concave edges, resulting from the sudden, long-delayed shrinkage of the interior. All necessary cutting and drilling operations must be carried out before the glass is tempered.

Though toughened glass is widely used in Great Britain for car windshields, it is not acceptable in the United States and some other countries. In certain situations its

very toughness can be a disadvantage; it is, for example, very much more rigid than the skulls of the car's occupants. For this reason, laminated glass, made to yield without disastrous results, is often considered to be a more suitable material for windshields.

Laminated glass

Laminated glass suffered a period of eclipse after its introduction because it was originally formed as a sandwich of two sheets of glass bonded to a celluloid interlayer. The celluloid quickly yellowed in some climates, first around the edges, but eventually over a wide zone. Today a very tough interlayer of polyvinyl butyral, which does not suffer from optical aging defects, is used. With flat glass, laminating is a very straightforward process, but modern curved windshields have necessarily complicated matters because the two layers of glass must be perfectly matched. This matching is achieved by placing paired sheets on a suitably shaped frame in an electrically-heated furnace in which they are caused to soften and take on their required contours. Next, the thoroughly dried plastic is introduced between the two glass layers in an air-conditioned room the temperature of which does not exceed 60° F (16° C) and which has a maximum relative humidity of 30 percent. Preliminary adhesion is obtained by means of mild electrical heating, and pressure supplied by rubber rollers. The assembly is then placed in an autoclave (closed vessel) in which the air pressure is raised to 50 pounds per square inch (3.5 kilograms per square centimetre) and the temperature to slightly above normal boiling point.

The vinyl layer in ordinary laminated glass, though very thin, is quite tough. Severe impacts will crack the glass but will not damage the interlayer; adhesion is such that splinters cannot fly. The process can also be extended to produce laminates that can resist almost any kind of attack; thicker layers of vinyl or glass (or more layers) can be used to make the material proof against deliberate assault by any means up to small-arms fire. Laminated glass also has the virtue of being a far more effective absorber of sound than unlaminated glass of equal thickness. A pistol shot, even at close range, will not penetrate ¾-inch- (19-millimetre-) thick laminated glass. Glass of this type is often termed bulletproof.

Glass in electricity and electronics. The uses of glass in electricity and electronics are too numerous to list; some common application problems, however, deserve mention. The lamp globes and tubes described above contain components largely constructed of glass, usually of a type very different from that of the outer shell. The high temperature frequently met in electrical use and the differing thermal expansions of the components involve the use of special glasses at junction points.

In sodium-vapour lamps, an envelope made of ordinary glass would be useless since the intensely active vapour would quickly destroy it. Chemically resistant glass containing large amounts of alkaline-earth oxides and alumina is not only costly to produce but also falls an easy victim to ordinary atmospheric attack. In consequence, sodium-vapour lamps have an envelope of cheap weather-resistant glass with a thin interior layer of the sodium-resistant material.

Glass in insulation

The great dielectric (electrically nonconducting) strength of glass makes it a valuable insulator; a great deal of glass yarn is used in the electrical industries for insulation purposes. High-voltage power lines are usually suspended from toughened glass insulators made up of as many subunits as circumstances demand. Their great mechanical strength withstands the considerable stresses imposed by the heavy cables as they swing in strong winds; failure, when it occurs, takes place not in the glass but in the high-tensile steel units that link the glass components together. Many specially formulated glasses are used in electronics for such common components as insulators, capacitors, and resistors and for many sophisticated devices employed in radar, television, and computer circuitry.

Miniature electronic components are often far too small for their glass parts to be molded by conventional means. In order to overcome this difficulty, thin glass rods are heated until they are sufficiently plastic to be shaped between precision dies. Sintering processes, which heat a substance to coherence without melting, are also used; finely powdered glass mixed with a removable binder is pressed to shape in dies, and then fired. Larger items can be fabricated by mixing the powdered glass with water and shaping it in a plaster mold; it emerges firm but still damp. After being thoroughly dried, it will retain its shape when fired at high temperature.

Heat-resistant glass. When ordinary glass fails as a result of thermal shock, it usually does so on abrupt cooling, because the surface is subjected to tensile stress by the still thermally expanded interior, which cools far more slowly than on the surface. Generally speaking, thick-walled articles fail more often from thermal shock than very thin ones. Microscope slides and cover slips, and even thin table glasses, can survive considerable thermal shock because there is in them so little interior material to apply a stress to the surface. Toughened glass, too, is fairly resistant. In many glass products, such as industrial chemically-resistant piping, high-temperature lamps, laboratory glassware, industrial thermometers, and ovenware, toughening after shaping is impracticable. These objects in consequence are fabricated from low-expansion glass, the smaller dimensional changes of which reduce thermally induced surface tensile stresses.

Borosilicate glass expands only about one-third as much as common glass for equal changes of temperature; it also has good chemical resistance and is an excellent electrical insulator. Like ordinary glass, it can be used to produce fibres and fabrics; because of their high melting point, these give excellent protection against fire, as well as being good heat insulators.

The low thermal expansion of Pyrex-type glass gives it a particular value in reflection optics because it reduces the changes of optical figure that arise in concave and convex surfaces as a result of their cross sections being different at different temperatures. The first, and most spectacular, use of low-expansion glass for this purpose was the creation of the 200-inch diameter mirror for the large reflecting telescope erected on Mt. Palomar in California. Reflecting telescopes of diameters as small as eight inches (20 centimetres) are being made today from Pyrex or other low-expansion material as a matter of course. Although its thermal expansion exceeds that of Pyrex, aluminosilicate glass may be employed in high temperature applications that also require specially high resistance to alkalies.

Shrunken glass

Shrunken glass in its finished form has less thermal expansion than any glass except pure silica. With its very high melting point, it can be used wherever extreme resistance to thermal shock is necessary. Articles are preformed from special borosilicate glass to a size larger than required; the nonsilicate ingredients are then leached out, leaving the material porous. It is then cleansed and heated until the pores vanish and the article assumes its desired dimensions. It is virtually immune to thermal shock and can be used in applications that involve working temperatures of well over 1,000° C (1,800° F).

Tableware. Among all the methods of glass production used at the present time, those employed in the making of fine tableware remain closest to traditional glassmaking techniques. Advantage is taken, however, of the improvement in furnaces and greater knowledge of glass chemistry. The range of colours from metallic oxides in the melt is greater than at any time in the past.

Offhand blowing is still an important process; tools are roughly identical with those of the past. Lead crystal is generally gathered, at temperatures around 1,000° C (1,800° F), from the middle of a fireclay ring, which ensures freedom from any scum or material of anomalous viscosity. The size of the glass gob needed is a matter of judgment, not measurement. The soft glass is then carefully shaped (marvered) on a charred wood block to ensure that it is suitably preformed before blowing, which cannot safely commence until the outer skin is viscous enough to resist expansion sufficiently so that glass flow remains controllable and proper distribution of the material can be assured. Stem and foot are added separately when the globe is firm, but still very hot.

Glassware stems

Stems can be plain or very complex. A hot gob of glass can be squeezed so that a number of evenly separated vertical grooves are formed in it. If these are then folded in they form tubes that remain when the gob is stretched; when the hot stem is twisted, they, too, are twisted, forming a helix of hollow tubes within the stem. In much the same way, the growing stem can be flattened to a square cross section and afterwards twisted. Coloured or opaque white-glass canes can be incorporated and twisted to form helices of almost any desired complexity.

Decoration of glass. Engraving processes of fine ware can be carried out on abrasive wheels if the article is substantial; water is used as a coolant and to wash away sludge. The cuts can be of various sections: flat, V-shaped, or curved (concave or convex). The cuts can be buffed to a polish afterward, on felt wheels, fed with very fine abrasive and water, or smoothed with acid.

Fine cuts on thin material and the very complex relief working on sculptured surfaces is carried out by holding the glass against the edges of rotating copper wheels fed with abrasive in an oil medium. Several dozen different sizes of wheels may be used to complete an engraving.

Diamond point stipple engraving is used occasionally; in this technique, minute specks are fractured from the surface. Considerable freedom of treatment is possible by the drypoint-etching technique, consisting of covering the article with an acid-resistant film, scratching through this coating, and then placing the glassware in hydrofluoric acid.

Masking enables broad areas to be etched by acid vapour or by air blasting with fine abrasives. In commercial glassware, considerable use is made of silk-screen printing methods to apply enamels of various colours to the glass; these are afterward rendered permanent by heat. Organic metals, similarly applied, can be used to create a thin metal film by firing the glass.

Diffusion methods

Diffusion methods can also be used, although their primary employment is in the marking of laboratory ware. A copper-silver stain is first applied by screen printing. When the article is fired, the silver and copper ions enter the glass structure to a depth determined by time and temperature control. The printing, being an integral part of the glass substance, is as permanent as the article itself.

Vacuum flasks and jars. The low thermal conductivity of glass does not prevent the contents of jars and bottles from responding as a result of heat transfer by radiation to changes in ambient temperature. Vacuum flasks, often made of low-expansion borosilicate glass, make direct conduction and radiation of heat through the walls of the container impossible. The flask consists of two spaced containers, fused together around their rims. Small pads between the components prevent flexural variations of the space between them, both during manufacture and use. After being united, the air-separated surfaces are silver plated and washed out. After being dried and heated the flasks are attached to a vacuum pump that exhausts the air from between the walls and the space is then sealed off.

The silvered surfaces are necessary to prevent losses by radiation across the vacuum space. Heat is simply reflected back into the contents and, conversely, to the air surrounding the flask. The only route of heat transfer available is by conduction along the very thin walls of the inner component. This process is so slow that fluids inserted at boiling temperature will still be hot 24 hours later; ice can remain unmelted after two days or more.

PRODUCTION OF SPECIAL GLASS PRODUCTS

Foam glass. A strange but nevertheless useful form of glass that has been called synthetic lava is made in molds from a mixture of ground glass and carbon. When heated, the mixture expands to many times its original volume and sets as a very light, rigid material. Because it is nothing but a mass of thin-walled bubbles, the foam glass can be sawed quite easily; and because the bubbles thus broken and exposed do not communicate with other bubbles, its extreme buoyancy is unaffected. Besides being useful as a buoyant medium, it is employed as insulation, an application in which its rigidity and immunity to insect and atmospheric attack make it superior to natural substances such as cork.

Glass bricks. Entire walls of steel-framed buildings have sometimes been constructed of hollow glass blocks laid with mortar bonding in the same way as ordinary household bricks. Glass bricks are made from molded halves fused together; the partial vacuum created by contraction of the residual trapped air on cooling after sealing makes them highly efficient sound insulators.

Fused spaced window-panes

A similar process is used to unite spaced panes in standard sizes as ready-made double glazing units. The complete and permanent nature of the seal ensures that moisture cannot condense inside the enclosure, makes satisfactory fitting independent of local weather conditions, and greatly simplifies installation problems.

Sulfated glass. The destructive effect of alkaline attack on ordinary glass was described above. Although resistant glasses can be fabricated, their cost or working characteristics can make them impractical in large-scale bottle making. Sulfating is not a glassmaking operation but a form of treatment that consists of changing the constitution of the glass at the surface. The bottles are filled with acid gases, sulfur dioxide and sulfur trioxide. By a process of replacement by diffusion, most of the sodium ions from the surface are removed, with a consequent raising of chemical resistance by factors of up to 100.

Photosensitive glass. By 1943 it had been well established that glass can be rendered photosensitive to ultraviolet light. It was found that the colour of certain types of clear ruby glass could be deepened by exposure to ultraviolet radiation before being subjected to the reheating process that made them acquire colour. The next step was the making of glass in which no colouring would occur during the reheating process without the preliminary ultraviolet irradiation; any desired pattern could be introduced deep inside the glass itself by interposing a suitably designed ultraviolet–opaque mask between the glass and the irradiating source.

Modern photosensitive glasses contain optical sensitizers, such as univalent cerium ions. Radiation readily removes an electron from these ions; when the glass is heated sufficiently, these can be captured by any gold ions that have been included in the glass, so that gold atoms form and remain in coalescence when the glass is cooled. Unlike the silver halides used in ordinary photographic processes, an image formed by these means is completely grainless because the opaque units are well beyond the threshold of resolution in visible wavelengths.

Photochromic glass. Unlike photosensitive glass, photochromic glass does not depend on heat treatment but on the action of ultraviolet or visible radiation in order to darken it; and unlike ordinary photosensitive materials the process of darkening is completely reversible. The action is due to the inclusion of silver chlorides, bromides, iodides, or their mixture in the glass material; these, in processing, form microscopic crystals thinly dispersed throughout the material. The darkening effect is similar to the action that occurs in ordinary photographic processes but is reversible principally, it is thought, because of the very different sizes of the crystals in the two different materials. A typical single crystal in a photographic emulsion is, by volume, many million times as great as the crystal units in photochromic glass and encourages the neutral silver alone to form stable particles. Additionally, in the impermeable glass material, the neutral chlorine atoms do not diffuse from the zones of reaction as in photo emulsions; the silver and chlorine are not removed by other reactions.

Opal glass. An ultraviolet-sensitive glass containing lithium silicate, which can be converted by heat into an opal glass, has special applications because it then becomes far more readily soluble in hydrofluoric acid. The use of a suitable mask makes it possible to irradiate the glass so that an opal print of almost any desired degree of complexity can be reproduced by heat treatment. This print can then be etched right through the glass with acid. The high resolution obtained makes it possible by this

means to "drill" accurately dimensioned, and accurately positioned, holes with distribution densities as high as 320,000 per square inch (50,000 per square centimetre).

Chemically toughened glass. The thermal-toughening process cannot be applied to very thin sections; where these need much more than ordinary flexural strength, chemically toughened glass is used. One method employed involves an ion-exchange process that produces a change in the chemical composition of the surface layer. If a glass that is rich in lithium ions is immersed in molten sodium chloride, many of the lithium ions escape and are replaced on a one-to-one basis by far larger sodium ions. This substitution results in considerable compressive stress being applied to the surface when the glass cools. Flexural strength of up to 98,000 pounds per square inch (6,900 kilograms per square centimetre) can be achieved by this treatment, much more than is possible by thermal toughening. Chemically toughened glass can actually be used to make springs.

Titanized glass. In common with many other forms of glass, bottles do not lend themselves to thermal toughening; maximum strength must be obtained by designing their shapes to withstand the impacts that inevitably result from mechanical handling. They can also be protected to some extent by suitable chemical treatment. The titanizing process, developed in Great Britain, consists of spraying bottles with a titanium compound before annealing, while the bottles are still red hot. Part of the compound volatilizes, leaving a very thin surface layer of titanium oxide, which is soluble in glass. The resulting increase in the working life of returnable bottles, which have to withstand being transported and passing through cleaning and filling plants, is due to the reduction of friction and resistance to the formation of tiny cracks in the treated glass.

Optical fibres. The phenomenon of total internal reflection described at the beginning of this article allows light to be piped around bends through the interior of a glass rod. As long as the ratio of the rod diameter and radius of curvature of the bend is not such as to allow internally reflected rays to arrive at another point of the interior surface at less than the critical angle, only absorption losses will occur.

Glass fibres are essentially very thin rods. It follows that a quarter-inch- (six-millimetre-) diameter bundle of fibres can convey light around a much sharper curve than a single rod of the same diameter. If the individual fibres are sufficiently fine and coherently arranged, the opposite ends of the bundle can be ground and polished. If one end is placed in contact with, or very close to, an object, the image of that object will appear at the distant face. Flexiscopes made from bundles of optical fibres for the investigation of otherwise inaccessible sites are used in medicine and industry. Very small bundles are called optical catheters.

The diameter of individual fibres must be extremely small for good resolution; very fine fibres are hard to handle. The task is simplified by first drawing together a bundle of coarser fibres, embedded for protection in a sheath of matrix glass. This group is then drawn together, like millefiori glass (described above) and can be stretched until it is of sufficiently small diameter, normally a few microns. The multiple fibres can be wound on a drum and afterward used to make up coherent bundles of fibrescopes.

Bundling optical fibres

Short bundles of fibres can be carefully stacked and fused together as a mass. The resulting slab or sheet has very different properties from those of ordinary glass. When its surfaces are ground and polished, light falling on a point at one surface will be transmitted without deviation to a corresponding point on the distant surface. Used as a faceplate on a cathode-ray tube it acts as an imaging screen that, because it directs the light as a bundle of parallel units, overcomes the difficulties normally arising because of light scatter when photographing the display.

Electrically conducting glass. Aircraft windshields, especially supersonic windshields, must withstand environmental conditions very different from those in automobiles. Air friction can raise the surface temperature of the glass to about 310° F (150° C) at low-level high-speed flight; and the ambient temperature may be as low as −200° F (−130° C) at some altitudes.

Laminated toughened glass with interlayers of vinyl up to one-quarter inch (six millimetres) thick is used; in military aircraft, four toughened glass sheets and three interlayers of plastic bonded together give a large degree of protection against bullets and flak splinters. A plating of gold, thin enough to be transparent, makes it possible to pass an electric current through the assembly to warm the glass and prevent ice formation. Very thin wires, less than $\frac{1}{1000}$ inch (0.03 millimetre) thick, can also be embedded for this purpose.

Optical windows. When photographing objects in supersonic wind tunnels, vacuum chambers, or any site into which the investigator or his instruments cannot easily be introduced, the camera is aimed through a special window fitted with glass the surfaces of which are exactly parallel, thus permitting undistorted photography. Even large optical windows are unlikely to deviate from true flatness by more than one-half wavelength of sodium light. For use in applications for which protection against X- and gamma radiation is required, very thick windows of extremely dense flint glass containing up to 80 percent of lead oxide are employed.

Pure silica glass. Ordinary glass is useless for ultraviolet transmission; large aluminized optical mirrors are necessary for ultraviolet spectroscopy in astronomy. Silica itself, however, is transparent to ultraviolet, although its great viscosity even at very high temperatures makes it unsuitable for molding, by ordinary means, of optical lens blanks.

The contemporary method of making 99.5 percent silica glass (fused quartz) satisfies most applications. Even

Table 2: International Glass Trade in Selected Countries
(1969*) Value in millions of dollars (U.S.)

	glass fibre		laboratory glass		lamp envelopes		safety glass		optical glass	
	export	import	export	import	export	import	export	import	export	import
United States	20.07	0.27	15.63	1.73	19.18	0.16	25.00	4.94	4.76	2.01
Canada	—	3.28	—	7.98	—	7.85	—	20.98	—	0.96
Belgium and Luxembourg	6.32	4.85	0.54	1.19	9.41	0.87	15.10	3.78	0.03	0.23
France	6.77	4.30	3.03	2.25	10.32	4.74	3.14	9.06	3.03	0.74
West Germany	5.75	11.10	13.20	0.69	7.34	4.79	12.11	2.94	5.45	1.39
Italy	3.95	2.80	0.45	1.06	0.23	8.55	2.74	4.78	0.05	0.95
The Netherlands	11.03	2.31	0.46	2.21	4.14	6.75	1.83	1.84	0.07	0.74
United Kingdom	8.79	2.03	—	0.60	5.89	1.26	3.92	1.28	2.83	0.59
Denmark	—	1.71	—	0.93	—	0.28	—	0.95	—	0.09
Norway	—	2.12	—	0.40	—	0.29	—	0.36	—	0.02
Sweden	2.03†	3.38†	0.69	1.49	—	0.47	—	3.96	—	0.08
Austria	—	1.16	0.40	0.38	0.01	0.33	—	0.74	0.02	0.68
Switzerland	—	2.22	1.29	1.85	—	0.58	0.75	0.43	—	0.71
Japan	1.51	1.83	0.90	0.28	0.70	4.25	5.03	0.23	0.73	1.31

*1969 figures except where otherwise noted. †1968 figure.

higher purity, however, can be obtained by first making disks of silicon tetra-chloride. Electrochemical methods (hydrolysis) are then employed at high temperatures, to cause the hydrochloric acid to be driven off in the form of a vapour and to leave a disk of virtually elemental glass. This is transparent to ultraviolet and near infrared light as well as to the entire visible spectrum.

Purity is such that the glass composition must be described as 100 percent silica because the impurities amount to no more than one or two parts per 100,000,000.

PRODUCTION STATISTICS

Table 2 shows the international trade in five glass categories for selected countries in 1969.

BIBLIOGRAPHY. T.K. DERRY and T.I. WILLIAMS, *A Short History of Technology from the Earliest Times to A.D. 1900* (1960), provides a useful summary of the development of glass technology. C.J. PHILLIPS, *Glass: The Miracle Maker* (1941), is intended to be of interest and use to such users of glass as architects and engineers. J. HOME DICKSON (ed.), *Glass: A Handbook for Students and Technicians* (1951); and F.J.T. MALONEY, *Glass in the Modern World* (1967), are both useful introductions to the subject for non-specialist readers with some scientific or technical knowledge, and the latter is probably the most up-to-date work of this type currently available in English. F. GIFFIN, *Glass* (1963), is suitable for non-scientific readers. Among more highly technical studies, G.O. JONES, *Glass* (1956) and G.W. MOREY, *The Properties of Glass* (1956), both contain a great deal of useful data; L. HOLLAND, *Properties of Glass Surfaces* (1964), is an exhaustive theoretical study of a narrower and more specialized field. B.E. MOODY, *Packaging in Glass* (1963), deals with the production of bottles and jars of most kinds as well as with their design and with the associated problems of filling and mechanical handling.

(Fr.J.M.)

Glassy State

Glass is a state of matter best defined by the process that produces it: when a substance in the liquid state is cooled to a rigid condition without any crystallization having taken place, the substance is said to be in the glassy state. As a liquid cools, the vibration of its atoms due to heat decreases and their spatial arrangement becomes less open. A steady contraction in the volume of the liquid results. At the freezing point, a liquid that crystallizes releases heat (called heat of fusion; the energy required to keep atoms in a liquid state is greater than the energy required to keep the same atoms in a crystalline state) with a sharp reduction in total volume as the atoms are rearranged into the crystal state. On further cooling, the crystal contracts much less rapidly because no more changes take place in the configuration of the atoms, even though the atomic, thermal, or heat, vibrations still continue. If the liquid can be cooled below its freezing point without crystallization, the supercooled liquid will be in internal thermodynamic equilibrium, in contrast to the condition of a crystallizing liquid at the freezing point. This means that the available free energy of the supercooled liquid is less than the available free energy in any other structure into which the atoms could rearrange themselves and, therefore, the supercooled structure will not by itself change. It needs energy to do so and it can remain in a precariously stable, or what is called a metastable, state until the energy is supplied. If complete nuclei form, or if crystals are dropped into the supercooled liquid, freezing will begin at those points with consequent release of heat of fusion. The system will no longer be in internal equilibrium and crystallization will spread through it.

GENERAL PROPERTIES OF GLASS

Volume and temperature relationship. Figure 1 shows typical volume changes plotted against temperature changes, and the resulting graphs give a general description of glassy state. A liquid at A is cooled to the freezing point at B. The volume shrinks as indicated from A to B and then contracts drastically from B to C without change in temperature (heat must be constantly removed from the system), as the liquid turns into a crystal. Further cooling results in moderate contraction as shown by CD. But if the liquid does not freeze at B, shrinkage continues smoothly through B on into the supercooled state to E. On further rapid removal of heat, the rate of contraction begins to change and from F to G it approximates that of the crystal. In the range E to F the atomic adjustments necessary for configurational contraction occur slowly enough to be measurable, and a glass transition temperature, T_f, is defined for a particular experiment by extending LF to cut the extended line BE.

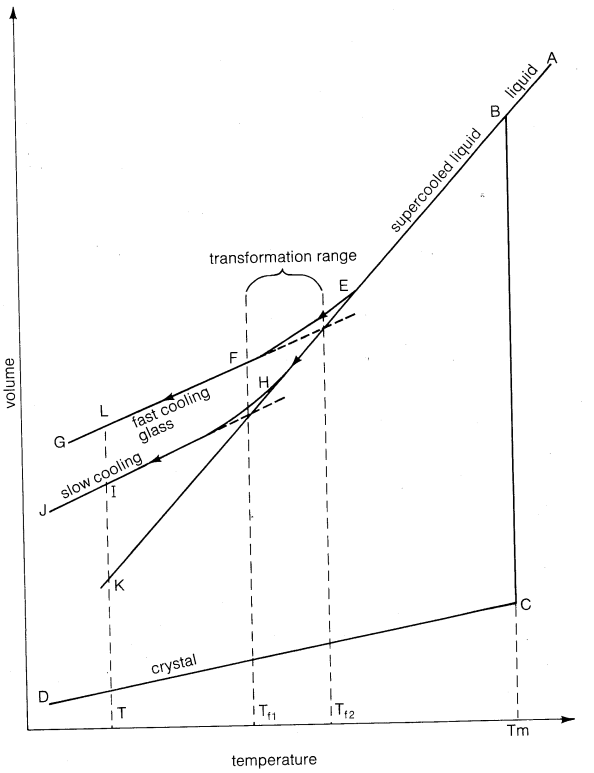


Figure 1: Volume–temperature plots for glasses, liquids, and crystals (see text).

It is more correct, however, to speak of a transformation range of temperature between the values T_{f2} and T_{f1}, which are defined by the attainable cooling rate limits. Below the transformation range the material is a glass and is in a thermodynamically unstable state. If it is cooled quickly enough to be following the line FG, and if the cooling is then stopped and the temperature held constant at T, say, the glass will slowly contract from L to K. The glass is then said to have been stabilized. The slower the rate of cooling, the further the line EK is followed, and BEHIJ is the curve obtained for a glass-forming liquid subject to a slower cooling rate than that experienced by the liquid represented by BEFG. Ordinarily, eventual departure from the equilibrium line will always occur because of the rapidly increasing times necessary for atomic rearrangement. Thus a particular glass composition has a "configurational temperature," or "fictive temperature" (T_{f1}, T_{f2}), which corresponds to the structural configuration frozen in at that temperature and persisting at lower temperatures.

Stabilized glass condition

Viscosity and temperature relationship. The variation of viscosity with temperatures indicates the ease of atomic rearrangements within the glass and is an essential property of glass that is utilized in industrial processes.

Molten glass may have a viscosity of about 10^3 poises (a poise is the unit of coefficient of viscosity and expresses the force required to maintain a unit difference in velocity between two parallel planes separated by one centimetre of fluid; one poise = one dyne-second per square centimetre) somewhat stiffer than honey on a cold day, and if cooled it gets much stiffer. At the temperatures at which molten glass can be shaped into useful articles its vis-

cosity lies between 10^4 and 10^8 poises, well below the 10^{13} poises characteristic of glass in the transformation range. At this viscosity, the time required for the stress of distortion to be relaxed at constant strain is about one minute and the time scale for atomic rearrangements is similar. In the transformation range the physical properties are time-dependent, and below the transformation range the configuration is fixed, essentially, and the glass has the general properties of an elastic solid.

The effect of composition on glass properties. The wide range of the properties of glasses depends on their composition, and special effects result from the presence of various modifying agents in certain basic glass-forming materials (see below, *The structure of glass*).

Silica glass formers

One of the most important glass-formers is silica (SiO_2). Pure crystalline silica melts at 1,710° C. In pure form, silica glass exhibits such properties as low thermal expansion, high softening temperatures, and excellent chemical and electrical resistance. In pure form, it is relatively transparent over a wide range of wavelengths to visible and ultraviolet light, and to ultrasonic waves.

The high viscosity and melting temperature of silica glass are affected by the presence or absence of other materials. For example, if certain materials called fluxes are added, the most important being soda (Na_2O), both the viscosity and melting temperature can be reduced. If too much Na_2O is added, the resulting glass is readily attacked by water, but if stabilizing oxides, such as lime (CaO) and magnesia (MgO), are present in suitable amounts, the glass becomes more durable. Most commercial glass has a soda–lime–silica composition and is produced in vast quantities for plate and sheet glass, for containers, and for lamp bulbs.

In soda–lime–silica glasses, if lime is replaced by lead oxide (PbO), and potash (K_2O) is used as a partial replacement for soda, lead–alkali–silicate glasses result that have lower softening points than lime glasses. The refractive indices, dispersive powers, and electrical resistance of these glasses are generally much greater than those of soda–lime–silica glasses.

Boric oxide (B_2O_3), itself a glass former, acts as a flux when present in silica and forms borosilicate glass and the substitution of small percentages of alkali and alumina increase the chemical stability. It also exhibits low thermal expansion, high dielectric strength, and high softening temperature.

Aluminosilicate glasses find applications similar to those of borosilicates, but the former can stand higher operating temperatures; glasses with relatively high alumina contents and no boric oxide are exceptionally resistant to alkalies.

Nonsilica glass formers

The above glasses all have silica as the glass former. With other glass formers, glasses have special properties. For example, if boric oxide (the basis of Lindemann glass) is present, X rays are transmitted and rare-earth glasses will exhibit low dispersion and a high refractive index. Phosphate glasses (used as optical glasses) based on phosphorus pentoxide (P_2O_5), are highly resistant to hydrofluoric acid, and act as efficient heat absorbers when iron oxide is added. The Table gives the compositions and physical properties of some typical commercial glasses of the types described.

THE STRUCTURE OF GLASS

Absence of long-range order. A glass is essentially a substance that has a frozen-in, liquid type of structure characteristic of that substance at a higher temperature than the actual temperature of its glassy state. Figure 2 shows, in two dimensions, the ways in which the atoms of an imaginary two-dimensional oxide (G_2O_3) would be arranged as a crystal (A) and as a glass (B). This imaginary oxide is the two-dimensional analogue of the three-dimensional tetrahedral silicon dioxide (SiO_2) in which each silicon atom is bonded to four oxygen atoms. In the crystal the structural unit is repeated regularly throughout the material and the crystal, therefore, shows long-range order. The structure of glass shows no such order. Although in both cases the number of bonds per atom is the same, in the amorphous structure of glass the bond angles are slightly distorted, and there is no long-range regularity.

X-ray evidence of structure. The broad X-ray-diffraction patterns for a glass are similar to those obtained for a liquid and they are in contrast to the crystal patterns that show sharp lines. The results are expressed in the form of radial distribution curves, two of which are shown in Figure 3. These graphs are plots of radial density; that is, they give the probability of finding a second atom at a given distance from the reference atom. For silica glass, a peak occurs at about 1.6 angstroms, which coincides with the average silicon-oxygen (Si–O) distance in crystalline silica and silicates. The area under this peak

Characteristics of Various Types of Glass

	fused silica	soda–lime–silica	borosilicate	alumino-silicate	lead
Approximate composition*	SiO_2 99.9% H_2O 0.1%	SiO_2 73% Al_2O_3 1% Na_2O 17% MgO 4% CaO 5%	SiO_2 81% Al_2O_3 2% B_2O_3 13% Na_2O 4%	SiO_2 62% Al_2O_3 17% B_2O_3 5% Na_2O 1% MgO 7% CaO 8%	SiO_2 56% Al_2O_3 2% Na_2O 4% K_2O 9% PbO 29%
Coefficient of thermal expansion (linear expansion per °C × 10^7)	5.5	93	33	42	89
Strain point† (°C) (viscosity about $10^{14.5}$ poise)	990	470	515	670	395
Annealing point‡ (°C) (viscosity about 10^{13} poise)	1,050	510	565	715	435
Softening point§‖ (°C) (viscosity about $10^{7.65}$ poise)	1,580	695	820	915	630
Young's modulus (lbs per sq in. × 10^6)	10.5	10	9.1	12.7	8.6
Refractive index (for sodium D line)	1.459	1.512	1.474	1.530	1.560
Dielectric constant (at 10^6 cycles per second and 20° C)	3.8	7.2	4.6	7.2	6.7
Density (g/cm³)	2.20	2.47	2.23	2.52	3.05

*These compositions are typical of the various glass types. †Temperature at which internal stresses are reduced significantly over a few hours. ‡Temperature at which internal stresses are reduced significantly over a few minutes. §Temperature at which glass will rapidly deform under its own weight. ‖The strain point and softening point roughly define the annealing range.

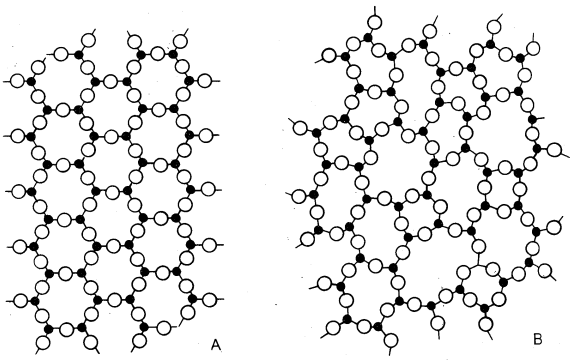


Figure 2: Structure of an imaginary oxide, G_2O_3, (A) as a crystal and (B) as a glass.

in the radial distribution curve shows that the number of oxygen atoms about each silicon atom is 4.3, which, within experimental error, is 4, and 4 is the coordination number of silicon by oxygen in crystalline silica and silicates. Other peaks and humps in the curve are thought to correspond to various characteristic interatomic distances in the silica. In the sodium–silica glass, for example, a new peak appears that probably represents the sodium–oxygen (Na–O) distance, giving an average of six oxygens near each sodium atom. Figure 3 also demonstrates an important property of glasses, namely, that they possess short-range order but no long-range order. The sharpness of the peaks is a measure of the amount of disorder present, and the peaks broaden to give a fairly smooth curve after a few interatomic distances confirming the lack of long-range order.

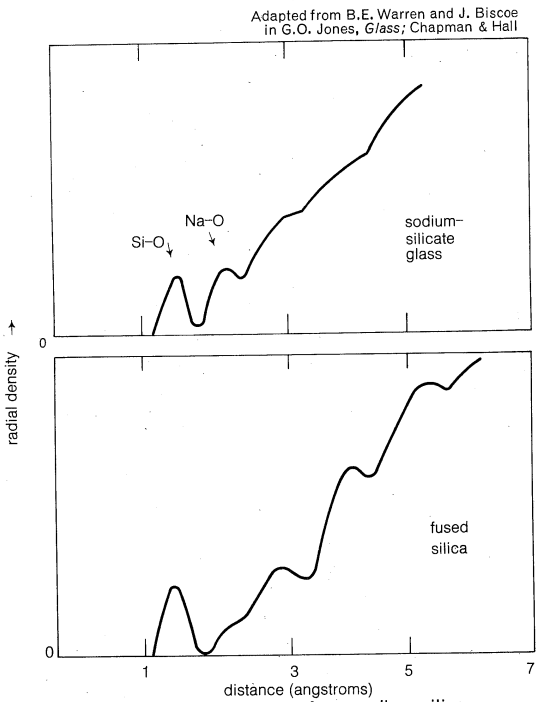


Figure 3: Radial distribution curves for a sodium–silica glass and for pure fused silica (see text).

Functions of network modifiers and glass formers. Network modifiers, in the form of large metallic cations (positively charged atoms), weaken the structure. Although it occupies a space in the open structure, the modifier removes one of the strong directional oxygen–silicon bonds from the basic network and replaces it with a weaker nondirectional bond, thereby lowering the viscosity of the glass. Alkali and alkaline-earth ions act as network modifiers.

Examples of elements other than silicon whose oxides, sulfides, selenides, and tellurides form strong directional bonds and act as glass formers are boron, phosphorus, germanium, vanadium, and arsenic. Network modifiers, with their low bond strengths and property of lowering the viscosity, include sodium oxide and calcium oxide. There is also a class of substances whose oxides are known as intermediates, including titanium, aluminum, thorium, and beryllium. Alone, intermediates will not form glasses but do so in combination with glass formers. They can enter the network if a suitable modifier oxide is available; if not, they act, themselves, as modifiers.

Spectroscopy of glasses. X-ray-diffraction studies of glasses give fairly reliable information about the number of oxygen atoms surrounding the small, highly charged, glass-forming cations, but far less about the positions of the modifying alkali ions. Spectroscopic techniques yield further useful information about the glass structure.

In spectroscopy, differences between allowed energies or quantum states of the system are measured, and these results are compared with theoretically calculated quantum states of a particular arrangement of atoms.

In the infrared region of the spectrum, the allowed energies, or quantum states, are associated with the vibrational motion of atoms but little theoretical progress toward understanding can be made because, in solids and liquids, the vibrations of all the atoms together must be considered. Spectral absorption peaks can, however, often be matched with observed peaks for crystals of known structure. Correspondence is taken to indicate the presence of particular groups of atoms, or "complexes."

In the ultraviolet region, the quantum states are those associated with the movement of an electron from a negative oxygen ion to a positive ion, or cation, but a large contribution is made by energy of vibrational modes (called phonons), which interact with the electronic energies to give intense broad absorption bands.

In the visible region, absorption bands arise from transition metal and rare-earth ions in solution in the glass. In this case the quantum states are only slightly modified by the nearest neighbours.

Deductions from ligand field theory

These nearest neighbours, which surround the central glass-forming ion, are known as ligands. Their modifications of the quantum states, and the symmetry of their arrangement around the central ion can be deduced from ligand field theory. This theory enables the relative energies of the absorption peaks to be derived and also allows the absorption intensities to be classified in groups according to the symmetry of arrangement of the nearest neighbours. It has been of great use in recent years in the study of coloured glasses.

Solution and colloidal colours in glass. Normally, in the visible region, only ions that give rise to colour need be considered when studying the transmission properties of glass because the characteristic absorptions due to vibration of the silicon–oxygen bond lie outside this region. The most common colouring ions are those of the transition metals cobalt (Co), chromium (Cr), copper (Cu), manganese (Mn), and iron (Fe). The colours are affected by three major factors, namely, the nature and concentration of the colouring ion, the composition of the base glass, and the oxidation state of the colouring ion.

Small quantities of the colouring ion, about one-tenth of a percent, are sufficient to colour glass. For this reason, the yellow-green colour caused by iron is a serious problem in the glass industry. This is alleviated by the choice of pure sands and by the addition of decolorizers. Ferrous iron absorbs energy in the near infrared, and its broad absorption band spreads into the visible spectrum causing a blue colour. The decolorizer contains an oxidizing agent, which converts iron cations with 2 charges (Fe^{2+}) to cations with 3 charges (Fe^{3+}), and also a little selenium and cobalt. Addition of the latter elements, together with the ferric iron, Fe^{3+}, give a residual absorption that is essentially neutral. The manganese dioxide, MnO_2, which was formerly used, oxidized the iron and gave a complementary purple colour, but was less stable.

Colour in glass can also be caused by colloidal precipitation of very small metallic crystals, such as those of copper or gold, to produce a "ruby" colour. The particles develop when a glass of suitable composition, previously

melted under certain reducing conditions, is heat treated. The particle size is critical, between 10 and 1,000 angstroms, and selective absorption then gives the glass its red colour.

Nucleation and crystallization in glass-forming liquids. It was noted above that a supercooled glass-forming liquid is in a metastable state (a stability that is maintained only by a slight margin over an unstable state) and that if crystals or crystal nuclei are introduced and freezing can begin, the crystallization will spread throughout the liquid.

Two aspects of the crystallization process must be considered separately, namely, the rates of formation of nuclei and their growth. So many nuclei are present in normal freezing processes that the crystallization process depends essentially upon the growth rate, and as this is negative above the freezing point, and very large and positive just below it, the liquid freezes at, or very near, the expected freezing point. For liquids (typically the highly viscous glass formers) that can be greatly supercooled, however, the crystallization rate also depends critically on the rate of formation of nuclei. Nuclei can be introduced from outside (heterogenous nucleation) or can appear spontaneously within the supercooled liquid (homogeneous nucleation). Homogeneous nucleation occurs because of thermal fluctuations within the liquid; consideration of the contribution of the surface energy of the embryonic nuclei to their total free energy shows that, at any given temperature, there is a critical size below which the nuclei are unstable and redissolve, and above which they can grow.

Homogeneous nucleation

When nucleation and crystal growth occurs in glasses, it is known as devitrification. It is most marked in glasses in which the rates of nucleation and crystal growth nearly coincide when plotted as functions of temperature. Closeness of these curves results in nucleation and crystal growth over a wide temperature range. More stable glasses have a high viscosity at the freezing point, and crystal growth, which proceeds by diffusion process at the surface of the crystal, is slowed down. Heterogeneous nucleation, however, can still occur, especially at glass surfaces, when impurity particles are present.

Studies of nucleation and crystallization in glasses have been made in recent years. One result of these efforts was the discovery of a glass containing a small amount of dissolved metal, such as gold or silver, which, after exposure to ultraviolet radiation and subsequent heat treatment, develops a colour caused by the growth of small metal particles in the glass. If this photosensitive glass is heat treated further, the particles act as nuclei, and opalescent regions of the glass devitrification products are formed wherever the glass was exposed to the ultraviolet radiation. Much less resistant to chemical attack than the parent glass, these regions are dissolved away while the parent glass remains. By the commercial use of this process, pieces and patterns of desired shapes can be obtained—for instance, fluid amplification elements.

Glass ceramics

A glass has been produced that can be changed almost completely into crystalline material—the glass ceramic. In application, an article of suitable glass composition fabricated by the usual method is subjected to appropriate heat treatment to convert it into a ceramic. Such materials are extremely inert chemically and can be designed to have either zero or a very small expansion coefficient.

SPECIFIC PROPERTIES OF GLASS

Mechanical properties. Glass rods with untouched surfaces have a tensile strength of 500,000 pounds per square inch. By considering the energy changes involved in the surfaces formed by fracture the theoretical strength of glass is estimated to be about 1,000,000 pounds per square inch. Glass articles normally show strengths that are much lower, about 30,000 lb/in^2; and this difference is due to submicroscopic surface flaws that, produced by the slightest contact with another body, act as stress magnifiers. Glass in ordinary use, therefore, is far less strong than it would be in a flawless state. Although the flaws have the same stress-reducing effect as dislocations in crystals, the exact identity of the flaws remains uncertain. Their presence, however, and number can be revealed by treating the glass with some modifying agent; for example, by immersing glass in molten lithium nitrate, the sodium in the surface is exchanged for lithium. The foreign layer that is produced is in tension when the glass cools and minute cracks develop at the flaws.

At temperatures below the transformation range, glass behaves as a solid and stresses will be present if the temperature varies within a glass object; a piece of hot glass cooled suddenly will contract more on the outside than on the inside; the further contraction of the outer layers is thus prevented and they are put into tension; the higher the thermal expansion coefficient of a glass is (the more it expands for each degree rise), the greater this tension will be. Some glasses expand very little. Thus, fused silica has excellent resistance to heat shock, while borosilicate glass can undergo repeated reheating below the transformation range without cracking. It makes excellent ovenware.

In the transformation range the glass behaves instantaneously as a solid, an immediate elastic strain being produced when an external force is applied, but this strain increases if the force is not removed. If the force is continually reduced to keep the strain constant, the internal stress gradually disappears. The rate of this change depends on the viscosity of the glass. When the coefficient of viscosity, η, equals 10^{13} poise, the stress is halved in about one minute. If a temperature gradient (a variation of temperature) exists in a piece of glass in the transformation range, the resulting stresses die away and the glass then has a temperature gradient but is stress free. On cooling to room temperature the hotter interior has to contract more than the outer layers and is, therefore, in tension while the outer layers are in compression. These stresses remain until the glass is reheated to the transformation or annealing range. The principle is applied in annealing, which is a process of minimizing the permanent stresses in glass by heating until they disappear and then cooling slowly so that only very small temperature gradients, and ultimately, therefore, only very small stresses, appear. Unless a glass article is properly annealed it might break spontaneously after cooling. Large volumes of stress-free uncracked glass are rare and difficult to achieve. Annealing for exacting uses, such as in some optical applications, moreover, involves more than the removal of thermal stresses; the atomic structure of the glass will vary within the piece unless all parts of it have had exactly the same thermal history through the transformation range.

Annealing

Chemical properties. Common glasses are chemically inert but reactions with many substances do occur slowly. In contact with aqueous media the reaction begins with an exchange of hydronium ions, H_3O^+, for sodium ions, Na^+. (Hydronium ions exist in equilibrium with hydroxyl ions, OH^-, in water.) Immediately, a surface layer starts to build up in which the sodium ions in the glass have been replaced by hydronium ions. For the reaction to continue, sodium ions must diffuse through this surface layer. While the silica itself dissolves slowly, eventually the rate of diffusion of sodium ions and the rate of removal of silica may be such that the surface layer, although moving into the glass, will remain at a constant thickness.

With commercial glasses these conditions are hardly attainable. The complexity of the process makes it very difficult to design accelerated test procedures. Components such as alumina may be added to increase the chemical durability of the glass. The action of alumina is to provide sites that hold sodium ions on the surface, thus producing a more uniform layer of sodium ions and lowering their rate of diffusion. The occupancy of atomic sites on the glass surface determines the behaviour of glass when it is used as an electrode in determining hydrogen ion (H^+) concentrations. When measuring hydrogen ion concentrations, it is important that the surface sites on the electrode be occupied by protons. Normally,

The effect of pH

with glass electrodes, as the hydrogen ion concentration decreases, and the pH value rises (pH is a logarithmic expression of hydrogen ion concentration) to a sufficiently high pH value, the electrode emf (electromotive force or voltage) is no longer proportional to the hydrogen ion concentration; this produces an "alkali error" because the sites have become occupied by sodium ions instead of by protons. If a glass contains a high proportion of alumina, the change point in the measurement can occur at quite low pH values so that over a large range of pH the electrode can be used to determine sodium ion concentrations. There is a close correlation between the electrode behaviour and variation of durability with the pH of the attacking solution.

Optical properties. Optical glass must be homogeneous. Inhomogeneities are of two general types, chemical and physical. Chemical inhomogeneities include gas bubbles and blisters, unmelted batch materials, and devitrification products, while physical inhomogeneity can result from the presence of strain in the glass caused by faulty annealing. Chemical and physical inhomogeneities may result in glass that differs in composition, and therefore in refractive index (the measure of the change in direction of light as it passes from one medium into another) from place to place in the glass.

Batch materials for optical glass manufacture must be of exceptionally high purity in order to give a highly transparent glass free from colour. In order to ensure chemical homogeneity the molten glass is stirred to promote thorough mixing. Physical inhomogeneity is minimized by extremely slow annealing (see above *Mechanical properties*).

The optical characteristics of any glass can be completely described by giving two values n and ν, n being the refractive index and ν the reciprocal of the dispersive power, a property related to the angular separation of two rays of light of different colours, passing through glass, and called the constringence. Thus any glass can be represented by a single point on an n-ν plot, as shown in Figure 4.

The tiny dark-shaded area (Figure 4) represents the entire range of optical glasses available in 1880. All earlier optical glasses were either soda–lime–silica crown glasses or potash–lead–silica flints, with a very limited range of optical properties and a dispersive power that increased with the refractive index. By controlled scientific experiments it was possible to add to these two glass types a range of optical glasses such as borosilicate crowns, barium crowns, barium flints, and borate and phosphate glasses. From these, pairs of crown and flint glasses could be used in the manufacture of lenses that gave images free of colour fringes. The new glasses also had a greater diversity in both refractive index and dispersive power and much more versatile optical systems became possible. The unshaded area of Figure 4 shows how far the range had been extended by 1934.

The remaining shaded area in Figure 4 has been filled in by glasses developed since 1934, owing largely to the introduction of rare-earth and fluoride types. New glasses of high dispersion combined with low refractive index, and of high refractive index combined with low dispersion, were especially sought. It was shown that optical glasses could be made by replacing silica with boric oxide, with additions of rare earths such as lanthana and thoria, and other oxides such as titania, and tantalum and tungsten oxides. These gave optical glasses with very high refractive indices but low dispersions.

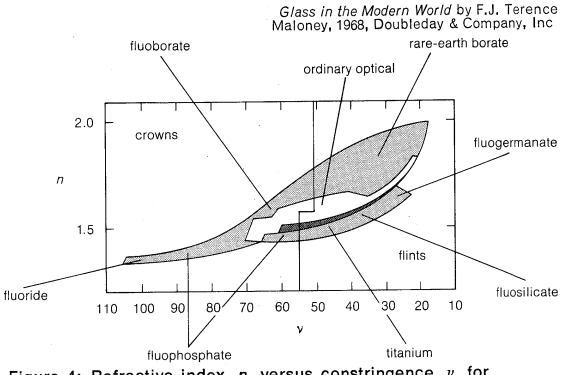


Figure 4: Refractive index, n, versus constringence, ν, for the range of optical glasses (see text).

The range of available dispersions and refractive indices was further extended by the use of fluosilicate flints, or "super flints"—glasses with dispersions greater than conventional flint glasses of the same refractive index. Figure 4 demonstrates the importance of glasses such as the fluoborates, phosphates and germanates, and even types of all-fluoride glasses. These glasses, with their very low dispersions and refractive indices, have an extended range of radiation transmission in both the infrared and ultraviolet.

Thus modern optical designers have available to them a very wide range of glasses with independently varying dispersions and refractive indices. This enables them to design versatile optical systems actually requiring fewer glasses with slightly differing compositions than previously; digital computers now aid in the work.

Electrical properties. All glasses containing alkali ions are electrolytic conductors and show great changes in resistivity with temperature. At room temperature, in addition to volume conductivity through the whole body of the glass, there is a conductivity restricted to the surface, caused by a layer of adsorbed moisture that decreases the surface resistivity. In a very humid atmosphere the surface conductivity of high-alkali glasses can be many times the volume conductivity. The effect is reduced by silicone surface treatment and disappears when the glass is tested in a vacuum.

In volume conductivity nearly all the current is carried by monovalent alkali cations. For this reason, the resistivity is determined mainly by the amount of alkali present and, to a lesser degree, by the network structure and thermal history of the glass. At room temperatures the volume resistivity can be as low as 10^6 ohm centimetres for certain high-alkali glasses; for glasses in which trace alkali conductivity is blocked by the addition of divalent ions, it can be as high as 10^{30} ohm centimetres.

Semiconductors

In recent years a new class of semiconducting glasses has been discovered that shows a decrease in resistivity with increasing temperature, but no polarization. Conduction occurs by electrons or holes (fictitious electron vacancies for which the electrical charge is positive) or both, and such glasses can have a resistivity as low as 10^4 ohm centimetres at room temperature. Typical glasses of this type contain arsenic, selenium, tellurium, and germanium. Glasses that contain large amounts of transition metal ions (for example, vanadium oxide) also conduct by the passage of electrons and show the same low resistivities.

Amorphous semiconducting materials, in particular some chalcogenide glasses, have been used in electronic switches, and a new technology is beginning to develop. An advantage claimed for amorphous switching devices, over conventional crystalline semiconductor switch types, is that they are basically less susceptible, or more insensitive to, the degrading caused by neutron or X-ray irradiation in "hot" environments or to the radiation of outer space.

BIBLIOGRAPHY. Modern general works include: F.J.T. MALONEY, *Glass in the Modern World: A Study in Materials Development* (1967), an excellent, comprehensive book for the general reader describing the history, technology, and properties of glass; and J.R. HUTCHINS and R.V. HARRINGTON, "Glass" in the *Encyclopaedia of Chemical Technology*, 2nd ed., vol. 10, pp. 533–604 (1966), a review article with an extensive bibliography, covering all aspects of modern glass science and manufacture. The following references contain a great deal of useful information: G.O. JONES, *Glass* (1956), a clear, short introduction to the physics of the glassy state; G.W. MOREY, *The Properties of Glass*, 2nd ed. (1954), a collection and discussion of the measurements of the physical properties of glasses and of their dependence on chemical composition; and F.V. TOOLEY (ed.), *Handbook of Glass Manufacture*,

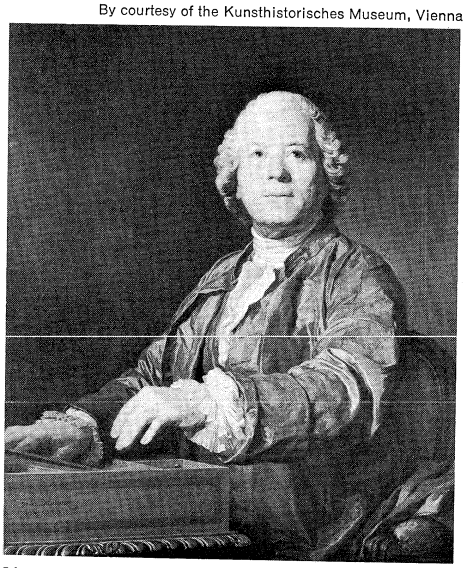
2 vol. (1953), a reference book on all aspects of glass technology. There are no detailed modern texts on the glassy state but recent advances may be found in the following journals: *Physics and Chemistry of Glasses;* and *Glass Technology* (bi-monthly), *Journal of the American Ceramic Society* (monthly), both containing extensive abstracts; and the *Journal of Non-Crystalline Solids* (bi-monthly).

(R.W.D.)

Gluck, Christoph

A leading composer of the 18th century, Gluck reversed the Italian opera tradition in which music took precedence over the drama and thus decisively influenced the development of musical drama in France, Italy, Austria, Sweden, and England. He was born on July 2, 1714, at Erasbach, near Berching, in the Upper Palatinate. His paternal forebears, mostly foresters, were of the border territory between the Upper Palatinate and Bohemia; nothing is known of his ancestors on his mother's side. His father, Alexander Gluck, had moved to Erasbach as a ranger in 1711–12; the family then moved to Reichstadt near Böhmisch-Leipa in Bohemia. Between 1722 and 1727 they lived near Böhmisch-Kamnitz and after this, until 1736, in Eisenberg (near Komotau), where Alexander Gluck held the post of master forester to Prince Philipp Hyazinth von Lobkowitz.

By courtesy of the Kunsthistorisches Museum, Vienna



Gluck, painting by Joseph Siffred Duplessis, 1775. In the Kunsthistorisches Museum, Vienna.

Early life

Christoph Willibald, whose father probably intended for him to continue in the family employment of forestry, at an early age showed a strong inclination toward music. In order to escape from disagreements with his father, the young Gluck finally left home and, supporting himself by his music, made his way to Prague, where he played in several churches, began university studies (1731), and continued his musical studies. He went to Vienna in the winter of 1735–36. There he was discovered by a Lombard nobleman who took him to Milan, where Gluck, apart from fulfilling his duties in the Melzi family chapel, spent four years studying composition with Italian organist and composer Giovanni Battista Sammartini, from whom he learned the new Italian style of instrumental music. Probably six trio sonatas, consisting of two movements with a minuet as conclusion, printed in London in 1746 were the fruits of his studies with Sammartini in Milan. Besides the six "London" sonatas, Gluck probably composed further trio sonatas under Sammartini.

On December 26, 1741, in the Teatro Ducal in Milan, Gluck had his first great dramatic success with his first opera, *Artaserse,* to a libretto by P. Metastasio. Until 1745 there then followed an annual succession of operas for this theatre: *Demofoonte* (1742), *Arsace* (in collaboration with G.B. Lampugnani; 1743), *Sofonisba* (1744), and *Ippolito* (1745). In addition Gluck wrote *Cleonice (Demetrio)* (1742) and *La finta schiava,* a *pasticcio* (1744), for Venice; *Il Tigrane* (1743) for Crema; and *Poro* (1744) for Turin. In these early works, of which mostly only fragments have survived, Gluck largely followed the existing operatic fashion that was Italian—melodic but never grand, charming without intensity. Occasional intensely passionate outbursts and the beginning of characterization, however, foreshadow the great dramatic composer he was to become. In 1745 Gluck, by then well known as an operatic composer, was invited to England at the instigation of Lord Middlesex, director of Italian opera at the Haymarket Theatre, London, in order to challenge Handel's solid hold on London music goers. The plan at first failed when, because of the political chaos caused by the Stuart rising, all theatres in London were closed before Gluck arrived in England. When the situation became calmer, theatrical activities recommenced with a performance of Gluck's opera *La caduta de' giganti* on January 17, 1746; the liberetto, by A.F. Vanneschi, glorified the hero of the day, the Duke of Cumberland, after his victory at Culloden over the forces of Prince Charles Edward, the Stuart claimant to the British throne. This work, as well as Gluck's second London opera, *Artamene,* produced on March 14, 1746, consisted largely of music from his own earlier works, lack of time having forced him to this device. Neither opera met with success. On March 25, shortly after the production of *Artamene,* Handel and Gluck together gave a concert in the Haymarket Theatre consisting of works by Gluck and an organ concerto by Handel, played by the composer. Gluck had won Handel's interest despite the latter's later much-quoted criticism of Gluck's lack of contrapuntal ability (Handel said that Gluck "knows no more counterpoint than my cook"). Gluck himself, according to the Irish singer Michael Kelly, tried to emulate Handel, whom he described as the "divine master of our art."

London: competition with Handel

After he left England (possibly in 1746) Gluck came into contact with two travelling opera companies, one of which, on June 29, 1747, performed his opera-serenade *Le nozze d'Ercole e d'Ebe* at Pillnitz Castle, near Dresden, on the occasion of the double wedding between the electoral families of Bavaria and Saxony. By early 1748 at the latest, Gluck was back in Vienna, at work on Pietro Metastasio's *Semiramide riconosciuta,* with which, on May 14, 1748, the Burgtheater was inaugurated. It proved a brilliant success for the composer. At that time Gluck met his future wife, Marianne Pergin, the 16-year-old daughter of a rich merchant, and in the same year as conductor of the P. and A. Mingotti Travelling Opera company, he travelled via Hamburg to Copenhagen, where he composed the opera-serenade *La contesa dei Numi* in celebration of the birth of the heir to the Danish throne. During the following two winters Gluck was in Prague, where he wrote *Ezio* (1750) and *Issipile* (1751–52). On September 15, 1750, he married Marianne Pergin in the Church of St. Ulrich in Vienna. Their marriage, a harmonious one, was childless. Later Gluck adopted his niece, Marianne. Before the young couple set up a permanent home in Vienna in the winter of 1752–53, Gluck took his wife to Naples for the summer of 1752, where he composed music for Metastasio's drama *La clemenza di Tito* after having rejected the text of *Arsace,* which he had already once set to music.

In Vienna, Gluck soon found a patron in the imperial field marshal Prince Joseph Friedrich von Sachsen-Hildburghausen, who engaged him first as *Konzertmeister* of his orchestra and later as *Kapellmeister.* Gluck gave successful performances of his symphonies and arias at weekly concerts in the prince's palace and made a particular impression with his opera-serenade *Le Cinesi,* which was performed on September 24, 1754, in the presence of the Emperor and Empress at a magnificent celebration at Schlosshof Castle. This success may well have contributed to the decision of the director of the court theatre to entrust the provision of the "theatrical and academic music" for the imperial court to Gluck. On May 5,

1755, Gluck's opera-serenade *La danza* was performed at the imperial Castle of Laxenburg, near Vienna and on December 8 of the same year followed *L'innocenza giustificata*. The following year (1756) Vienna saw *Il rè pastore*, while the first performance of the opera *Antigono* was given during a visit to Rome. In Rome Gluck was created Knight of the Golden Spur, and after his return to Vienna he set to work to provide music for a number of French *vaudeville* comedies imported from Paris. *Tircis et Doristée* (1756) may have been a first attempt at this genre. In these Parisian comedies the dialogue was spoken or sung in the manner of street songs, so-called *timbres*. After 1758 Gluck proceeded more independently and composed for such works as *La fausse esclave*, *L'ile de Merlin* (1758), *La Cythère assiégée* (1759), *Le diable à quatre*, *L'arbre enchanté* (1759), *L'ivrogne corrigé* (1760), and *Le cadi dupé* (1761), which contained, in addition to the overture, a steadily increasing number of new songs in place of the stock *vaudeville* tunes. In *La rencontre imprévue*, first performed in Vienna on January 7, 1764, no *vaudeville* elements remain at all, with the result that the work is a perfect example of *opéra comique*. Gluck gave the scores of *Le cadi dupé* and *La rencontre imprévue* particular charm by using "oriental" instrumental effects. In many of the arias, tuneful melody and programmatic writing foreshadow later developments in Gluck's operatic style: in, for example, the first examples of complex scene development in *L'ile de Merlin* and *L'ivrogne corrigé*.

French *vaudeville* comedies

In February 1761 Raniero Calzabigi, a friend of the adventurer Giovanni Giacomo Casanova, visited Vienna. His libretto for *Orfeo ed Euridice*, partly based on the theories and practices of such literary men as D. Diderot, F.M. von Grimm, Rousseau, and Voltaire, was enthusiastically greeted by Gluck's friends, who immediately brought the two together. On October 17, 1761, the first performance of their first work of collaboration, the dramatic ballet *Le festin de pierre* (Don Juan) was presented. Gluck later composed the music for the dance dramas *Semiramide* and *Iphigenie* (both 1765) to a scenario by Calzabigi and *L'orfano della China* (1774). The choreography for these works was created by the Viennese ballet master G. Angiolini. Together with Calzabigi, Gluck also wrote the three Italian "reform operas," *Orfeo ed Euridice* (1762), *Alceste* (1767), and *Paride ed Elena* (1770).

Gluck himself, in the foreword to *Alceste*, described his and Calzabigi's aims with the words "simplicity, truth and naturalness," demands that primarily affected the libretto. In place of involved plots in the older manner, there was to be a simple, true, and natural action in the tradition of the classical drama; in place of courtly conventions, there was to be a purely human element. The chorus, again on the classical pattern, was to have equal importance with the main characters of the action and participated directly in the dramatic events. The function of the music was, in Gluck's own words (foreword to *Alceste*), "to serve poetry by means of expression and by following the situations of the story, without interrupting the action or stifling it with a useless superfluity of ornaments." The *recitativo secco* ("unaccompanied recitative") was banished (except in *Alceste*); the *recitativo accompagnato*, arioso, aria, chorus, and pantomime were welded together with declamatory style and expressive orchestral writing to form scenes and groups of scenes as parts of a great work of architecture. As Gluck himself confessed, the impulse toward opera reform came from Calzabigi, but it must also be recognized that Calzabigi proceeded largely from the ideas put forward after 1750 by the Parisian poetic and literary circles mentioned above, while important new musical features (*e.g.*, the complex scene development) were the contributions of Gluck's own genius.

Aims of the reforms

Besides the three Italian "reform operas," which were not written as the result of a particular request, there appeared a series of commissioned works, partly after librettos by Metastasio: *Il trionfo di Clelia* (Bologna, 1763), the second version of *Ezio* of 1750 (Vienna, 1763) and, after a short visit to Paris in the spring of 1764, *Il Parnasso confuso*, *Telemacco o sia L'isola di Circe*, and the dance drama *Semiramide*, all written for the second marriage of the Holy Roman emperor Joseph II in 1765. The opera-serenade *La corona*, written in the same year, was not performed owing to court mourning for the death of the emperor Francis I. In Florence on February 22, 1767, Gluck gave performances of his festival opera *Il prologo*, together with T. Traetta's *Ifigenia in Tauride; La Vestale*, the revised version of *L'innocenza giustificata* of 1755, followed in Vienna in 1768; and in Parma in 1769, he presented *Le feste d'Apollo*.

On August 1, 1772, the Paris Opéra was encouraged to stage Gluck's newly completed opera, *Iphigénie en Aulide* (the text, after Racine's tragedy, was by François-Louis Leblanc, bailli Du Roullet); and as Gluck had undertaken to form from the genial Italian style to the more serious opera cultivated by French composers as well as to provide six more similar operas, he went to Paris in the autumn of 1773. The performances of *Iphigénie* on April 19, 1774, and of the French version of *Orfeo* in the summer of the same year met with tremendous success. In Vienna, Gluck was appointed official court composer, but he soon took leave to return to Paris, where the new version of *L'arbre enchanté* in 1775 brought him little success, and the completely rewritten *Cythère assiégée* proved a failure. The French version of *Alceste*, which was produced during his third visit to Paris on April 23, 1776, also met with disapproval. Deeply distressed by this, and the death of his niece, Marianne, Gluck left Paris in May 1776 and returned to Vienna.

In Paris Gluck left both friends and enemies, who began to form two opposing parties: his adherents, the Gluckists, under the leadership of the French writers and music critics L'abbé François Arnaud and Jean-Baptiste-Antoine Suard; his opponents, called Piccinnists after the Italian composer N. Piccinni, who had been prevailed upon to come to Paris in the summer of 1776 to write opera in opposition to Gluck's style. The struggle, which reached its full fury in 1777, never drew either Gluck or Piccinni into active participation in the dispute. Gluck, in Vienna, had completed *Armide* but had destroyed his sketches for *Roland* on hearing that Piccinni was setting the same text for Paris. At the end of May 1777, Gluck returned to Paris.

The opera "war"

At the first performance of *Armide* on September 23, 1777, the war of the theatres reached a climax, but soon after the performance of Piccinni's *Roland* on January 27, 1778, the struggle abated again. Gluck retired to Vienna and his last visit to Paris began at the end of 1778, where he arrived with his two latest completed dramatic works, *Iphigénie en Tauride* and *Écho et Narcisse*. The performance of *Iphigénie* on May 18, 1779, brought him his greatest success in Paris, but *Écho* (which was first performed on September 24, 1779) met with little appreciation. Gluck, who had suffered a stroke during the rehearsals of *Écho*, left Paris for the last time at the beginning of October 1779.

Gluck's great French "reform operas" are more strongly governed by the principle of contrast than are the Italian works; the declamatory style of the vocal line is more marked than in the Viennese operas, and the power and orchestral colour are more intense. The works are constructed in shorter sections, which frequently follow each other without a break, and the spacious conception of the scenes is partly sacrificed in order to achieve a greater degree of dramatic and psychological flexibility.

Gluck spent the last eight years of his life in Vienna and in Perchtholdsdorf nearby, in the care of his wife, continuing to work tirelessly. His attention turned again to F.G. Klopstock's *Hermannsschlacht*, which had occupied him as early as 1770; he revised *Écho et Narcisse* and, together with a Viennese poet, J.B. von Alxinger, produced a German version of *Iphigénie en Tauride*, first performed in Vienna on October 23, 1781, on the occasion of the visit by the Russian grand duke Pavel Petrovich, later Emperor Paul I. At this time the paths of the aging Gluck again crossed those of Mozart, as had already occurred once in Paris; they met on several occa-

sions, but no close personal relationship developed between them. In 1781 Gluck suffered a second stroke, which partly paralyzed him, and his physical powers began to decline. Only a few years before his death he published his *Klopstocks Oden und Lieder* (seven numbers), which must have been written *c.* 1770.

On November 15, 1787, Gluck had a further stroke, from which he died. Two days later he was buried in the central cemetery in Vienna amid general mourning. The requiem mass included a performance of Gluck's own *De profundis*, conducted by A. Salieri.

MAJOR WORKS

OPERAS: *Artaserse* (1741); *Sofonisba* (or *Siface*) and *Ipermestra* (both 1744); *Le nozze d'Ercole e d'Ebe* (1747); *Semiramide riconosciuta* (1748); *Ezio* (1750, second version 1763); *La clemenze di Tito* (1752); *Orfeo ed Euridice* (1762); *Alceste* (1767); *Paride ed Elena* (1770); *Iphigénie en Aulide* (1774); *Orphée et Eurydice* (French version of *Orfeo ed Euridice*, 1774); *Alceste* (French version, 1776); *Armide* (1777); *Écho et Narcisse* (1779); *Iphigénie en Tauride* (1779); *Iphigene auf Tauris* (German version, 1781).

COMIC OPERAS (OPERAS COMIQUES): *L'ile de Merlin* (1758); *L'arbre enchanté* (1759, second version 1775); *La Cythère assiégée* (1759, second version 1775); *Le cadi dupé* (1761); *La rencontre imprévue* or *Die Pilgrime von Mekka* (1764).

DRAMATIC BALLETS AND PANTOMIMES: *Don Juan ou le festin de pierre* (1761); *Semiramide* (1765); *Iphigenie* (1765); *Achille* (1770?).

ODES: *Klopstocks Oden und Lieder* (seven odes for voice and piano with texts by Klopstock, composed *c.* 1770).

CHURCH MUSIC: *De profundis* (after 1785?).

INSTRUMENTAL MUSIC: Six trio sonatas; about 15 sinfonies (overtures).

BIBLIOGRAPHY

Catalog of printed works and bibliography: ALFRED WOTQUENNE, *Thematisches Verzeichnis der Werke von Chr. W.v. Gluck, 1714–1787* (1904); and *Ergänzungen und Nachträge*, ed. by JOSEF LIEBESKIND (1911); C.W. HOPKINSON, *A Bibliography of the Printed Works of C.W. von Gluck, 1714–1787*, 2nd rev. ed. (1967); STEPHAN WORTSMANN, *Die deutsche Gluck-Literatur* (1914); A.A. ABERT, "Gluck, Christoph Willibald," in *Die Musik in Geschichte und Gegenwart*, vol. 5, col. 376–380 (1956).

Editions: *Sämtliche Werke*, 16 vol. (1951–72, in progress); *Denkmäler der Tonkunst in Bayern*, vol. 14/2, *Le nozze d'Ercole e d'Ebe*, ed. by HERMANN ABERT (1914); *Denkmäler der Tonkunst in Österreich:* vol. 21/44a, *Orfeo ed Euridice*, ed. by HERMANN ABERT (1914); vol. 30/60, *Le festin de pierre* (*Don Juan*), ed. by ROBERT HAAS (1923); and vol. 44/82, *L'innocenza giustificata*, ed. by ALFRED EINSTEIN (1937); *Iphigénie en Aulide, Orphée et Euridice, Alceste, Armide, Iphigénie en Tauride, Echo et Narcisse*, ed. by F. PELLETAN *et al.* (since 1873); *The Collected Correspondence and Papers of Christoph W. Gluck*, ed. by HEDWIG and E.H. MUELLER VON ASOW (1962), review with corrections and additions by KLAUS HORTSCHANSKY in *Die Musikforschung*, 17:469–471 (1965).

Biographies: ANTON SCHMID, *Christoph Willibald Ritter von Gluck* (1854), the first comprehensive biography (in German); MARTIN COOPER, *Gluck* (1935), a biographical study based on modern research; ALFRED EINSTEIN, *Gluck: Sein Leben, seine Werke* (1936), Gluck seen as the reformer of Italian opera and French opera in Paris; RUDOLF GERBER, *Christoph Willibald Gluck*, 2nd ed. (1950), an authentic and comprehensive work (in German); ANNA A. ABERT, *Christoph Willibald Gluck* (1959), a popular, reliable account (in German).

Studies: FRIEDRICH J. RIEDEL, *Über die Musik des Ritters C. von Gluck* (1775); ADOLF B. MARX, *Gluck und die Oper*, 2 vol. (1863); ERNST KURTH, *Die Jugendopern Glucks bis Orfeo* (1913); MAX AREND, *Zur Kunst Glucks* (1914); ROBERT HAAS, *Gluck und Durazzo im Burgtheater* (1925); RUDOLF GERBER, "Unbekannte Instrumentalwerke von Christoph Willibald Gluck," *Die Musikforschung* 4:305–318 (1951); DANIEL HEARTZ, "From Garrick to Gluck: The Reform of Theatre and Opera in the Mid-Eighteenth Century," *Proceedings of the Royal Musical Association*, pp. 111–127 (1968); GERHARD CROLL, "Gluckforschung und Gluck-Gesamtausgabe," in *Musik und Verlag: Karl Vötterle zum 65. Geburtstag am 12 April 1968*, pp. 192–196 (1968).

(G.Cr.)

Gnosticism

Gnosticism is a syncretistic theosophical and philosophical religious movement especially important in the 2nd century AD that contributed to the development of Christianity by forcing the early church (in its reaction to Gnosticism) to develop a scriptural canon, a creedal theology, and an episcopal organization. Its emphasis on *gnōsis* (esoteric knowledge) was in sharp contrast to orthodox Christianity's emphasis on *pistis* (faith).

NATURE AND SIGNIFICANCE

Gnosis and Gnosticism

The term Gnosticism is derived from the Greek word *gnōstikos*, one who knows; what he knows is *gnōsis*, or knowledge of an esoteric nature. The context in which *gnōsis* is used determines whether or not it is different from *sophia*, wisdom, or *epistēmē*, generally knowledge acquired by learning or empirical observation. The gnosis claimed by the Gnostic may therefore be different from other kinds of knowledge and viewed as derived not from ordinary sources but from special divine revelation. Gnosticism is a modern term ordinarily used to describe the theological or theosophical systems prominent in the 2nd century of the Christian Era and attacked by leading early Christian theologians.

In essence, the Gnostic sects believed that there is a divine spark in man that has come from the divine realm above and has fallen into the world of fate, birth, and death; it can be awakened by its divine counterpart because of a revelation and can then be reintegrated into the spiritual world. This kind of view, which has obvious antecedents in Hellenistic philosophy and Jewish-Christian religion, came to full expression only after the rise of Christianity. On the other hand, ideas that seem to point toward it already were present in pre-Christian times and in non-Christian circles. The basic question is whether there were pre-Gnostic ideas expressed in Jewish apocalyptic thought (views about God's intervention in history accompanied by cataclysmic events) or even among the Pharisees (Jewish sect, *fl.* 1st century BC–1st century AD), as well as by some early Christians and perhaps by religious thinkers in Egypt or Mesopotamia, or whether there were proto-Gnostic ideas, involving the essence of Gnosticism, already present in Iran or India or even early Greece (Platonism, Orphism, Pythagoreanism). An attempt to clarify scholarly terminology in regard to this question was made at the Colloquio di Messina (Colloquium of Messina) on the origins of Gnosticism in 1966, and a distinction was drawn between Gnosticism, as defined above, and gnosis, which has the more general meaning of "knowledge of the divine mysteries reserved for an élite." Obviously, modern theological concerns are involved, especially as regards the influence of gnosis on early Christianity and the extent to which Christianity itself was, or became, Gnostic.

Ideally, the definition of Gnosticism should follow, not precede, the investigation of the sources available for its study. Unfortunately, the sources are difficult to obtain and to assess. In modern times a whole library of Gnostic documents, Coptic translations from Greek originals, was discovered near Chenoboskion (Naj' Ḥammādī) in Egypt; but the publication of them has been impeded chiefly by political problems related to the establishment of the State of Israel and the difficulties between Israel and the United Arab Republic. All of the documents have been read, but relatively few have been published. There are 53 documents (48 of them different) bound in 13 leather volumes; of this total about 25 have appeared in print. In addition, three Gnostic documents in Coptic in a Berlin collection (two of them also at Chenoboskion) were published in 1955. Other Gnostic documents that have been published are considerably later in date. In ancient times the Christian opponents of Gnosticism had access to documents but often paraphrased or distorted (or both) their meaning, so that considerable caution is needed in dealing with the accounts provided by such anti-Gnostic writers as Irenaeus (*c.* 185), Hippolytus (*c.* 230), and Epiphanius (*c.* 375). They were convinced that their opponents were immoral heretics, and they were trying to present Gnostic views in the worst possible light. In spite of these difficulties, it is possible to reconstruct the Gnostic systems, even in con-

Coptic library near Naj' Ḥammādī

siderable detail, and it is possible to determine their general characteristics.

HISTORY

Iranian and Hellenistic influences

Roots. Iranian dualism, perhaps mediated through heterodox (sectarian) Judaism or Hellenistic thought, may well have provided one of the essential preconditions for the rise of Gnosticism. As early as the 2nd century, the anti-Christian writer Celsus compared the teaching of the Ophite Gnostics about heavens and gates with the mysteries of Mithra. More important is the thoroughgoing dualism of Iranian thought, with its conflict between the good god Ormazd and the evil god Ahriman and the astrological setting later provided. There is also a primordial man who seems also to be the saviour. Clearly, there are remarkable similarities between Gnosticism and the kind of Iranian religious thought known in the Roman world. One should not try, however, to ascribe a single source or impetus to the whole complex movement called Gnosticism. There are analogies with Egyptian and Mesopotamian thought as well as with Iranian.

Gnosticism arose and flourished in the Greco-Roman world during a period when philosophers, especially Pythagoreans and Platonists, were seeking to transcend diversity and contradiction and to find a religious foundation for their thought. Middle Platonists attempted to reconcile Plato with Aristotle and to interpret both as teachers of ancient Greco-Oriental wisdom, liberally employing the allegorical (symbolical) method in their quest. They laid emphasis on the absolute transcendence of the One, or the Good, and sometimes separated it from the creative Mind, identified with Plato's Demiurge (creator). It could be held that knowledge of the One was given only through intuition of revealed truth and that this knowledge was so valuable that it had to be kept secret. Emphasis on transcendence, revelation, and secrecy occurred among Gnostics and philosophers alike. In addition, the very idea of a philosophical or theological system is itself Greek. It should be noted, however, that the Hermetic writings (astrological, occult, and medical documents attributed to Hermes Trismegistos) of the 2nd century and later, often used as witnesses for a "Greek" Gnosticism or gnosis, are essentially expressions of Middle Platonic commonplaces. Some of them expressed pessimism about the world, but this was not uniquely a Gnostic idea—even though the Neoplatonist Plotinus attacked the Gnostic doctrines as anticosmic and, indeed, blasphemous.

Jewish and Christian influences

Similar dualistic tendencies were present in heterodox Judaism, partly among Hellenized Jews like Philo of Alexandria and in more rigorously dualistic form among adherents of apocalyptic thought. Among apocalyptists, alienation from political, social, and religious structures led to belief not only in two ages (the present, evil; the future, good) but also in two "powers" and in knowledge of the future through secret revelation. Themes from Jewish apocalyptic angelology are to be found in several Christian Gnostic systems, and there were at least a few Jewish Gnostics. Generally, however, Gnostics were much more anti-Judaic than Judaic, and it is not possible to find in Jewish circles the idea of an imperfect, subordinate, or evil demiurge (creator).

The Gnostic idea that a redeemer figure provided saving gnosis to the elite is sometimes regarded as pre-Christian, but no evidence is as yet available to support this view. Simon Magus, whom 2nd-century Christians viewed as the founder of Gnosticism, was apparently regarded as "the great Power" of God (Acts 8:10), and the earliest account of Simonian doctrine, about 150, describes him as the supreme god. An expansion of this view by Irenaeus, bishop of Lyon, however, clearly reflects Christian doctrine. Simon appeared to the Jews as Son, to the Samaritans as Father, to others as the Holy Spirit. It seems likely that the idea of a redeemer figure was derived from Christian sources or at least was given impetus by the concurrent Christian proclamation. The constant use of New Testament books among 2nd-century Gnostics at least shows that they regarded them as media of revelation if rightly understood.

Origins and early history. The first Gnostic teachers about whom anything is known were Simon and his so-called disciple Menander, who held that he was sent down himself as the saviour. Both men probably lived within the 1st century, and both were described as teaching the existence of a subordinate female creative principle, superior to world-creating angels, and as encouraging the use of magic. A little later, Saturninus of Antioch drastically reinterpreted the story of Jesus in relation to world- and woman-denying asceticism. There is neither female principle nor magic in his thought. These two types of early Gnosticism, the one leading to "permissive" behaviour, the other to extreme asceticism, set the stage for later developments. During the first third of the 2nd century, Basilides presented a more philosophical system at Alexandria, possibly with some Indian sources, in which the illusory nature of "the nonexistent god" and his nonexistent cosmos was stressed. At the same time, Basilides drew close to Christian texts and provided interpretations of the Old Testament and the Gospels. He was succeeded by his son Isidore, who emphasized ethical problems and allegorical exegesis.

Basilides, Valentinus, and other Gnostic teachers

The most prominent figure of the time, however, was a certain Valentinus, who claimed that his revelation had come when the Logos appeared to him as an infant. Fragments of letters, sermons, and a psalm show that Valentinus expressed mythological ideas in semiphilosophical dress. His later disciples moved in diverse directions. The greatest of them, Ptolemaeus, apparently taught at Rome and created a systematic philosophical theology, ostensibly based on biblical texts. Another, Marcus, was active in Asia Minor or Gaul, or both, and developed the master's teaching in the direction of ritual practices and numerology. In the East, Theodotus laid emphasis on theology, exegesis (interpretive studies), and cultic rites. A disciple of Ptolemaeus, Heracleon, became the first systematic exegete of the Gospel According to John, interpreting it as a Valentinian book. The ideas of the Valentinians aroused the opposition of the early Church Fathers, and by the early 3rd century their influence was vanishing. They were succeeded by other schools, largely anonymous, in which permutations and combinations of Gnostic doctrines were provided.

Within the 2nd century also, anonymous or, rather, pseudonymous, Gnostic doctrines flourished. There were Gnostic gospels and epistles and revelations, the most important of which seems to have been the *Apocryphon* (secret book) *of John*. Irenaeus knew something like it; three copies have turned up at Chenoboskion.

It is a question whether or not Marcion, an important Bible critic and heretic of the mid-2nd century, should be called a Gnostic. The basic difference between him and the others, however, lies only in his rejection of secrecy and secret revelation; otherwise he shared the major tenets of ascetic Gnosticism. In his view there was a supreme, good Father god, unknown but by revelation and superior to the merely just demiurge. The revelation of the demiurge is contained in the Old Testament; the revelation of the Father was given by Jesus, the "saving spirit" who came down from the highest heaven in the 15th year of Tiberius. The teaching of Jesus was corrupted by his disciples, who were trying to present it to Jews. A new revelation was then given to Paul, who alone understood Jesus' gospel and taught it to Luke. Later on, however, both the Gospel and Paul's letters were interpolated. Only Marcion was able to restore their primitive form and content. He also wrote *Antitheses*, a book contrasting the Old Testament with his *Gospel* and *Apostle*. Apparently he was able to remain within the Roman Church between 137 and 144, then being excommunicated. His initial success, or at least toleration, shows how impressive Gnostic teaching could seem in Christian circles before the rise of systematic theology.

Decline. By the early 3rd century the great days of Gnosticism were over. At Alexandria theologians such as Clement and Origen created philosophical theologies

that made Gnosticism seem naïve; Origen converted Valentinians to Christianity. At Antioch, according to Tertullian, one true Valentinian teacher remained. At Rome Hippolytus wrote against Gnostics but relied on archaic sources. The earlier Gnostic doctrines did not survive the publicity that church writers gave them, and the Gnostics' lack of orthodoxy and organization caused their sects to fall apart. Christian writers of the later 3rd century paid hardly any attention to Gnosticism, and when Eusebius wrote his *Ecclesiastical History*, beginning about 303, he knew almost nothing about it. Only a few conventicles survived, at Rome and elsewhere. Epiphanius wrote against heresies toward the end of the 4th century and investigated many old Gnostic ideas; Gnostic women had once invited him to join their group, but his primary targets were not Gnostics but unorthodox Christians like Origen.

Manichaeism and the Mandaeans

The new threat to Christian orthodoxy was presented by Mani (a 3rd-century Iranian prophet who was executed for his teachings) and by his numerous followers, the Manichees. After the year 297, when an imperial rescript denounced this semi-Gnostic group, Christian leaders attacked it on scriptural and philosophical grounds. Eusebius said that Mani made use of "the countless, long-extinct, godless heresies." His words illustrate the general tendency to use Gnosticism as a stick for beating more modern ideas. Among the 65 decrees against heretics cited in the Theodosian Code (439), only two even mention Gnostic groups. Gnosticism was not a significant option in the 4th and 5th centuries.

Thus, Gnostic survivals after the 3rd century exist as gnosis, in the Gnostic sense of esoteric knowledge, but not in Gnosticism as a movement. To be sure, Gnostic documents survived and were imitated in Egypt and Syria. The existence of the "library in a jar" at Chenoboskion does not prove that a Gnostic sect flourished there in the 4th and 5th centuries, much less that it would have had any influence in major centres.

Survivals. Among the Mandaeans, a small sect still existing in Iraq, Gnostic books are still preserved, dating from the 7th or 8th century but using earlier materials. Their mythology is not altogether consistent and seems to reflect the gradual appropriation of Gnostic themes by a group whose centre lay in the cult, especially baptism. The question of Mandaean origins, hotly debated during the early 20th century, is still not solved.

Medieval Christian heretical groups sometimes made use of supposedly primitive apocryphal (noncanonical) books, and it has often been claimed that they represent survivals from early Gnosticism. Generally speaking, such is not the case. In so far as they were influenced by earlier movements, they were acquainted with Manichaeism, not with its forerunners. In more modern times there have been self-consciously Gnostic groups, especially in Europe; the basic source of their ideas seems to lie not in early Gnosticism but in a kind of Faustian Romanticism. The tendency to classify opponents, whether religious, political, or social, as "Gnostics" may have some polemical value, as it seems to have been the case in ancient times, but it is historically invalid. The attempt, for example, to trace aspects of the 20th-century German philosopher Martin Heidegger's thought to Gnostic sources was unsatisfactory since an existentialist description of Gnosticism, influenced by Heidegger, was being used.

A more complex approach to Gnosticism is found among certain Roman Catholic authors who view various decisions of the early church as necessary at the time because of the threat of Gnosticism, but therefore historically conditioned and not binding now. They follow the line taken (*e.g.*, by the American Protestant scholar A.C. McGiffert in 1902), arguing that what underlies the Apostles' Creed is an early formulation directed exclusively against Marcion and therefore not fully relevant afterward.

MYTHOLOGY

Cosmology. The essence of Gnostic knowledge was concerned with man and his place in the universe, a universe understood as vastly larger than the world of sense perception and superior to it. The universe was viewed as consisting of concentric spheres, of which the earth was the centre. These spheres were marked by the circular orbits of the planets, each governed by a deity hostile to spiritual men. Beyond them was the sphere of the fixed stars, equivalent to the Valentinians' *plērōma*, or totality. It was made up of 30 "aeons," or primary spiritual powers, each corresponding to the 30 days of the month. Valentinus' disciple Marcus found astrological significance in the subdivisions of the 30: the highest 8 above were reflected in fire and heat, water and wetness, earth and cold, air and dryness; the next 10 in the 7 heavens, the sphere about them, the sun, and the moon; the last 12 in the zodiac and in the "zones" of the earth. Gnostics knew about the spiritual "doubles" of all these, confirming their calculations by numerology, or religious symbolism: ". . . the Apostles were substituted for the twelve signs of the Zodiac, for, as birth is directed by them, so is rebirth by the Apostles" (Theodotus). The primary function of Gnostic revelation was to free spiritual men from astral determinism: "Until baptism . . . Fate is real, but after it the astrologers are no longer right." Thus the Gnostic knew the whole story of his origin in the *plērōma*, and this liberated him from the power of planets and stars. Some sects even provided diagrams to show what the spiritual world was like. After death, the Gnostic passed upward through the "barrier of evil" to the supreme power of which he knew he was a part.

The astral *plērōma*

Creation of the world. If the supreme unknown power above is perfect and good, why does the universe, experienced as imperfect or evil, exist? Various explanations were given. The most widespread view seems to have been that the spiritual world came into existence out of unions ("syzygies") between pairs of aeons. The 30th (sometimes the 12th) aeon, named Sophia ("wisdom") produced, or her desire produced, an imperfect and inferior demiurge, sometimes called Ialdabaoth (perhaps a parody of Yahweh, the name of the Jewish God). This evil creator made the world or the angels who (sometimes with him) in turn made the world. He is almost always described as the God of the Old Testament, passages from which were cited to show that he was jealous (of the supreme power) but claimed to be the only God. The six or seven (sometimes 12) angels he created bear names like Michael, (S)Uriel, Raphael, and Gabriel (archangels in Jewish and Christian thought) or divine names from the Old Testament (Iao, Sabaoth, Adonai, and other such names). The flourishing angelology of the 2nd century later provided further names. Some systems do not state what the world was made of; the Valentinians held that it came from the ignorance, grief, fear, and consternation of Sophia or her (individuated) desire. This doctrine is clearly mythological, not philosophical. Like the Neoplatonic philosopher Plotinus, the Christian theologian Irenaeus criticized the Gnostic doctrine as anticosmic and blasphemous. In the Gnostic view, the universe was imperfect and hostile to man and could not have been made by the supreme being above.

Role of Sophia

Man. Gnostics were not concerned primarily with the universe as such, however, but with the nature of man, who for them is essentially a spirit dwelling in a body. For some, man's spirit is held prisoner in the body; for others, the situation is more complicated. Saturninus held that the creator angels tried to copy an image of the power above but were too weak to make the copy live. Man's life is due to the pity of the power, which sent down the spark of life (*cf.* Gen. 2:7). Others held that the spark or spirit was given by Sophia or even, by accident, through Ialdabaoth. In any case, it was the spirit alone that constituted true humanity; it alone was capable of salvation. Another way of explaining the human situation was by saying that man was made of matter "in our image" (Gen. 1:26), and possessed soul "after our likeness" and, in addition, the "spirit of life" (2:7); his "coat of skin" (3:21) was his flesh. It is obvious that whereas in Judaism and Christianity man's condition was

explained by Adam's fall, for the Gnostic it was not the first man who fell but his distant ancestor Sophia.

Classes of men

In Valentinian thought there were three classes of men: spiritual; psychic (animate or possessing soul: ordinary Christians); and material, or fleshly. Spiritual men (Gnostics) were, it was believed, perfect and would be "by nature saved"; psychic men possessed free will and could go upward or, in decay, downward; material men could not be saved, since they were "by nature sons of the devil." Obviously Pauline language was important for making such a distinction (*cf.* I Thess. 5:23; I Cor. 2:6–3:3), though Paul's differentiations were moral, not ontological (concerning grades or levels of being). According to the Valentinian Gnostic Ptolemaeus, the three classes also were to be found in sayings of Jesus, but obviously his claim was the result of allegorical exegesis.

Some Gnostics were concerned with the origin of various parts of the body and control over the parts by creator angels, by the signs of the zodiac, or by 360 or 365 angels related to the days of the solar year. Such doctrines are astrological in origin.

Redemption. The spirit of man was believed to dwell in an alien body in an alien world. How was he, like the Prodigal Son of Luke 15, to "come to himself" and go to his true father? The answer the Gnostics offered was that man needed salvation through knowledge. The Simonians held that Simon, the supreme deity himself, came down through the spheres and rescued his "first thought," apparently related to human spirits generally. He had earlier been present among men as Son, indeed as Jesus, but suffering would have been irrelevant and he was not crucified. (Basilidians even held that Simon of Cyrene was crucified in Jesus' place.) Generally, however, the Gnostics claimed that Jesus was a man into whom the spiritual Christ came from above; Christ was produced in the spiritual world. Because the Christ came from above he could reveal the mysteries of spiritual cosmology (order of the world) and cosmogony (origin of the world). He

Role of Christ

was responsible for Jesus' miracles and as spirit departed from Jesus at the crucifixion (*cf.* Luke 23:46). Others held that Christ was a psychic being like his father the creator and spoke of him through the Hebrew prophets; what descended upon him was the spiritual Saviour. (The Gnostic Menander held that he himself was the saviour.) According to Basilides, Jesus became the Saviour when illuminated by light from above; he was subject to astral influences but ascended in order to lead other spiritual beings upward. Obviously, the revelation of knowledge was what mattered. The cross existed as a symbol of the dividing line between the spiritual world and what lies below. Even when Jesus was described as crucified, his crucifixion was simply the occasion for his ascension and not, as in Christian thought, a sacrifice. Christ, as in the Gospel According to John, is the revealer. In the *Gospel of Truth* Jesus was "nailed to a cross" but as teacher. His mission was to speak "in a mystery privately to his disciples and apostles" (Carpocratians).

In some measure Jews and Christians held similar views. The death of the Qumrān Teacher of Righteousness in the Jewish sect revealed in the Dead Sea Scrolls was not redemptive, and Christian apologists could compare Jesus' death with that of the good teacher Socrates. Generally, however, Christians believed that Jesus' death, not just his teaching, had redemptive significance.

Eschatology. Though mystical experience was possible during life, the true ascent of the spirit took place after death when it imitated the saviour or Christ and moved upward through the spheres governed by the planetary angels. (Menander taught that his disciples would not die, but time proved him wrong.) The spirit had learned formulas to proclaim that he traced his origin to the preexistent one and was returning to his own. Valentinians held that "the spirituals will put off their souls and will become intelligent spirits . . . given as brides to the angels about the Savior." In the system of Saturninus the same point is made very simply: "after death the spark of life returns to what is of the same nature as itself, and the other elements of man's composition are dissolved into what they were made from." There is thus a cosmic tendency for the spirit to rise (as Basilides says), and it is simply reinforced by what the Saviour taught.

Fate of the material world

But what would happen to the universe after the withdrawal of the spiritual elements in it? Ptolemaeus held that "then the fire hidden in the universe will shine forth and ignite and become effective in consuming all matter along with itself and finally will become non-existent." (This resembles the Stoic doctrine of cosmic conflagration.) On the other hand, Basilides taught that the world would remain but God would bring ignorance upon it, so that souls remaining below will not desire anything contrary to their nature and thus experience torment. "Everything that remains in its place is imperishable; it is perishable only if it wants to pass beyond its natural limits."

WORSHIP, PRACTICES, AND INSTITUTIONS

Sacred writings. The Gnostic sects of the 2nd century generally made use of Hebrew and Christian religious writings, employing the allegorical method to extricate Gnostic meanings from them. Some could find three classes of men in various parables of Jesus. Apocryphal gospels, letters, acts, and apocalypses were often employed. The Neoplatonist philosopher Porphyry attacked Gnostics for composing fraudulent apocalypses ascribed to Zoroaster, Zostrianus, and Allogenes; a "discussion by Zostrianus" and a book called *Allogenes* are among the books found at Chenoboskion. The *Apocryphon of John*, supposedly based on a revelation received by John the son of Zebedee, reinterprets the first few chapters of Genesis and says the truth is "not as Moses said."

Prominent among the Gnostic books is the *Gospel of Thomas*, first known in modern times from Greek fragments found at Oxyrhynchus in Egypt and extant in a Coptic version from Chenoboskion. This gospel consists of nothing but sayings of Jesus, derived partly from the synoptic gospels or traditions and partly from the Gnostics' own traditions. Another "gospel" is the *Gospel of Philip*, an esoteric collection of Valentinian doctrines. The *Gospel of Mary* (Magdalene) contains postresurrection teachings of the Saviour (Jesus) to the disciples, especially to Mary (whom he loved more than the others). Andrew and Peter doubt, but Levi persuades the others that she received the true Gnostic doctrine.

The Chenoboskion collection includes apocalypses ascribed to Adam, James, Peter, and Paul, and other apocalypses are known because of mention by church writers.

In addition, Gnostics generally believed that by allegorization they could interpret all religious literature in relation to their systems. They claimed to know the true meaning of the mystery religions and of Greek and oriental mythology; they could explain the hidden significance of Homer, Hesiod, and other poets. Anyone who knew the story of "Elohim and Eden" in the Gnostic *Baruch* could understand the labours of Hercules and the stories about Zeus and Leda and Danaë. Hermetic writings were included in the Chenoboskion collection.

Teachers. The picture the Church Fathers give of Gnostic teachers in succession is wrong. Menander was not the disciple of Simon; Basilides and Saturninus were not their followers. There were real "successions" among the Gnostics, however. Basilides claimed that he had been

Basis of Gnostic teachers' authority

taught by a certain Glaukias, who (like the evangelist Mark) had been an "interpreter" of the Apostle Peter. In turn, his principal pupil was his own son Isidore. Similarly, Valentinus said that he had learned from Theodas, a companion (like Luke) of the Apostle Paul. The names of five of his chief disciples are known, but early in the 3rd century Tertullian could claim that only one teacher still preserved Valentinus' authentic doctrine. At an earlier date, Ptolemaeus was writing to Roman Christians and claiming that Valentinians too had "received the apostolic tradition by succession." Other Gnostics had ideals of secret traditions derived from apostles, but in general the links in the tradition, like the later teachers, remained anonymous. A certain Justin, who wrote a book entitled *Baruch*, provides one of the few exceptions.

Organization. Most Gnostic groups seem to have been organized as schools, in which the authoritative teaching was transmitted, interpreted, and kept secret. There must have been regular meetings of some sort, for Gnostics were in danger of arrest during times when Christians were persecuted. Among some groups there were regular meetings for sacramental and prophetic worship. In the late 4th century such meetings were forbidden and Gnostic books were burned. Gnostics seem to have gone underground, although an early 4th-century inscription refers to "the synagogue of the Marcionites of the Lord and Savior Jesus Chrestos" (the Greek word *chrēstos* means "good"), and as early as the 2nd century there were Marcionite bishops and presbyters.

Rites. Gnostics disagreed among themselves as to the importance of rites. Many rejected conventional worship, including prayer, fasting, and almsgiving, not to mention Baptism and Eucharist. Some insisted that "true redemption" was given the "inner spiritual man" through gnosis, while others held that a rite called "redemption," often involving "Hebrew names" and unguents, was important and necessary. Among the liturgically minded there were Eucharists with bread and wine and Baptisms with water and chrism, just as among Christians of the time. The followers of Basilides celebrated the day of Christ's Baptism, presumably because then the spiritual Christ descended upon the human Jesus, and the day of his death, because then Christ left him. It is hard to tell whether or not there was actually a rite called "spiritual marriage," for the language of the *Gospel of Philip* is ambiguous; certainly Irenaeus believed they practiced this rite. "Redemption" and other rites seem intended primarily to aid the divine spark in its ascent, following the death of the body, to the divine realms above. Formulas were provided for use in passing through the various planetary spheres.

In addition, some groups had rites involving "prophecy." The leader in worship, according to one account, would inform his hearer that grace from above had descended upon her and would command her to open her mouth and prophesy. If she replied, "I have never prophesied and do not know how to prophesy," he would say, "Open your mouth and say anything, and you will prophesy." The notion of Epiphanius and other church writers that some Gnostics participated in sex orgies as religious rites seems somewhat exaggerated but cannot be disproved. There is not much evidence for it in the early days of Gnosticism.

Libertinism and asceticism

Ethics. Gnostic ethics ran the gamut from compulsive promiscuity to extreme asceticism. What the Gnostic normally rejected was conventional law or wisdom, especially as expressed in the Old Testament. This kind of law, they held, had been imposed on men by the world-creating angels, who desired to enslave them. The Carpocratians and the Cainites (two Gnostic sects) insisted that it was necessary to experience "everything" in order to be freed from the control of these angels. Even the Valentinians argued that ordinary Christians, who were "from the world" as Gnostics were not, could not love a woman and possess her, while Gnostics, "in the world" but not "from" it, had to do so. Church writers accused many Gnostics of employing love magic. On the other hand, there were Gnostics who claimed that marriage and generation originated with Satan (as in Gen. 2–3, taken literally); some also abstained from eating meat.

Basilides explicitly taught that suffering took place only in consequence of sin, actual or potential, and quoted Job (14:4) to prove that "no one is pure . . . from defilement." His son Isidore claimed that sin was always voluntary and (like Epicureans) argued that sin could involve what was natural but not necessary. Epiphanes followed radical Platonic lines and insisted that equality and sharing were intended by God; men should not treat wealth or women as private possessions.

GNOSTICISM AND CHRISTIANITY

It is difficult to disentangle the relations between early Christianity and Gnosticism. The chronology is rather obscure, and it is hard to differentiate causes from effects. Two main views have been held: (1) that Gnosticism arose before Christianity and influenced such Christians as Paul and John even though they argued against it and (2) that Gnosticism was a form of Jewish heterodoxy or Christian heresy or both, probably not known to the New Testament writers but flourishing from the early 2nd century onward. Conceivably a third view could be supported: that the beginnings of the Gnostic and the Christian movement were virtually simultaneous and that the story is therefore one of cross-fertilization. It would appear that neither the Church Fathers nor the Chenoboskion documents provide evidence for a final decision. The fact that the major Gnostic teachers claimed to have been taught by disciples of Jesus proves no more than that 2nd-century Gnosticism owes much to Christianity. It does not indicate what underlay the doctrinal struggles reflected in I Corinthians, Colossians–Ephesians, John, and I John.

Contributions of Gnosticism to Christianity

One thing is clear, however. In developing its organizational and doctrinal structures the church owed much to the Gnostics. The structures were not Gnostic in origin, but without the Gnostic crisis they would have taken longer to evolve. The formulation of creedal symbols, the canonization of the New Testament Scriptures, and the emphasis on episcopal authority were made necessary by the Gnostics' claims. Moreover, in some measure the Gnostics were the first theologians. They made it necessary for Irenaeus to develop what systematic thought he had and prepared the way for treatises like Origen's *On First Principles.* Origen owed much to church tradition and philosophical teaching, and much also to his Gnostic forerunners, some of whom raised the kinds of questions he tried to answer.

The Gnostics influenced Christianity positively as well as negatively. They kept alive the great issues of freedom, redemption, and grace, which, after the times of Paul and John, were not adequately discussed by 2nd-century Christian writers. The Gnostic solutions did not prove tenable in the Christian framework, but at least the Gnostics kept such issues alive. At a later date the theology of Augustine owed a good deal to the experience he had had as a Manichee.

INTERPRETATIONS OF THE ROLE OF GNOSTICISM

Patristic. For the Church Fathers Gnosticism was a perverse heresy; particular Gnostic sects, they supposed, often originated when unstable Christians sought church office and failed to obtain it. Sometimes they explained the consequent heresy as derived from Jewish sectarianism; sometimes they ascribed it to the influence of one Greek philosophical school or another. Irenaeus suggested that the Gnostic simply loved novelty: "they pretend to find something new every day and to produce what no one ever thought of." On the other hand, most of the fathers believed in the genetic development of Gnosticism: Simon founded it, and one could trace its history through disciples who came after him. Unfortunately this explanation is quite wrong. There was no uniform original doctrine that spread like the branches of a tree. Instead, the various systems reflect a common Gnostic attitude but, along with it, remarkable diversities in doctrine. What must underlie the diversity is a variety of approaches to the world, to human nature, and to ideas of knowledge. Again, when the Church Fathers related Gnosticism to philosophy they were apparently trying to bring it under the judgment of a common criterion. All they could say of the Gnostic myths is that they were myths. They had to engage in rational argument in order to show that the myths were irrational.

Modern. Modern man, more aware of the role of myth than were the Church Fathers, has found it possible to understand aspects, at least, of the Gnostic world views. The leading interpretations have been made in psychological and in existential terms. The principal student of Gnostic ideas among psychologists has been C.G. Jung, who was long interested in Basilides and Carpocrates and with Karl Kerényi wrote "essays on a science

of mythology." In his theory of "archetypes," primordial images at work in the human psyche and presented as symbolic representations, Jung was fairly close to the Gnostic depiction of aeons and their unions and separations. Speaking of the "body schema as archetype of the first man" in Gnostic thought, his pupil Erich Neumann wrote that ". . . early man lived in the middle of . . . psychophysical space in which outside and inside, world and man, powers and things, are bound together in an indissoluble unity." According to "modern consciousness," Gnostic myths are "projections, i.e. . . an outside experience of archetypal images." Of Sophia, Neumann wrote: "In the patriarchal development of the Judaeo-Christian West . . . the goddess, as a feminine figure of wisdom, was disenthroned and repressed. She survived only secretly, for the most part on heretical and revolutionary by-paths." One may add that among the Valentinians there is a double Sophia, in some measure reflecting the "elementary characters" of the primordial goddess that Neumann calls positive and negative.

Psychological and existential interpretations

Relying to some extent on the thought of Martin Heidegger, Hans Jonas started from the relatively late Mandaean literature and described "Gnostic imagery and symbolic language" in terms reminiscent of Existentialism. He spoke of alienation and estrangement, of "fall, sinking, capture," of "forlornness, dread, homesickness," and "numbness, sleep, intoxication." Such terms are indubitably present in Gnostic thought, and Jonas rightly endeavoured to show what bound Gnosticism together, laying less emphasis on its origins and on diversities within it. He was concerned with morphology, not genetics. And there are indubitable similarities between "Gnosticism and modern nihilism," to which Jonas has drawn attention.

CONCLUSION

The basic question, as Jonas pointed out, is "what is Gnosticism?" But it must be answered in modern times by asking "what was Gnosticism?" The subject under discussion, whether or not alive today, is a phenomenon of the past, and therefore the problems of origins, sources, and possible development need to be considered, as well as the problem of definition. With Jung, Jonas, and others, one finds the goal in an understanding of what gnosis and Gnosticism were as differentiated from their sources and even their influences. It is not enough to call Gnosticism "the acute Hellenizing of Christianity" (Adolf von Harnack) or to trace its ingredients to Greek or Oriental ideas without explaining how and why men found it meaningful. It arose in an age of syncretism, but it was not merely syncretistic. It was not precisely Greek, Jewish, or Christian, though elements of all existed in it. To a greater degree it was an anti-Greek, anti-Jewish, anti-Christian movement; in the 2nd century, and in Manichaean form still later, it captured at least for a time the imaginations of such theologians as Basilides, Valentinus, and Augustine.

BIBLIOGRAPHY

Collections of Gnostic works: W. VOLKER (ed.), *Quellen zur Geschichte der christlichen Gnosis* (1932), Greek and Latin texts from the Church Fathers; R.M. GRANT (ed.), *Gnosticism: An Anthology* (1961), Eng. trans. of early Gnostic writings, including some Coptic documents, with discussion and references; J.M. ROBINSON, "The Coptic Gnostic Library Today," *New Testament Studies*, 14:356–401 (April 1968).

Studies: W. BOUSSET, *Hauptprobleme der Gnosis* (1907), classical but outdated; A. VON HARNACK, *Marcion: das Evangelium vom fremden Gott*, 2nd ed. (1924), the only collection of sources for Marcion, with full discussion; H. JONAS, *Gnosis und spätantiker Geist*, 2 vol. (1934, 1954), full treatment, both phenomenological and existentialist, *The Gnostic Religion*, 2nd ed. (1963), abbreviated but indispensable, and "Gnosticism and Modern Nihilism," *Social Research*, 19: 430–452 (Dec. 1952); G. QUISPEL, *Gnosis als Weltreligion* (1951), Jungian; R.M. WILSON, *The Gnostic Problem: A Study of the Relations between Hellenistic Judaism and the Gnostic Heresy* (1958), a valuable survey.

Particular aspects: G.G. SCHOLEM, *Major Trends in Jewish Mysticism*, 3rd ed. (1954) and *Jewish Gnosticism, Merkabah Mysticism, and Talmudic Tradition* (1960); R.M. GRANT, *Gnosticism and Early Christianity*, rev. ed. (1966).

The Messina Colloquium: U. BIANCHI (ed.), *Le origini dello gnosticismo* (1967), a collection of papers (in French, German, English, and Italian) presented at the Colloquium of Messina, April 1966, and *Studi di storia religiosa della tarda antichità* (1968), additional papers.

(R.M.G.)

Gobi (Desert)

One of the great desert and semidesert regions of the globe, the Gobi (from Mongolian *gobi*, meaning "waterless place") stretches across the vastness of Inner Asia over huge stretches of both the Mongolian and Chinese People's republics. Contrary to the perhaps romantic image long associated with what—at least to the European mind—was a remote and unexplored region, much of the Gobi is not sandy desert but bare rock. One may drive over this surface by car for long distances in any direction; northward, say, toward the mountains of the Altai and Hangayn ranges or eastward toward the city of T'ien-shan or southward toward the Pei Shan-mo. To the west, 1,000 miles from the Gobi's eastern limits, lies the Sinkiang region, a great basin enclosed by the Plateau of Tibet to the south and the Tien Shan ranges to the north. The desert as a whole occupies a vast arc of land 1,000 miles long and 300 to 600 miles wide. The region first became known to Europeans through the remarkable 13th-century descriptions of Marco Polo. Modern geographical study of the Gobi has mainly been undertaken by Russian scholars, though a series of works has appeared in Mongolian and Chinese since the 1960s. For the purposes of the present article, the Gobi is defined as lying between the Altai Mountains and Hangayn Nuruu (Hangayn Mountains) in the north; the eastern Tien Shan in the west; and the A-erh-chin Shan-mo, Pei Shan-mo, and Yin-shan Shan-mo in the south. For related information on physical geography, see TAKLA MAKAN DESERT, which covers a western region sometimes included in the Gobi; TIEN SHAN (MOUNTAINS); ALTAI MOUNTAINS; for information on human geography, see SINKIANG UIGHUR; KANSU; INNER MONGOLIA; and MONGOLIA.

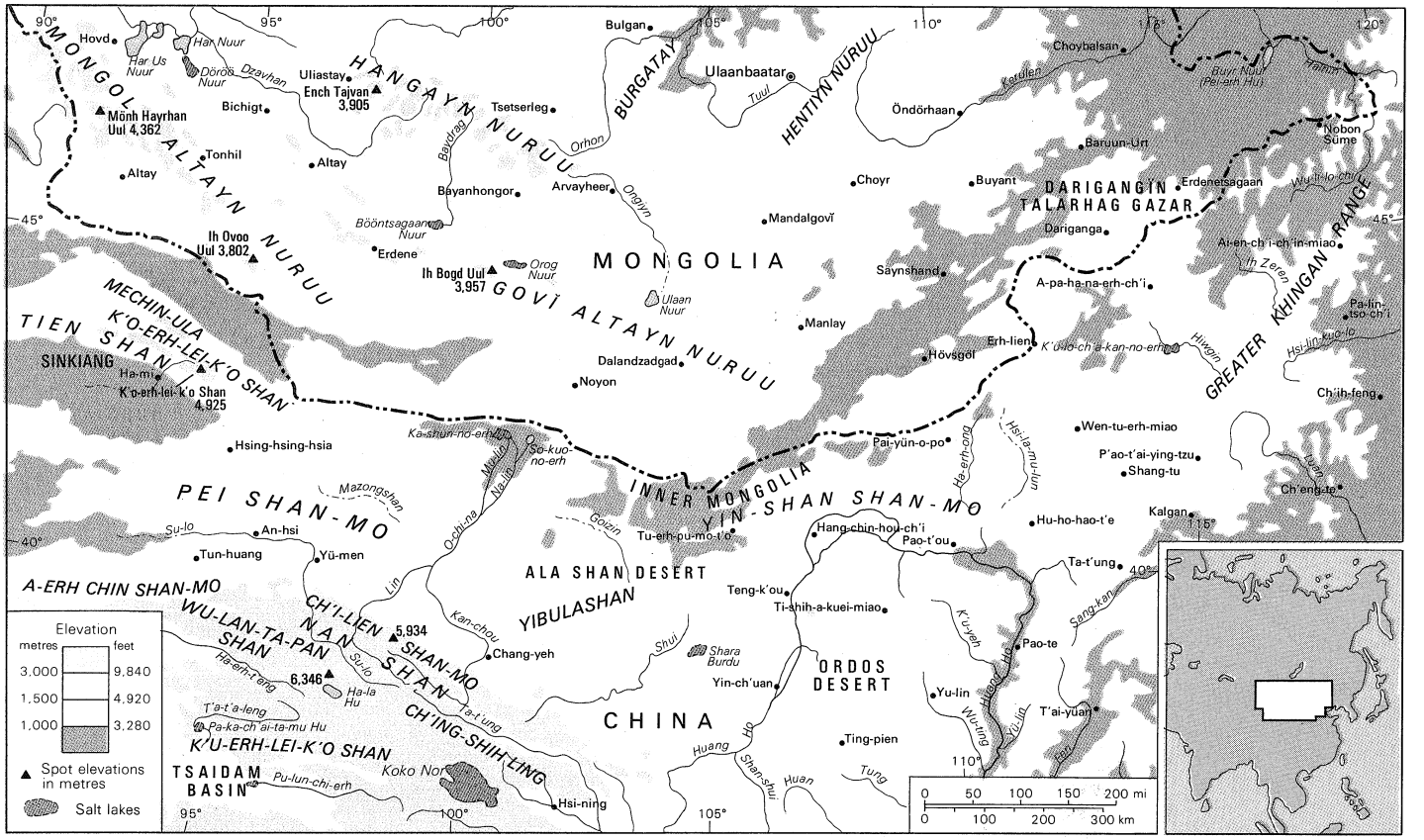
Physical features. The Gobi may be subdivided into the Ka-shun, Dzungarian, and Trans-Altai Gobi (south of the Mongolian Altai Mountains) in the west and the Eastern, or Mongolian, Gobi in the centre and east.

Regional subdivisions

The Ka-shun is bounded by the spurs of the Tien Shan to the west and the Pei Shan-mo in the south and rises as high as 5,000 feet. It is gently corrugated, with a complex labyrinth of wide hollows separated by flat hills and rocky crests sometimes rising more than 300 feet above the plain. The desert is stony and waterless, though salt marshes lie in the secluded depressions. The soil is grayish brown and contains gypsum. Vegetation is very rare, though richer in the riverbeds, where there are individual shrubs of tamarisk, zaysansky haloxylon and nitre bush (both saltworts), and annual halophytes (plants growing naturally in soils containing certain salts).

The Trans-Altai Gobi is situated between the eastern spurs of the Mongol Altayn Nuruu (Mongolian Altai), the Govĭ Altayn Nuruu (Gobi Altai) in the north and east, and Pei Shan-mo in the south. The plain is elevated, sharp, and rugged. Alongside the plains and the isolated group of low, rounded hills is a fairly extensive mountain area, extending over six miles out into the plain. The mountains are very barren and broken up by dry ravines. The western section of the Trans-Altai Gobi is basically a plain, too, but interspersed with small raised areas and furrowed by dry riverbeds and, again, with extensive salt marshes. In the central portion this fragmentation increases, and mesas (flat-topped, steep-sided hills) appear along with dry gullies ending in flat depressions, occupied by *takyr* (clayey tracts). The Trans-Altai Gobi is quite parched, with annual precipitation of less than four inches, though there is always water underground. There are virtually no wells and springs, however, and vegetation is very sparse and almost useless for livestock.

The Dzungarian Gobi is situated north of the Ka-shun Gobi, between the eastern spurs of the Mongol Altayn



The Gobi

Nuruu and the eastern extremity of the eastern Tien Shan (the K'o-erh-lei-k'o Shan). It is like the Trans-Altai Gobi, and its edges are fractured by ravines, alternating with residual hills and low mountain ridges.

The Eastern Gobi is of similar character, with altitudes varying from 2,300 to 5,000 feet, but enjoys rather more precipitation—up to eight inches a year—though with virtually no rivers. The underground waters are relatively abundant and only partly mineralized. They are also near the surface, feeding small lakes and springs. The vegetation, however, is sparse: herb wormwood in coarse, grayish-brown soil. In the moister depressions there are the usual salt marshes and grassy swamps. In the northern and eastern outlying regions, where more precipitation occurs, the landscape of the desert gradually mellows, sometimes even becoming steppes.

Geologic character and fossils. The Gobi's various chalky plains are chiefly Paleocene to Recent (up to 65,000,000 years old), though some of the low, isolated hills are older. The terrain contains small masses of shifting sands. In the central Gobi, Mesozoic remains of dinosaurs (65,000,000 to 225,000,000 years old) and Paleogenic and Neogenic fossils of mammals have been found. The desert also contains Paleolithic and Neolithic sites occupied by ancient man.

Fossils

Climate. The climate is acutely continental and dry, ranging from January's −40° F (−40° C) to 113° F (45° C) in July. Winter is severe; spring is dry and cold; and summer is warm. The annual total precipitation varies from 2.7 inches in the west to more than eight inches in the northeast (the maximum falls in summer), and in the eastern regions the impact is quite monsoonal. North and northwestern winds prevail over the Gobi.

Drainage. Aridity in the Gobi depends on the relative chalkiness of the soil and has been aggravated by the strong mountain structure to the west. The lakes have correspondingly shrunk, leaving a series of terraces considerably farther from and higher than the present shorelines. Indeed, Wu-lan Hu (east of the Pei Chan-mo), Orog Nuur, and Bööntsagaan Nuur (in the easternmost Mongolian Altai) are but shadows of their former selves.

The drainage of the desert is largely underground; surface rivers have very little constant flow. Mountain streams are confined to the Gobi's fringes and even then quickly dry up, as they disappear into the loose soil or the salty, enclosed depressions. Many rivers flow only in summer. On the other hand, subterranean water is widespread and of sufficient quality to allow cattle raising.

Soil types and plants. The soil is chiefly grayish-brown and brown carbonaceous (rich in carbon), gypseous (containing gypsum), coarse gravel, often combined with sandy salt marshes and *takyr*. Vegetation is sparse and rare. On the plateau, and on the plains beneath the mountains, small bushlike vegetation occurs: echinochloa (a type of succulent grass found in warm regions), yellowwood bean caper, winterfat (a shrub covered with densely matted hairs), Dzungarian reaumur, nitre bush, and bushlike halophytic vegetation. In the salt marshes, too, halophilous groups prevail: potash bush, Siberian nitre bush, tamarisk, and annual halophytes; in the sands grow zaysansky haloxylon, the sandy wormwood, and sparse perennial and annual herbs such as the Gobi kumarchik and timuriya. In semidesert tracts vegetation is richer, belonging to the herbaceous-wormwood groups: Gobi feather grass; the annual Gobi kumarchik (*Agriophylum gobicum*); the perennials timuriya (*Timouria villosa*) and snakeweed (*Cleistogenes* sp.); and cold wormwood. There are herb meadows with rhizome Mongolian onions and herb salt marshes with sparse beds of bushlike *Caragana*. In the Gobi Altai and other high mountains, desert-grass steppes completely cover the lower slopes, and, on the upper parts, mountain versions of the feather-grass steppes appear.

Areas with fairly rich vegetation

Animal life. Animal life is varied, with such large mammals as the wild camel, the ass kulan (*Equus asinus ferus*), the Przhevalsk wild horses (named after a town in the Kirgiz region of the Soviet Union), the dzheiran gazelle, and the dzeren (an antelope). Rodents include marmots and gophers, and there are reptiles.

The population density is small—fewer than three persons per square mile—mostly Mongols with Chinese in Inner Mongolia. The Chinese have increased greatly in

recent years. The main occupation of the inhabitants is nomadic cattle raising, though, in regions where the Chinese are concentrated, agriculture is predominant. The living quarters of the Mongol nomads are felt yurts (types of tent), while the Chinese farmers live in *fanzy*, clay homes built from crude brick.

Farming and land use. The province of the Gobi and its semidesert sections is mainly a livestock region, sheep and goats comprising 57 percent of the total herds. Next come the large-horned cattle, accounting for 25 percent. Horses only make up about 4 percent of the total and, together with the large-horned cattle, are concentrated in the lusher semidesert of the southeastern region. About 15 percent of the livestock still consists of two-humped camels, still used for transportation in some areas. Pasturage for cattle is available all year round because of underground waters. Livestock raising is nomadic, and herds move ten times a year, migrating as much as 120 miles between extreme points.

Natural resources

Useful mineral deposits are scant, but there are pockets of petroleum around the city of Saynshand (about 250 miles southeast of Ulaanbaatar), and, in the Mongolias as a whole, salt and light metal ores are mined. Agriculture is developed only along the river valleys.

Transportation. The Gobi is intersected by railroads in the east and west. There are several highways: from the town of An-hsi to the town of Ha-mi across Pei Shan-mo and the Ka-shun Gobi, from the town of Kalgan (northwest of Peking) to Ulaanbaatar, and from Ulaanbaatar to Dalandzadgad (some 300 miles south-southwest of Ulaanbaatar). In addition, various ancient caravan tracks crisscross the Gobi in all directions.

Future prospects for the development of the Gobi area rest upon industrial development of the mineral wealth and on cattle and sheep breeding. In the early 1970s, the rate of economic development in those Gobi areas in the Mongolian Peoples Republic was faster than that for the desert portions of the Inner Mongolia regions of China.

BIBLIOGRAPHY. Additional information on the Gobi Desert may be found in MILDRED CABLE, *The Gobi Desert* (1942); SVEN A. HEDIN, *Åter till Asien* (1928; Eng. trans., *Across the Gobi Desert*, 1931, reprinted 1968); and ALONZO W. POND, *Climate and Weather in the Central Gobi of Mongolia* (1954).

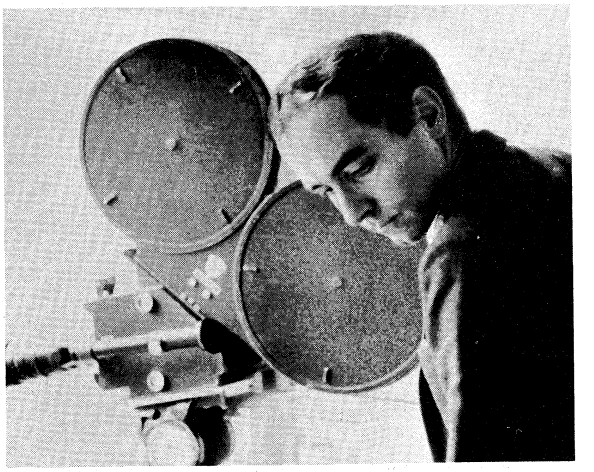
(M.P.Pe.)

Godard, Jean-Luc

Of all the former critics of the journal *Cahiers du Cinéma* who, in the late 1950s, became the major French film makers of the "New Wave" (*e.g.*, François Truffaut, Alain Resnais), none has had a more prolific or more controversial career than Jean-Luc Godard. No less boldly radical in his political views than in his aesthetic convictions, he has written and directed highly personal films, such as *Breathless*, that are fascinating in expressing so forcefully his readiness to sacrifice both his art and his way of life to a tortuous conscience. Although Godard's films have provoked a wide range of critical responses, it is indisputable that they have exercised an immense influence on film making throughout the world.

Though he was born in Paris, on December 3, 1930, Godard spent his formative years on the Swiss side of Lake Geneva, where his father directed a clinic. His higher education comprised study for a degree in ethnology at the University of Paris, interminable student cafe conversations, and a labouring job on a dam, which inspired his first short film, *Opération Béton* (1954). His ethnological interests link with the influence on his work of Jean Rouch, an anthropologist who became the first practitioner and theoretician of the documentary-like film style *cinéma vérité* ("cinema truth"). Film makers of this school employ lightweight television equipment in order to observe their subject with the utmost informality and so completely without preconceived bias that the theme and motifs of the film emerge only while shooting, or even later, at the editing stage. Godard (often writing under the pseudonym of Hans Lucas) also evinced a then unfashionable devotion to such studio-bound German expressionist film makers as Fritz Lang.

Improvisatory techniques



Godard.
By courtesy of The Museum of Modern Art Film Stills Archive, New York

Godard's first feature film, *Breathless*, which was produced by Truffaut, his colleague on *Cahiers*, won the Prix Jean Vigo in 1960. It inaugurated a long series of features, all celebrated for the often drastic nonchalance of Godard's improvisatory film-making procedures. *Breathless* was shot without a script; Godard sketched the dialogue overnight and revised it between and during rehearsals. In subsequent films he even resorted to speaking the characters' replies to the actors from behind the camera during takes. Thus, he used improvisatory techniques sometimes to observe reality, sometimes to impose his own vision, and often to interrelate the two so as to create a strangely abstract effect.

His 1963 film *Contempt*, based on a story by the Italian novelist Alberto Moravia, marked his only venture into orthodox and comparatively expensive film making. Afterward, he maintained an almost unique position as an absolute, independent creator, using extraordinarily cheap alfresco production methods and enjoying repeated success on the international "art cinema" circuit.

Breathless recounts the misadventures of a petty crook (played by Jean-Paul Belmondo, often Godard's alter ego on screen) who admires Humphrey Bogart and is betrayed to the police by an American girl. Being uncertain whether or not she loves him, she informs on him simply to see if she can. For some years, Godard's work showed an increasingly desperate obsession with themes of fickleness (both male and female), indignity, caprice, and the impossibility of distinguishing a meaningful reality from the imposture perpetrated by others, by one's own mind, by ideology, and by art. Godard used the face of the actress who was then his wife, Anna Karina, as a sphinxlike icon representing this existential duplicity in several films, notably *The Little Soldier* (1960), an ironically flip tragedy, banned for many years, about torture and counter-torture. *My Life to Live* (1962), a study of a young Parisian prostitute, used, with ironical solipsism, pastiches of documentary form and clinical jargon. And on the strength of *Pierrot le fou* (1965), he was asked to direct what was to be an immensely successful U.S. film, *Bonnie and Clyde* (he refused it because of his suspicion of the Hollywood system).

Early themes of fickleness

Godard offered his visual and verbal images as delusive counterfeits for a life whose meaning has become irretrievably lost, or, perhaps was always intrinsically absurd. These images are endowed with additional depth by his extensive culture. Increasingly, his films came to include shots of books brandished or read from and suggestive street signs or posters and dialogue that is delivered as if the performers were alienated from their roles, merely reading texts. Historically impossible or subtly fantastic settings are juxtaposed, and his films compel disrupting awareness of the medium itself. Thus, in two of his films, Belmondo, as the hero, addresses the audience directly, through the camera, while driving. His allusions to other films in themselves constitute an

intricate maze. The heroine of *The Little Soldier*, for instance, is surnamed "Dreyer" after Carl Dreyer, the director whom Godard admires; an extract from one of Dreyer's films is watched by the heroine of *My Life to Live*. *Alphaville* (1965) features scenes from *Metropolis* (1926), whose director, Fritz Lang, plays a film director in *Contempt*. In these ways, Godard's films become intellectual essays: in them, the acted, experienced fictions of earlier motion pictures are transformed into the illustrative ideological cinema of the late 1960s. In 1966 two features—*Made in U.S.A.*, devoted to America, and *Two or Three Things I Know About Her*, devoted to Paris—marked a nadir of Godard's generalized despair, which by then was aimed at society as well as at interpersonal relationships. An increasing interest in left-wing thought was implicit in *La Chinoise* (1967; its title is slang for Parisian Maoists) and was confirmed by Godard's active participation in the Paris student riots of 1968 and other demonstrations.

Revolutionary militancy

By then married to the actress Anne Wiazemsky, he moved from fiction and aesthetic preoccupation to the Marxism of Herbert Marcuse, "Che" Guevara, Frantz Fanon, and others. *Le Gai Savoir* (1968) is a flatly illustrated text spoken by two students named Émile Rousseau and Patricia Lumumba. His subsequent texts have exhibited a complete indifference to their appeal to the public and have been intended as intellectual agitprop (*i.e.*, agitation–propaganda): in Godard's own words, they are "not a show, a struggle." With this ideological twist Godard disconcerted those who had admired him, whether their particular enthusiasm was for the dexterity of his film form, for his reputation as a second creator of the intellectual cinema (as adumbrated by the Soviet director Sergey Eisenstein a generation earlier), for his skill in posing complex cultural riddles, or for his cool but sad recording of Western man's crises of identity. His evolution has posed a problem, also, for his detractors, whether they criticized him for solipsism, for nihilism, or for his suspiciously complacent celebration of the ignominies of bourgeois man under the shadow of revolution. Even the minority that declared his earlier films to be honourable failures, or exercises in intellectual tedium, have had to agree that, of all directors, he remains the most recklessly volatile, and his development the most fascinatingly unpredictable.

MAJOR WORKS

À bout de souffle (1960; *Breathless*); *Le Petit Soldat* (1960; *The Little Soldier*); *Une Femme est une femme* (1961; *A Woman is a Woman*); *Vivre sa vie* (1962; *My Life to Live*); *Les Carabiniers* (1963; *The Soldiers*); *Le Mépris* (1963; *Contempt*); *Bande à part* (1964; U.S. title, *Band of Outsiders;* British title, *The Outsiders*); *Une Femme mariée* (1964; U.S. title, *The Married Woman;* British title, *A Married Woman*); *Alphaville* (1965); *Pierrot le fou* (1965; *Crazy Pete*); *Masculin-féminin* (1965); *Made in U.S.A.* (1966); *Deux ou trois choses que je sais d'elle* (1966); *La Chinoise* (1967); *Loin de Viêt-nam* (1967; *Far from Vietnam,* episode); *Week-end* (1967); *Le Gai Savoir* (1968); *One plus One* (1968; alternate title, *Sympathy for the Devil*); *British Sounds* (1969); *Le Vent d'est* (1969); *Pravda* (1970).

BIBLIOGRAPHY. The English bibliography is dominated by two collections of essays: *The Films of Jean-Luc Godard*, 2nd ed. by IAN CAMERON (1970); and *Jean-Luc Godard: A Critical Anthology*, ed. by TOBY MUSSMAN (1968). Other English material includes RICHARD ROUD *Jean-Luc Godard*, 2nd ed. (1970); and the following scripts: *Le Petit Soldat, The Married Woman, Alphaville, Pierrot le fou, Masculin-féminin*, and *Made in U.S.A.*

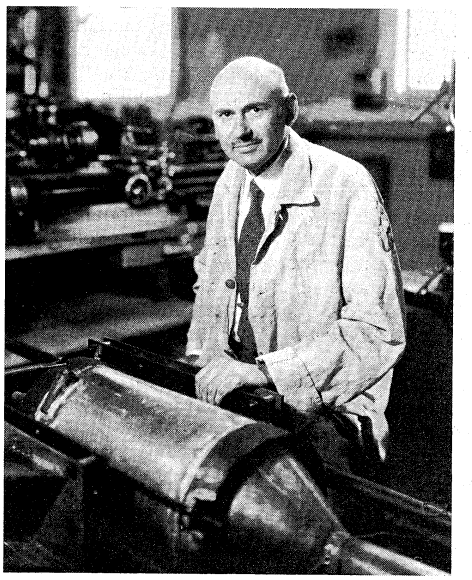
(R.Du.)

Goddard, Robert Hutchings

Robert Hutchings Goddard, a retiring New England professor who dreamed of interplanetary travel at the turn of the 20th century and became America's earliest inventor of rocket engines, is generally acknowledged to be the father of modern rocketry. Among the more than 200 rocket patents issued to Goddard from 1914 to the mid-1940s are methods of propulsion in airless space anticipated by the physicist as a young man. In 1919, the publication of his classical treatise, *A Method of Reaching Extreme Altitudes* (*Smithsonian Miscellaneous Collections*, vol. 71, no. 2) quietly ushered in the age of space flight and man's travel to other planets.

Born October 5, 1882, in Worcester, Massachusetts, Goddard was the only child of a bookkeeper, salesman,

By courtesy of Esther C. Goddard



Goddard in his workshop, 1935.

and machine-shop owner of modest means. The boy had a genteel upbringing and in early youth felt the excitement of the post-Civil War Industrial Revolution when Worcester factories were producing machinery and goods for the burgeoning country. From childhood on he displayed great curiosity about physical phenomena and a bent toward inventiveness. He read in physics and mechanics, and dreamed of great inventions.

In 1898, young Goddard's imagination was fired by the H. G. Wells space-fiction novel, *War of the Worlds*, then serialized in the *Boston Post*. Shortly thereafter, as he recounted, he actually dreamed of constructing a workable space-flight machine. On October 19, 1899, a day that became his "Anniversary Day," he climbed a cherry tree in his backyard and ". . . imagined how wonderful it would be to make some device which had even the *possibility* of ascending to Mars . . . when I descended the tree . . . ," he wrote in his diary, "existence at last seemed very purposive."

Goddard's fascination with space flight and the means of attaining it continued into his college years at the Worcester Polytechnic Institute. In an assigned theme, "Travelling in 1950," he was also intrigued with the notion of "the fastest possible travel for living bodies on the earth's surface" and projected a plan for travel inside a steel vacuum tube in which cars were suspended and driven by the attraction and repulsion of electromagnets. Patents on a vacuum-tube system of transport were later granted the inventor, with thrust—acceleration and deceleration—the chief principle.

In 1908 Goddard began a long association with Clark University, Worcester, where he earned his doctorate, taught physics, and carried out rocket experiments. In his small laboratory there, he was the first to prove that thrust and consequent propulsion can take place in a vacuum, needing no air to push against. He was the first to explore mathematically the ratios of energy and thrust per weight of various fuels, including liquid oxygen and liquid hydrogen. He was also the first to develop a rocket motor using liquid fuels (liquid oxygen and gasoline), as used in the German V-2 rocket weapon 15 years later. In a small structure adjoining his laboratory, a liquid-propelled rocket in a static test in 1925 "operated satisfactorily and lifted its own weight," he wrote. On March 16, 1926, the world's first flight of a liquid-propelled rocket engine took place on his Aunt Effie's farm in Auburn, Massachusetts, achieving a brief lift-off.

World's first rocket flight

In achieving lift-off of his small but sophisticated rocket engine, Goddard carried his experiments further than did the Russian and German space pioneers of the day. As is frequently the case with scientific theory and invention, developments proceeded in various parts of the world. While Goddard was engaged in building models of a space-bound vehicle, he was unaware that an obscure schoolteacher in a remote village of Russia was equally fascinated by the potential for space flight. In 1903 Konstantin E. Tsiolkovsky wrote "Investigations of Space by Means of Rockets," which many years later was hailed by the Soviet Union as the forerunner of space flight. The other member of the pioneer space trio—Hermann Oberth of Germany—published his space-flight treatise, *Die Rakete zu den Planetenräumen*, in 1923, four years after the appearance of Goddard's early monograph.

Goddard's early tests and others were modestly financed over a period of several years by the Smithsonian Institution, whose secretary, Charles G. Abbot, had responded to Goddard's appeal for financial support. In 1929, following an aborted and noisy flight test that brought unwanted press notice to the publicity-shy inventor, Charles A. Lindbergh was instrumental in procuring greater financial assistance for Goddard's experiments. From 1930 to the mid-1940s, the Guggenheim Fund for the Promotion of Aeronautics financed the work on a scale that made possible a small shop and crew and experimental flights in the open spaces of the American southwest, at Roswell, New Mexico. There, Goddard spent most of his remaining days in the unending trial-and-error reach for high altitudes.

Experiments at Roswell

In the course of his experiments there he became the first to shoot a liquid-fuel rocket faster than the speed of sound (1935). He obtained the first patents of a steering apparatus for the rocket machine, and of the use of "step rockets" to gain great altitudes. He also developed the first pumps suitable for rocket fuels, self-cooling rocket motors, and other components of an engine designed to carry man to outer space. Furthermore, his experiments and calculations took place at a time when any news of his work drew from the press and the public high amusement that "Moony" Goddard could take seriously the possibility of travel beyond Earth. His small rockets, early prototypes of the modern Moon thrusters, achieved altitudes of up to one mile above the prairie.

During World War II, Goddard offered his work to the military, but lack of interest in rocket development led to his closing down the Roswell establishment and participating in the war effort through a small Navy contract for work at Annapolis, Maryland, on the development of a jet-thrust booster for seaplane takeoff. Lindbergh and the industrialist and philanthropist Harry F. Guggenheim remained staunch advocates of the Worcester inventor and the feasibility of space exploration.

Goddard died of throat cancer August 10, 1945, at the threshold of the age of jet and rocket. Years later, his work was acknowledged by the United States government when a $1,000,000 settlement was made for the use of his patents. Honours paid posthumously included a Goddard Memorial Library at Clark University.

BIBLIOGRAPHY. The main biographical source is MILTON LEHMAN, *This High Man: The Life of Robert H. Goddard* (1963). See also *The Papers of Robert H. Goddard*, 3 vol. (1970); and R.H. GODDARD, *Rocket Development* (1948), both ed. by ESTHER C. GODDARD and G. EDWARD PENDRAY.

(M.K.L./M.Ln.)

Goebbels, Joseph

Paul Joseph Goebbels, the leading propagandist of the Nazi Party and minister of propaganda of Adolf Hitler's Third Reich, mastered, to the point of near perfection, every contemporary device for leading and misleading the public. It was largely due to his extraordinarily persuasive propaganda that the German people followed Hitler into the abyss, long after they knew where he was leading them.

Early years

Goebbels was born on October 29, 1897, the third of five children of a factory clerk in Rheydt, in the Rhineland. The parents, eager to see their talented son rise in the world, not only provided him with a high school education but also helped support him during the five years of his undergraduate studies. In World War I he was exempted from active wartime duty and all other military service because of his clubfoot, which later enabled his enemies to draw a parallel with the cloven hoof and limp of the Devil. This defect, presumably not congenital but rather the result of a childhood disease, played a disastrous role in his life by engendering strong desires for compensation.



Goebbels, *c.* 1935.
Interfoto-Friedrich Rauch, Munich

After graduating from Heidelberg University in 1921, with a doctorate in German philology, Goebbels engaged in—largely unremunerative—literary, dramatic, and journalistic efforts. Although not yet involved in politics, Goebbels, in common with most of his contemporaries, was imbued with a nationalistic fervour made more intense by the frustrating outcome of the war. During his university days, a friend also introduced him to socialistic and communistic ideas. Antibourgeois from his youth, Goebbels remained so in spite of all his later upper class affectations. On the other hand, he was initially not anti-Semitic. The high school teachers he valued most were Jews, and he was, during that time, engaged to a half-Jewish girl. At that point his options, if he chose to enter politics, were still wide open. An accident determined the party he was to join.

Rise in the party

In the autumn of 1924 he made friends with a group of National Socialists. A gifted speaker, he was soon made the district administrator of the Nationalsozialistische Deutsche Arbeiterpartei (National Socialist German Workers' Party [NSDAP]) in Elberfeld and editor of a biweekly National Socialist magazine. In November 1926 Hitler appointed him district leader in Berlin. The NSDAP, or Nazi Party, had been founded and developed in Bavaria, and, up to that time, there had been practically no party organization in Berlin, the country's capital. Goebbels owed his new appointment to the prudent choice he made in a conflict between Gregor Strasser, representing the "left-wing," anti-capitalist faction of the NSDAP, and the "right-wing" party leader, Hitler. In this conflict, Goebbels, against his own inner convictions, took Hitler's side.

Personally courageous and never shirking danger, Goebbels proceeded to build Nazi strength in Berlin until Hitler's accession to power at the end of January 1933. In 1928 Hitler gave the successful orator, well-versed propagandist, and brilliant journalist (he was editor of *Der Angriff* ["The Assault"] and later, from 1940 to 1945, *Das Reich*) the additional post of propaganda director for the NSDAP for all of Germany. With great diligence Goebbels began to create the Führer myth around the person of Hitler and to institute the ritual of party celebrations and demonstrations that played a decisive role in converting the masses to Nazism. In addition, he

spread propaganda by continuing his rigorous schedule of speechmaking.

Control of propaganda

After the "seizure of power," Goebbels was also able to take control of the national propaganda machinery. A "National Ministry for Public Enlightenment and Propaganda" was created for him, and in addition he became president of the newly formed "Chamber of Culture" for the *Reich*. In these positions he controlled, in addition to actual propaganda, the press, the radio, films, publication, the theatre, music, and the visual arts. In the case of foreign propaganda, the press, publications, and the theatre, he engaged in occasionally bitter jurisdictional struggles with other functionaries; he controlled music and art without too much interest. He did not, however, succeed in extending his power into other areas, such as the high schools.

His cultural policies were relatively liberal, but he often had to capitulate to the demands of nationalist extremists. Even his propaganda messages were limited by the rational argument that ceaseless agitation only dulls the receptive powers of the listener. As far as Goebbels was concerned, efficiency always took precedence over dogmatism, expediency over principles.

Goebbels' influence decreased in the years 1937 and 1938, when he came within a hair's breadth of giving up his career and family because of a deep love affair with a Czechoslovakian film star. (Since 1931 he had been married to a woman of good family who eventually bore him six children.) His position underwent little change with the outbreak of World War II (which he did not welcome): in times of victory, the propagandist's services are not much in demand. Goebbels' hour came with the turn in fortunes of the war after the defeats in Stalingrad and Africa, when he was to prove himself a master of the clever propaganda of holding out in the face of defeat.

Contrary to common belief, Goebbels did not at that time falsify the facts of the situation. On the contrary, the main thrust of his propaganda—which he carried on personally and without respite in the press and over the radio—was to continually raise hopes by citing historical parallels and making other comparisons, by conjuring up allegedly immutable laws of history, or even, as a last resort, by referring to some secret miracle weapons. Here, too, he demonstrated personal courage by appearing constantly before the public long after the other prominent Nazis had retreated to their bunkers and fortifications. His public appearances in these years did much to improve an image that had until then been overwhelmingly negative. Goebbels' work was especially effective in intensifying the efforts of the home front: he became the protagonist of total war. After several false starts, the attempted assassination of Hitler on July 20, 1944, brought him within view of his goal. On July 25 he was "invested by the Reich with full powers to implement total war"—but it was too late.

On May 1, 1945, the only one of the original Nazi leaders to remain with Hitler in the besieged bunker in Berlin, Goebbels, together with his wife, took his own life and those of his six children. This was the last and, if not the bloodiest, at least the most macabre production of this talented stage manager. The day before, he had been named chancellor of the *Reich* in Hitler's will. For one day, on a few square metres, he thus became the last successor to Otto von Bismarck.

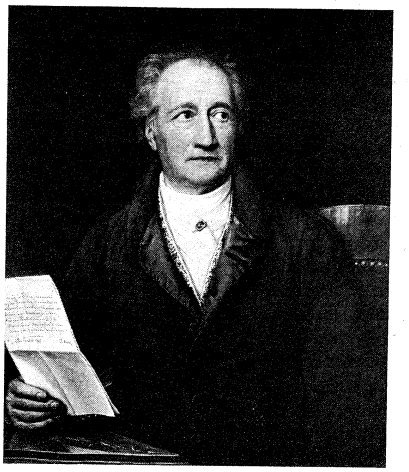
BIBLIOGRAPHY. HELMUT HEIBER, *Joseph Goebbels* (1962), in German; ROGER MANVELL and HEINRICH FRAENKEL, *Doctor Goebbels: His Life and Death* (1960); HELMUT HEIBER (ed.), *Das Tagebuch von Joseph Goebbels, 1925/26* (1960; Eng. trans., *The Early Goebbels Diaries, 1925–1926*, 1962); LOUIS P. LOCHNER, *The Goebbels Diaries* (1965); VIKTOR REIMANN, *Dr. Joseph Goebbels* (1971), in German.

(H.Hei.)

Goethe, Johann Wolfgang von

A German writer, universally acknowledged to be one of the giants of world literature, Goethe was perhaps the last European to attempt the many-sidedness of the great Renaissance personalities: critic, journalist, painter, theatre manager, statesman, educationalist, natural philosopher. The bulk and diversity of his output is in itself phenomenal: his writings on science alone fill about 14 volumes. In the lyric vein he displayed a unique variety of theme and style; in fiction he ranged from fairy tales, which have proved a quarry for psychoanalysts, through the poetic concentration of his shorter novels and *Novellen* to the "open," symbolic form of *Wilhelm Meister;* in the theatre, from historical, political, or psychological plays in prose through blank-verse drama to his *Faust*, one of the masterpieces of modern literature. He achieved in his 82 years a wisdom often termed Olympian, even inhuman; yet almost to the end he retained a willingness to let himself be shaken to his foundations by love or sorrow. He disciplined himself to a routine that might armour him against chaos; yet he never lost the power of producing magical short lyrics in which the mystery of living, loving, and thinking was distilled into sheer transparency.

By courtesy of the Bayerische Staatsgemaldesammlungen, Munich



Goethe, oil painting by Joseph Karl Stieler, 1828. In the Bayerische Staatsgemäldesammlungen, Munich.

And at the last there was granted him a gift, uncanny even to himself, of tapping at will the springs of creativity in order to complete the work he had carried with him for 60 years. When, a few months before his death, he sealed his *Faust*, he bequeathed it with ironic resignation to the critics of posterity to discover its imperfections. Its final couplet, "Das Ewig-Weibliche/Zieht uns hinan" ("Eternal Womanhead/Leads us on high"), epitomizes his own feeling about the central polarity of human existence: woman was to him at once man's energizer and his civilizer, source of creative life and focus of the highest endeavours of both mind and spirit.

There was in Goethe a natural, if not always painless, swing between poles of existence often thought to be mutually exclusive and an innate commitment to change and process. And in the last letter he was to write, he rounded off what has sometimes been called his greatest work, his life, by setting the seal of his approval on a mode of growth that sees the art of living as the intensification of inborn talents through a judicious surrender to the natural rhythm of opposing tendencies.

Early life and influences. Born on August 28, 1749, in Frankfurt am Main, Goethe came of middle class stock, the *Bürgertum* that he never ceased to praise as a breeding ground of the finest culture. His father, Johann Kaspar Goethe, was of north German extraction. A retired lawyer, he was able to lead a life of cultured leisure, travelled in Italy, and amassed a well-stocked library and picture gallery in his handsomely furnished house. His mother, Katharine Elisabeth Textor, daughter of a *Bürgermeister* of Frankfurt, opened up to her son valued connections with the patriciate of the free city. Thus even in his heredity Goethe unites those opposing tendencies that have always prevailed in German lands: the intellectual and moral rigour of the north and the easygoing artistic sensuousness of the south. Of eight children, only Wolfgang, the firstborn, and a sister, Cornelia, survived.

In his autobiography, *Dichtung und Wahrheit* ("Poetry and Truth"), Goethe left an unforgettable picture of a happy childhood. Here are set out with acute psychological insight the emotional complexities of his bond with Cornelia, which found expression in numerous portrayals of the brother-sister relationship in his works; his passionate attachment to a barmaid, Gretchen, which foreshadowed the rejection pattern of many of his loves; the broadening of outlook that came with French occupation during the Seven Years' War; the coronation of Joseph II in the Frankfurt *Römer*, with its indelible impressions of medieval pageantry; the fervent religiosity of Pietistic circles, which led him to declaim F.G. Klopstock's *Messias* as a kind of Lenten exercise, to write a prose epic on Joseph and a poem on Christ's descent into hell. The French army had brought its own troupe of actors, and their performances intensified a passion for the stage first kindled in him by his grandmother's gift of a puppet theatre, and inspired a lifelong devotion to Racine. A love of things English was fostered by friendship with a young clothier from Leeds (Goethe's paternal grandfather was a fashionable tailor) with whom Cornelia, seeing herself as the heroine of a Richardsonian novel, fell hopelessly in love. Wolfgang's reaction was the inception of a novel in letters, a kind of linguistic exercise in which four brothers correspond in different languages.

Study at Leipzig

In October 1765 Goethe was sent to study law at his father's old University of Leipzig, though he himself would have preferred to read classics in the newly founded university at Göttingen, where English influence prevailed. In Leipzig, or "little Paris" as he calls it in *Faust*, by contrast, a world of elegance and fashion made the young provincial feel like a fish out of water. The Frenchifying influence of the critic J.C. Gottsched still dominated the theatre and provided a repertory of the best plays of contemporary Europe. But C.F. Gellert, poet and author of fables and hymns, now in the heyday of his fame, presented the new sensibility of Edward Young, Laurence Sterne, and Samuel Richardson. Goethe praised Gellert's lectures as "the foundation of German moral culture" and learned from them invaluable lessons in epistolary style and in social conduct. Gellert's literary influence was reinforced by the robust elegance and ironic sagacity of the novels, tales, and epics of C.M. Wieland. Wieland's work was brought to Goethe's notice by A.F. Oeser, friend and teacher of the archaeologist and art historian J.J. Winckelmann, who profoundly influenced European fashions in art. From Oeser, Goethe learned a love of Greek art and two things that stood him in good stead all his life: to use his eyes and to master the craft of whatever he undertook. A visit to Dresden, "the Florence of the north," as the poet and critic J.G. Herder called it, opened his eyes to the splendours of Rococo architecture as well as classical statuary. Nor was music neglected in his education; a new 18th-century concert society, under the direction of the musician and composer J.A. Hiller, provided splendid performances, which became world famous as the *Gewandhaus* concerts.

Early works in the Rococo fashion

The literary harvest of Goethe's Leipzig period manifested itself in a songbook written in the prevailing Rococo mode—songs praising love and wine in the manner of the Greek poet Anacreon. Appropriately titled *Das Leipziger Liederbuch* (*The Leipzig Song Book*), it was ostensibly inspired by the daughter of the wine merchant at whose tavern he took his midday meal. But neither his 1766–67 poems *Das Buch Annette* (as he called her in Rococo fashion) nor the *Neue Lieder* of 1769 made any pretense of real passion. Yet it was in connection with these literary trifles that he subsequently made the famous and much abused statement that all his works were "fragments of a great confession." The same note is struck in two plays written in alexandrine verse (a twelve-syllable iambic line borrowed from the French), *Die Laune des Verliebten* ("The Mood of the Beloved") and a more sombre farce, *Die Mitschuldigen* ("The Accomplices"), which foreshadows the psychological preoccupations of later works. From then on, Rococo was one element in Goethe's repertoire to be drawn on as occasion demanded. It was to reappear in the setting of *Torquato Tasso* and *Die Wahlverwandtschaften* ("Kindred by Choice"); he was to pay tribute to its charm in *Anakreons Grab* ("Anacreon's Grave"; 1806) and amalgamate it with Eastern influence in enchanting poems of the *West-östlicher Divan* ("The West-East Divan").

Works of the storm and stress period. Goethe's stay in Leipzig was cut short by severe illness, and by the autumn of 1768 he was back home. A long convalescence fostered introspection and religious mysticism. He played with alchemy, astrology, and occult philosophy, all of which left their mark on *Faust*. On his recovery it was decided that he should pursue legal studies in Strassburg as a first stage on the way to Paris and the Grand Tour (never actually completed). His stay there proved a turning point for his whole life and work. In this German capital of a French province, he experienced a reaction against the cosmopolitan atmosphere of Leipzig and under the impact of the great cathedral proclaimed his conversion to the Gothic German ideal. More decisive still was the influence of J.G. Herder, who spent the winter of 1770–71 there undergoing treatment for his eyes. From him Goethe learned the role played by touch, the haptic sense, in the growth of the mind; a new view of the artist as a creator fashioning forms expressive of feeling; a new theory of poetry as the original and most vital language of man; the virtues of a new style, that of the *Volkslied* (folk song) and the poetry of "primitive" peoples as enshrined in the Bible, the epics of Homer, and the poems attributed (probably falsely) to Ossian, a 3rd-century Celtic poet. It is this new sense of felt immediacy, and of the plasticity of his linguistic medium, that informs the lyrics Goethe wrote to one of his early loves, Friederike Brion, the pastor's daughter of Sesenheim. They mark the beginning of a new epoch in the German lyric. Such poems as "Mailied" and "Willkommen und Abschied" are still the most popular, though not the greatest, of his *Lieder*. The latter, especially in its revised form of 1790, touchingly expresses the guilt he felt that this time he himself had the role of deserter and rejecter, and the whole idyl as recounted in *Dichtung und Wahrheit* reveals that cross-fertilization of life and literature that he increasingly saw as a potent factor in human development.

Influence of J.G. Herder

If, as Herder maintained, energy was one of the marks of poetry, it was clearly in the passions acted out on the stage that it could find its most vital expression. And where more vital than in the colossal figures of the "Gothic Shakespeare"? In writing the *Geschichte Gottfriedens von Berlichingen mit der eisernen Hand dramatisiert* ("The History of Gottfried von Berlichingen Dramatized with the Iron Hand"; 1771), Goethe was deliberately vying with Shakespeare. For the real Götz, who died two years before Shakespeare was born, was near enough in time to represent that bustling spacious 16th century, the animal vitality of which contrasted so forcibly with the straitlaced affectations of Goethe's own day. With the publication in 1773 of *Götz von Berlichingen*, a radically tautened version of that "History," the Shakespeare cult was launched and the Sturm und Drang provided with its first major work of genius. The manifesto of the movement, heralded by Goethe's enthusiastic *Rede zum Schakespears Tag* ("Conversation from Shakespeare's Day"), had appeared after Goethe's return to Frankfurt in August 1771. "Von deutscher Art und Kunst" ("Concerning German Nature and Art"), as it was called, contained a defense of German nationality by the historian J.M. Möser, two essays by Herder championing Ossian and Shakespeare, and a rhapsody on Gothic architecture by Goethe.

The success of *Götz von Berlichingen*

Though ostensibly in practice as a lawyer, the young poet now found himself caught up in a whirl of literary and social duties—helping to edit the *Frankfurter Gelehrte Anzeigen* ("Frankfurt Scholarly Reviews"), for instance—and it was to break loose from this that he left for Wetzlar, seat of the supreme court of the Empire. But again literature won the day over law, and an impassioned yet self-ironic ode in free verse, "Wandrers Sturmlied," is testimony both to a recently inspired admiration for Pindar, the greatest lyric poet of ancient

Greece, and to a hesitant certainty that he himself might be destined for greatness. And in Wetzlar he experienced a new passion, this time for a girl safely out of reach from the start, Charlotte Buff. Her betrothed, Johann Christian Kestner, showed great understanding until, as it seemed to him, he found the affair exposed to public gaze in *Die Leiden des jungen Werthers* (*The Sorrows of Young Werther;* 1774).

But much besides the Wetzlar experience had gone into the making of this novel: Herder's scathing comments on his young pupil's lack of formal- and self-mastery; the recent indictment by G.E. Lessing of the Neoplatonic doctrine of artistic creation in *Emilia Galotti;* a passing attraction to Maximiliane, the daughter of the German novelist Sophie von La Roche, who probably endowed his heroine with her black eyes. And it was only when Kestner reported the suicide of a Wetzlar acquaintance who had killed himself out of hopeless love that all this was precipitated into a plot. If *Werther* took the world by storm it was because, in Thomas Carlyle's words, it gave expression to "the nameless unrest and longing discontent which was then agitating every bosom." But this first novel is no sentimental tearjerker. Nor is disappointed love its real theme. It is rather what the 18th century called Enthusiasm: the fatal effects of a predilection for absolutes, whether in love, art, society, or the realm of thought. The mind that conceived its symmetry, wove its intricate linguistic patterns, and handled the subtle differentiation of hero and narrator was moved by a formal as well as a personal passion. Even the title has been trivialized in translation: *Sorrows* (instead of "Sufferings") obscures the allusion to the Passion of Christ and individualizes what Goethe himself thought of as a "general confession," in a tradition going back to St. Augustine.

Lyrical and dramatic hymns

Besides *Werther* and *Götz*, the period 1771–75 saw the appearance of a number of magnificent hymns—lyrical or dramatic, according to whether the influence of Pindar or Shakespeare prevailed—"Cäsar," "Mahomets Gesang," "Der Ewige Jude," "Prometheus," "Sokrates," "Satyros," "Der Wandrer"; the inception of *Egmont* and *Faust* (this so-called *Urfaust*, or "original" version of *Faust* was discovered by a lucky chance in 1887); the completion of *Clavigo*, a play of more "regular" form on a theme of the French playwright Beaumarchais, and of *Stella* (1775), with its conciliatory ending of a *mariage à trois*, subsequently conventionalized into tragedy. Two operettas, *Erwin und Elmire* and *Claudine von Villa Bella*, reflect a return to the elegance of Rococo inspired by Goethe's betrothal to Lili Schönemann, daughter of a rich banker, who moved in fashionable circles that were soon to prove unbearably restrictive to the young Stürmer und Dränger. From the conflicts of this love he took refuge, as so often, in nature; and in a poem written on the lake of Zurich, "Auf dem See," created the first of those many short lyrics in which language of radiant simplicity is made the vehicle of inexhaustible significance. With his departure for Weimar in November 1775, the engagement was allowed to lapse.

The mature years at Weimar. Going to Weimar was the major turning point of Goethe's life. He went on a visit to the reigning duke, Charles Augustus. It remained his home—despite Napoleon's invitation to Paris—until his death there on March 22, 1832. From now on, mastery of life became his chief concern; and *Wilhelm Meisters Lehrjahre* (*Wilhelm Meister's Apprenticeship;* 1824), the title he eventually gave his next novel (1795–96), suggests the long apprenticeship such mastery involves. He served his own in the innumerable and ever increasing official duties the young duke heaped on his willing shoulders until, as indispensable minister of the little state, he was inspecting mines, superintending irrigation schemes, and even organizing the issue of uniforms to its tiny army.

Friendship with Charlotte von Stein

He served his apprenticeship, too, in his passionate devotion to the wife of a court official, Charlotte von Stein. For the first time he found himself in love with a woman who could also meet him on the intellectual plane. From the 1,500 or so letters he wrote her we can see her become the guiding principle of his life, teaching him the graces of society, dominating the details of his daily existence, engaging his imagination and desire, yet insisting on a relation governed by decorum and conventional virtue. She would be his sister and nothing more, and the sublimation she increasingly enforced on him, though irksome, could inspire the almost psychoanalytical probings of "Warum gabst du uns die tiefen Blicke?", the tortures of Orest and their assuagement by Iphigenie, the delicate one-act play, *Die Geschwister* ("Brother and Sister"; 1776), and such well-loved lyrics as "An den Mond," "Der Becher," "Jägers Abendlied," "Seefahrt," and the two exquisite "Wandrers Nachtlieder."

In these and other poems of this period—"Grenzen der Menschheit," "Gesang der Geister über den Wassern," "Das Göttliche," "Harzreise im Winter," "Ilmenau"—nature has ceased to be a mere reflection of man's moods and has become something existing in its own right, a setting for an idea or a force indifferent, even hostile to him. This new "objectivity" is in tune with Goethe's growing scientific preoccupations. Yet such is his versatility that he could, when he chose, revert to the temper of "Der König in Thule" (written in 1774) and compose ballads such as "Erlkönig" or "Der Fischer," in which nature bears the projection of unconscious forces; while a number of *Singspiele* or musical plays betoken his readiness and ability to provide light entertainment for the court. *Der Triumph der Empfindsamkeit* ("The Triumph of Sensibility") even satirizes the sensibility his own *Werther* had helped to foster.

But neither the cares of state nor those of a frustrating love affair were conducive to the peace and leisure required to complete works of such magnitude as *Egmont*, *Faust*, *Tasso*, and *Iphigenie* (a prose version of this last was sufficiently advanced to be put on before the court in 1779 with Goethe himself in the role of Orestes). And in September 1786, in dramatic secrecy and with the haste of one pursued, he set out on his long-postponed Italian journey. This flight was at once a death and a rebirth. And it was in these terms that he wrote of it in his letters. He sought the renewal of himself, both as man and artist, and so deliberately cut himself off from his emotional, literary, and cultural past, scorning the "Gothic follies" he had once acclaimed, rejecting Juliet's tomb in Verona in favour of the Greek steles in the museum, finding delight in Palladio's churches rather than in San Marco or the doge's palace, devoting barely three hours to Florence, and ignoring completely the medieval glories of Assisi for the sake of its temple of Minerva, feverishly bent on arriving in Rome, "capital of the ancient world," but seeing even that as a prelude to Magna Graecia, to the temples of Paestum, and the revelation of classical grandeur in Sicily, "key to the whole," a prelude to the world of Homer, which he recaptured in a glorious dramatic fragment, *Nausikaa* (1787). And just as he sought and found the *Urmensch*, or archetypal man, in the forms of Greek antiquity, so in these landscapes there came to his mind the extension of this idea to plants as well. In his literary work these pursuits led to the creation of beings who are individual manifestations but of a clearly discernible type; to themes that are universal and timeless but treated in a highly differentiated way; to the measured cadences of verse that are yet vibrant with personal passion.

Italian influence

This new conception of form is apparent in the revision of the four plays he had taken with him to Italy. *Faust, Ein Fragment*, published in 1790, is quite clearly, by its excisions as well as its additions, a step in the direction of the stupendous cultural symbol the play would eventually become rather than any attempt to weld into dramatic unity the sharply individualized episodes of the original version, the *Urfaust*. *Egmont*, though not actually cast into verse, is raised to the level of poetic drama not by virtue of its frequent iambic rhythms but by a thickening of the verbal texture, so that when music finally takes over it seems the inevitable culmination of a gradual convergence and sudden contraction of themes rather than the "*salto mortale* (*i.e.*, somersault) into the world of opera" Schiller was to dub it. By such means,

the personal and the political aspects of the problem become completely interfused—Egmont and his beloved Klärchen, the most lovable characters Goethe ever created, are embodiments of an inner freedom that is a heightened form of the easygoing independence of the Netherlands people—and what had started as a dramatic portrayal of a daemonic individual is transformed into a tragedy of the very idea of freedom, of its fate in a world ruled not just by calculation or intrigue but by unpredictable conjunctures of persons and events.

Torquato Tasso

In *Torquato Tasso* such linguistic density is carried to lengths possible only in verse. Goethe spoke of having expended a positively "unlawful care" on it. But this is not inappropriate to a play about a poet, an artist whose medium is the ordinary vehicle of communication between men. The tragic conflict here arises from misunderstandings about the various modes of language, and the temperamental clashes are presented as concomitants of this rather than as the prime focus of interest (though there is enough psychology to justify the description by the French writer Mme de Staël of Goethe as "*le Racine de l'Allemagne*"). The slightness of the outward action in *Torquato Tasso* has been much criticized, but it can be justified in a study of the "poetical character" per se—a creature for whom "any little vexation grows in five minutes into a theme for Sophocles." By placing him in a society that, far from being indifferent or hostile, cherishes him and values his work, Goethe has thrown into sharpest relief the incurable "discrepancy" between poet and world, and this rift is not healed by Tasso's discovery that even the extremes of anguish can be transmuted into imperishable verse.

Iphigenie

But it was perhaps *Iphigenie auf Tauris* (1787) that benefited most from his encounter with classical antiquity. And yet Schiller was right in calling it "astonishingly modern and un-Greek." Like *Tasso*, it too treats of the problems of communication: of the unforeseeable power of words once they are released into the world; of the double face of language, which conceals as much as it reveals; of truth, whose opposite is not just an outright lie but the withholding of self. But it treats, too, of man's power to free himself from his myths by recognizing them as projections of his own unconscious, of his power to break the chain of events that seems to determine his present (symbolized in the monotonously regular crime sequence of the race of Tantalus) by a reorientation of outlook. The conciliatory ending, which Euripides contrived by the sudden appearance of the goddess Athena, here comes with the apparent suddenness of new insight: the words of the oracle are susceptible to a different interpretation. In its synthesis of Greek and Christian values, its elevation of the physical to the spiritual through the identification of Iphigenie with the divine sister, Diana, this play represents the highest achievement of 18th-century humanism.

Römische Elegien

The chief lyrical product of the Italian journey was the *Römische Elegien* (the "Roman Elegies"; written 1788–89). In their plastic beauty and unabashed sensuality, their blending of erotic tenderness with an enhanced sense of our cultural heritage, these pagan, highly civilized poems are unique in any modern language. Had they been written in the metre of Byron's *Don Juan*, Goethe acknowledged, they might easily have been offensive; but the classical distichs (couplets) lend them that veil of aesthetic distance that reveals even as it shrouds. The true begetter of these elegies was not some passing Roman amour but Christiane Vulpius, daughter of a humble official, whom Goethe had taken into heart and home soon after his return from Italy in April 1788. Christiane bore him several children; but it was not until 1806, when life and property were threatened by the French invasion, that the nonconformist eventually conformed and in grateful recognition of its indissoluble bonds regularized their union in the eyes of society.

His first Italian journey finally brought home to Goethe that, for all his interest and talent, he was not destined to be a painter. Despite diligent practice with his artist friends in Rome, he was never able to master this medium to the point at which it became expressive of his deepest feeling, and with rare exceptions his numerous drawings have no more than the charm of a sensitive amateur. But his abiding preoccupation with the visual arts left an indelible mark on his literary as well as his scientific work and gave added precision to his many critical and aesthetic essays. And it was on this first visit to Italy, too, that he finally reached the decision that he must shed his administrative duties and devote himself henceforth to his true vocation of literature and science.

Political writings during the French Revolution

A return visit to Italy in 1790 brought nothing but disappointment, and a restlessness aggravated by the revolutionary events in the outer world. The *Venetian Epigrams* of 1790 reflect something of this discontent. In 1792 Goethe accompanied his duke on the disastrous campaign into France, was present at the battle of Valmy, and wrote up his experiences in two still very readable war books, *Campagne in Frankreich 1792* and *Belagerung von Mainz*. His liberal-conservative attitudes found expression in *Reineke Fuchs* ("Reynard the Fox"), a recasting of the Low German satire, the *Unterhaltungen deutscher Ausgewanderten*, and three plays. *Der Gross-Cophta*, *Die Aufgeregten* and *Der Bürgergeneral*, which, though artistically unsuccessful, are of interest in being among the few examples of political literature produced by German poets. But it was only as the French Revolution receded that he was able to transmute its overwhelming actuality into timeless poetry. It still forms the background of his Homeric treatment of the refugee problem, *Hermann und Dorothea* (1797). It fills the whole canvas of *Die Natürliche Tochter* ("The Natural Daughter"; 1804). Planned as a trilogy but never completed, this was Goethe's final reckoning with the greatest event of his time. Beneath the coolness of its formal perfection there stirs a profound concern with revolutionary phenomena, with the role of death and destruction in the perpetuation of social and cultural, no less than of natural, forms of life.

Schiller and the classical ideal. The human and spiritual isolation in which Goethe found himself on his return from Italy was unexpectedly relieved by the development of a friendship with Schiller. His acceptance of a formal invitation to contribute to a new journal, *Die Horen* (1795–97), called forth Schiller's now-famous letter of Aug. 23, 1794, in which, with marvelous insight, he summed up Goethe's whole existence. Here, it seemed to him, was the very embodiment of the naïve poet—but consciously naïve, moving from feeling to reflection and then transforming reflection back into feeling, concepts of the mind back into percepts of the senses. It was this conscious assent to a mode of thinking different from Schiller's own more abstractive reflection that made possible their immensely fruitful partnership, and the four volumes of their daily correspondence offer not only an invaluable commentary on the ideals and achievements of the greatest period of German literature but astonishing insight into the processes of artistic creation. Some of the works Goethe produced during the next few years are embodiments of their classical ideal. *Hermann und Dorothea*, one of the best loved, is his attempt to "produce a Greece from within." In it he claimed to have "separated the purely human from the dross." The characters are types—except for the hero and heroine, they have no proper names, and even theirs are symbolic—and like those of the *Odyssey* they vindicate peace and home and the domestic virtues. Yet, as always in Goethe's works, these are shown as never secure for long, as constantly in need of being fostered by man's efforts to be human and humane. In the Helena act of *Faust, Part II*, in which the meeting and mating of Faust and Helen of Troy marks the synthesis of paganism and Christianity, of Greece and Germany, he captured the Greek spirit so successfully that competent critics hold that if translated into Attic Greek it might well pass for a lost fragment of the Athenian stage.

A never completed epic, *Achilleis*, is his last attempt to "be a Greek after his own fashion." Other works of this period are in tune with Schiller's growing conviction that the only future for literature in a world that increasingly clamoured for the naturalistic and the tendentious lay in

a hermetic closing of the poetic world by a frank introduction of symbolic devices. *Wilhelm Meisters Theatralische Sendung* ("Wilhelm Meister's Theatrical Mission"; a manuscript of this version turned up in 1910) is now widened to a vocation for life, a theme dear to the heart of Schiller, who had himself just completed a treatise *On the Aesthetic Education of Man* (1795) and wholly in tune with their joint conviction that art, though not the handmaid of either truth or morality, has nevertheless its own peculiar part to play in making better men and better citizens. Fictional realism is now blended with abstraction; characterization, however psychologically acute, subordinated to an overall poetic significance; and the presence in a novel of contemporary society of such mysteriously compelling figures as the Harper and Mignon seems to justify Goethe's claim that his novel is "thoroughly symbolic."

Schiller's influence on *Faust*

It was Schiller, too, who turned his thoughts to the continuation of *Faust* and discerned the difficulties involved in reconciling this "barbarous composition" with their classical ideal, in blending the evident seriousness of its "idea" with that element of "play" that was the prerequisite of the art of the future. By his insistence on such problems, he inspired the fictional framework of *Faust's* "Prelude on the Stage" no less than the philosophical framework of the "Prologue in Heaven." If in spite of such indications, the world insisted on reading *Faust, Part I* (1808) as a love story, which stamped its author as a Romantic; it was because at this stage the almost unbearable pathos of the Gretchen tragedy had not yet found its place in the wider tragedy of Western man.

Goethe and Schiller blamed the failure of the journals in which they strove to propagate their ideals of art and literature (Goethe's *Propyläen*, 1798–1800, was a quasi-successor to Schiller's *Horen*) on the indifference of an uncultivated public and vented their disappointment in *Xenien*, about 400 mordant distichs in the manner of Martial. A more positive reply to their detractors was a wonderful harvest of ballads. Goethe's own—"Der Schatzgräber," "Die Braut von Korinth," "Die Zauberlehrling"—differ from his earlier ones in that man rather than nature now holds sway. The "white" magic of reflection is consciously, even ironically, introduced. And in the ballad, with its blend of lyric, epic, and dramatic elements, Goethe now discerned the *Urei*, or archetypal form, of poetry by analogy with the *Urpflanze* he had discovered in the vegetable world.

Goethe's relation to the Romantics. With Schiller's death in 1805, Goethe felt he had lost "the half of his existence" and he wrote a magnificent tribute to his great friend in *Epilog zu Schillers Glocke*. His intellectual loneliness was eased in some measure by his relations to the new school of Romantics then flourishing in Jena. For they had much in common. Friedrich von Schlegel had begun his career with a book extolling Greek culture and gone on to praise the Orient as the summit of Romantic thought and poetry. His brother Wilhelm's absorption in form and metre was after Goethe's own heart, and he could not be indifferent to their enthusiastic praise of *Wilhelm Meister* or to Novalis' description of him as "the viceregent of poetry upon earth." In Bettina Brentano, daughter of his old love, Maximiliane von La Roche, he found an ardent response to both his genius and his humanity, and her *Briefwechsel Goethes mit einem Kinde* (1835) remains one of the most readable books in German literature, whatever doubts maybe cast on its reliability. Though Goethe decried the Romantics as "forced talents," amateurishly oblivious of the virtues of form, though he deplored their catholicizing tendencies, their uncritical addiction to all things medieval, their attempts to blur the literary genres and confuse the boundaries between art and life, he yet remained open to many of their enthusiasms, even letting himself be moved to a renewed interest in Gothic architecture. And in *Die Wahlverwandtschaften* ("Elective Affinities"; 1809) he drew heavily for his thematic material upon their preoccupation with "the night-side of nature," with the animal, magnetic affinities that attract human beings to each other, as elements are in the chemical world.

Die Wahlverwandtschaften

But this novel offers no support at all for a superstitious surrender to forces natural or supernatural, for a subhuman abdication of moral responsibility. Catastrophe follows inexorably upon the arbitrary interpretation of signs and portents; the heroine enters upon a path of renunciation that brings her near sainthood; marriage may be presented with ruthless realism as "a synthesis of impossibilities," but it remains nevertheless "the beginning and end of all civilization." The Romantics were here taught a lesson of social behaviour—and of artistic form. The narrative is conducted with a serene impartiality, and all the classical values of plasticity, restraint, and symmetry are brought to bear on a subject that is sensational to the point of improbability.

By their translations—Romanticism *is* translation, Clemens Brentano declared—the Romantics were opening up the literary treasures of the world, and *Weltliteratur* was to become one of Goethe's most treasured concepts. Its aim was, as he put it, to advance civilization by encouraging mutual understanding and respect—whether through translation or criticism (his own attempts to interpret Serbian poetry to the Germans is an excellent example of this latter) or through the blending of different literary traditions. Two great ballads, "Der Gott und die Bajadere" and "Paria," and two exquisite cycles, the late and lesser known *Chinesisch-Deutsche Jahres- und Tageszeiten* ("Chinese-German Hours and Seasons"; 1830) and the *West-östlicher Divan* ("Divan of West and East"; 1819), are his own outstanding attempts to marry East with West. This latter is a book of love in all its aspects—tender, playful, sensuous, ironic, wise, and wanton—all of it irradiated by that quality of *Geist*—of intellect, spirit, wit—which he discerned as "the predominant passion" of Persian poetry. His living muse this time, Marianne, the young wife of his friend von Willemer, was perhaps the most completely satisfying of all his loves, so attuned to him in spirit that she could even take a hand in the creation of some of these poems.

The last decade. But the world vision of the aging poet did not only find expression in a silent communing with the past. In his last years, Goethe found himself a world figure, and little Weimar became a Mecca that drew a constant stream of pilgrims from both the Old World and the New. Reports of his stiffness and reserve in the face of almost daily invasions are far outweighed by the testimony of those to whom he showed warmth, understanding, an insatiable curiosity about what was going on in the outside world, and an abiding openness to the present and the future. This is nowhere more apparent than in *Wilhelm Meisters Wanderjahre* (1821–29; "*Wilhelm Meister's Travels*"), with its commitment to social and technological progress (what he would most like to see before he died, Goethe once said, was the completion of the Panama and Suez canals), to a type of education better adapted to modern specialization than the old humanistic studies, to a world no longer centred wholly in Europe—a major "complication" of his plot is a resettlement plan for emigrants in the land of the future ("Amerika, du hast es besser!" ["America, you are better!"]). *Wilhelm Meister* points the truth that mastery of life is not conferred at the end of the "apprentice years" and henceforth an inalienable possession, but a ceaseless wandering in which the goal turns out to be the way, and the way the goal.

Wilhelm Meister

At first sight the subtitle, *Die Entsagenden* ("Renunciation"), seems curiously at odds with such purposeful unrest. But renunciation for Goethe implies no passive resignation to the status quo. It is a growing acceptance of the limits imposed by life itself, limits arising from the nature of space and time and from the conflict of interests and potentialities. The apparent formlessness of the novel reflects the duality of its title. It meanders, its narrative interspersed with tales, anecdotes, episodes and maxims, having but the loosest connection with the plot but a formal, if often subterranean, connection with the poetic significance. These interpolations, like the increasingly symbolic characters, display the whole spectrum of human modes of renunciation. The "whole man" is here represented not by any single individual but by a constel-

lation of many, and the informing principle is the spatial one of configuration rather than the temporal one of succession.

Faust

Faust, too, is often decried as formless, though the climate of criticism is now more propitious to the discovery of its "law." The array of lyric, epic, dramatic, operatic, and balletic elements, of almost every known metre, from doggerel through terza rima (an Italian form of iambic verse consisting of stanzas of three lines) to six-foot trimeter (a line of verse consisting of three measures) of styles ranging from Greek tragedy through medieval mystery, baroque allegory, Renaissance masque, commedia dell'arte, and the "temerities of the English stage," to something akin to the modern revue, all suggest a deliberate attempt to make these various forms a vehicle of cultural comment rather than any failure to create a coherent form of his own. And the content with which Goethe invests his forms bears this out. He draws on an immense variety of cultural material—theological, mythological, philosophical, political, economic, scientific, aesthetic, musical, literary—for the more realistic Part I no less than for the more symbolic Part II (first published posthumously in 1832): if Faust's wooing of Helena in the "Classic-Romantic Phantasmagoria" (as the first publication of the scene in 1827 called it) is accomplished by teaching her the unfamiliar delights of rhymed verse, his seduction of Gretchen is firmly set in the long tradition of erotic mysticism going back to the Song of Solomon. The Faust myth is here made the medium of a profoundly serious but highly ironic commentary on our cultural heritage, presented not as historical pageant—Faust's "progress" from his 18th–16th-century beginnings back through the Middle Ages and classical antiquity to the origins of life, and beyond that to the "Mothers," timeless source of all forms of being, annuls the historical time sequence—but as a drama of the diverse potentialities that coexist in Western civilization.

This Faust, unlike his creator, is the very type of Western man, with two souls warring within his breast and a restlessly inquiring spirit. To the 19th century his ceaseless striving seemed a good thing in itself. To a generation shocked into doubts about progress and the value of action, the disastrous consequences of his attempts to experience "the weal and woe of all mankind" (the *libido sciendi* of Marlowe's Faustus is here but briefly indulged and as swiftly transcended) loom larger than the quotable "message" of any of the speeches, and his ultimate "salvation" becomes correspondingly suspect. Yet the love that bears his mortal remains to "higher spheres" does not mitigate the ironic defeat of his highest mortal endeavour. If the seal of approval is set on a spirit that has eluded Mephisto's every effort to lull him into sloth, the evil into which it led him is not condoned. It needs the combined intercession of human wisdom and human suffering, human innocence and human experience, before compassionate verdict is passed on the erring and straying of this soul "in ferment." Indeed, none of Goethe's conciliatory endings, except that of *Iphigenie*, really removes the sting of tragedy. Critics have tended to excuse or deplore them by reference to his own *konziliante Natur* (his "conciliatory nature"). But at least as relevant is his preoccupation with the form of Greek trilogies and tetralogies and his unorthodox interpretation of Aristotle's catharsis as an effect only likely to be produced in the spectator if there is a corresponding element of "reconciliation" in the structure of the play itself. The apotheosis of the hero, whether Faust's, Egmont's or Ottilie's in the *Wahlverwandtschaften*, is always set in a context reminiscent of a theophany and of the ritual origins of tragedy.

Interest in music

Nor can his interest in the cathartic effect of music be ignored. Unlike the German Romantic poet Novalis, for whom music was "the key to the universe," Goethe was profoundly aware of its dual nature and as suspicious as Plato of its orgiastic power. As in every art he looked for the taming of the Dionysiac by the Apolline, nowhere more movingly symbolized than by the taming of the lion through the piping of the little child in his *Novelle* of 1828, a theme he had already discussed with Schiller as far back as 1797. And increasingly he turned to music for assuagement of his own suffering. His *Trilogie der Leidenschaft* ("Trilogy of Passion"; 1823–27) is at once the lyrical precipitate of an old man's anguished love for a girl of 18 and a tribute to the cathartic effect of this "heavenly art," which restores to life even as it soothes. His *Zauberflöte, Zweiter Teil* is a tribute to his favourite Mozart's *Magic Flute:* Mozart would, he thought, have been the ideal composer for *Faust*. And one of the comforts of his later years was an intimate friendship with the composer K.F. Zelter, whose most brilliant pupil, the young Mendelssohn, afforded him hours of musical delight and deepened his musical understanding—though he never succeeded in reconciling him to the daemonic aspects of Beethoven's music.

Scientific work

By common consent, *Faust* is one of the supreme, if as yet unclassified, achievements of literature. But there were moments when Goethe rated his scientific work higher than all his poetry. His predilection for his *Farbenlehre* ("Theory of Colour"; 1805–10) has something of the love of a parent for a problem child, and nothing is easier than for the physicist to pick holes in his systematic attempt to prove Newton wrong, or for the psychologist to find the cause of his stubbornness in his sense of mathematical inadequacy or in his neurotic attachment to the doctrine that light is one and indivisible and never to be explained by any theory of particles. On the other hand, the usefulness of the Psycho-Physiological Section, together with his study, *Entoptische Farben* ("On Entoptic Images"), is generally acknowledged, while the Historical Section is something of a pioneer work in the writing of the history of science. His work in botany and biology is less controversial. His *Metamorphose der Pflanzen* ("Attempt to Explain the Metamorphosis of Plants"; 1790) is a model of presentation, and the drawings in it are a botanist's delight. His main thesis, that all the parts of the plant are modifications of a type-leaf, has met with a measure of acceptance, though his categorical neglect of the root is regarded as an unscientific exclusion of a possible area of relevance. His hypothesis of a type-plant, by contrast, commands no interest among orthodox botanists today. His discovery in 1784, arrived at independently even if he was not the first to make it, of a recognizable os intermaxillare (the premaxilla of modern anatomists) in the human species was yet another result of his sustained quest for unity and continuity in nature and caused Darwin to hail him as a forerunner.

But what makes for the continuing interest of Goethe's science is not his discoveries: he could not always claim priority for them at the time, nor was he in the least interested to do so. It is his insight into his methods of arriving at them. Few have been as aware of the mental processes involved in the study of natural phenomena; few have been more alive to the hazards that beset the scientist, at every level, from sheer observation to the construction of a theory; and few have been more conscious of the unwitting theorizing involved in even the simplest act of perception. And no one has argued more convincingly that the only way of coping with this inescapable involvement of the observer in the phenomena to be observed is to let "knowledge of self" develop with "knowledge of world."

Such scrupulous awareness of his own mental operations was, of course, of paramount importance in morphology, the science Goethe founded and named. Morphology, as he understood it, was the systematic study of formation and transformation—whether of rocks, clouds, colours, plants, animals, or the cultural phenomena of human society—as these present themselves to sentient experience. He did not propose it as a substitute for the quantitative sciences, which break down forms as we know them and by converting them into mathematical terms ensure a measure of prediction and control. He was not, contrary to common belief, opposed to analysis —one of his favourite maxims was that analysis and synthesis must alternate as naturally as breathing in and breathing out—and his only objection to physics was its increasing tendency to claim monopoly of understanding.

What he was aiming at was rather a humanizing supplement, an understanding of nature in all its qualitative manifestations; and one of his most impassioned pleas is for a concert of all the sciences, a cooperation of all types of method and mind.

This impulse, to find a scientific as well as an aesthetic corrective to the inevitably esoteric tendencies of specialization, is nowhere more apparent than in his two elegies on plant and animal metamorphosis in which he tries to present to imagination and feeling what has been understood by the mind. They eventually took their place in a cycle of philosophical poems entitled *Gott und Welt* ("God and World"). Though no orthodox believer, Goethe was by no means the pure pagan the 19th century liked to imagine. Spinoza's pantheism certainly struck a sympathetic chord, for the deist idea of a God who, having created the world, then left it to revolve, was repugnant to him. But he was and remained a grateful heir of the Christian tradition—*bibelfest*, rooted in the Bible—as his language constantly proclaims. And it was from this centre that he extended sympathetic understanding to all other religions, seeking their common ground without destroying their individual excellences, seeing them as different manifestations of an *Ur*, or archetypal, religion and thus giving expression, in this field as elsewhere, to the essentially morphological temper of his mind. "Panentheism" has been proposed as a more exact term for his belief in a divinity at once immanent and transcendent, and he rebuked those who tried to confine him to one mode of thought by saying that as poet he was polytheist, as scientist pantheist, and that when, as a moral being, he had need of a personal God, "that too had been taken care of." This was one of the meanings he attached to the biblical text: "In my father's house are many mansions."

Religious views

Appraisal. A day will come, Carlyle predicted in a letter to Ralph Waldo Emerson, when "you will find that this sunny-looking courtly Goethe held veiled in him a Prophetic sorrow deep as Dante's." And since World War II there have been many attempts to replace the image of the serene optimist by that of the tortured skeptic. The one is as inadequate as the other—as inadequate as T.S. Eliot's conclusion that he was sage rather than poet—though this is perhaps inevitable when a writer is such a master of his own medium that even his prose proves resistant to translation. Even his Werther knew that the realities of existence are rarely to be grasped by Either-Or. And the reality of Goethe himself certainly eludes any such attempt. If he was a skeptic, and he often was, he was a hopeful skeptic. He looked deep into the abyss, but he deliberately emphasized life and light. He lived life to the full at every level, but never to the detriment of the civilized virtues. He remained closely in touch with the richness of his unconscious mind, but he shed on it the light of reflection without destroying the spontaneity of its processes. He was, as befits a son of the Enlightenment, wholly committed to the adventure of science; but he stood in awe and reverence before the mystery of the universe. Goethe nowhere formulated a system of thought. He was as impatient of the sterilities of logic chopping as of the inflations of metaphysics, though he acknowledged his indebtedness to many philosophers, including Kant. But here again he was not to be confined. Truth for him lay not in compromise but in the embracing of opposites. And this is expressed in the form of his *Maximen* ("Maxims"), which, together with his *Gespräche* ("Conversations"), contain the sum of his wisdom. As with proverbs, one can always find among them a twin that expresses the complementary opposite. And they have something of the banality of proverbs too. But it is, as André Gide observed, "*une banalité supérieure*." What makes it "superior" is that the thought has been felt and lived and that the formulation betrays this. And for all his specialized talents, there was a kind of "superior banality" about Goethe's life. If he himself felt it was "symbolic" and worth presenting as such in a series of autobiographical writings, it was not from arrogance but from a realization that he was an extraordinarily ordinary man in whom ordinary men might see themselves reflected. Not an ascetic, a mystic, a saint, or a recluse, not a Don Juan or a poet's poet but one who to the best of his ability had tried to achieve the highest form of *l'homme moyen sensuel*—which is perhaps what Napoleon sensed when after their meeting in Erfurt he uttered his famous "*Voilà un homme!*"

MAJOR WORKS

VERSE: *Das Buch Annette* (1767); *Neue Lieder*, so-called *Leipziger Liederbuch* (1770); *Römische Elegien* (1793); *Venezianische Epigramme* (1795); *Xenien*, written with Schiller (1796); *Balladenjahr* (1797); *Epilog zu Schillers Glocke* (1805); *West-östlicher Divan* (1819); *Trilogie der Leidenschaft* (1823–27).

DRAMATIC WORKS: *Die Laune des Verliebten* (1767, printed 1806); *Die Mitschuldigen* (1768, printed 1787); *Götz von Berlichingen* (1773); *Götter, Helden und Wieland* (1774); *Clavigo* (1774); *Erwin und Elmire* (1775), operetta; *Stella* (1776); *Claudine von Villa Bella* (1776), operetta; *Die Geschwister* (1776); *Der Triumph der Empfindsamkeit* (1777–78); *Iphigenie auf Tauris* (1779; final version, 1787); *Egmont* (1788); *Torquato Tasso* (1790); *Faust, Ein Fragment* (1790); *Der Gross-Cophta* (1792); *Der Bürgergeneral* (1793); *Die Natürliche Tochter* (1804); *Faust I* (1808); *Pandoras Wiederkunft* (1808); *Des Epimenides Erwachen* (1815); *Faust II* (1832).

NOVELS AND EPICS: *Die Leiden des jungen Werthers* (1774); *Reineke Fuchs* (1794), satire; *Unterhaltungen deutscher Ausgewanderten* (1795); *Wilhelm Meisters Lehrjahre* (1795–96); *Hermann und Dorothea* (1797); *Achilleis* (1808), epic fragment; *Die Wahlverwandtschaften* (1809); *Wilhelm Meisters Wanderjahre* (1821–29); *Novelle* (1826).

AUTOBIOGRAPHY: *Aus meinem Leben, Dichtung und Wahrheit* (1811–22); *Italienische Reise* (1816–17); *Campagne in Frankreich, 1792* (1822); *Die Belagerung von Mainz 1793* (1822).

SCIENCE: *Versuch die Metamorphose der Pflanzen zu erklären* (1790); *Beiträge zur Optik* (1791–92); *Zur Farbenlehre* (1810); *Zur Naturwissenschaft überhaupt, besonders zur Morphologie* (1817–23), journal.

CRITICISM: *Von deutscher Baukunst* (1773); *Die Propyläen* (1798–1800), journal ed. by Goethe; *Winkelmann und sein Jahrhundert* (1805); *Über Kunst und Altertum*, 6 vol. (1816–32), periodical.

TRANSLATIONS: There is still no uniform edition in English though the chief works have been frequently translated and many of them are to be found in Bohn's Standard Library, 14 vol. (1846–90). For details see Lina and Eugen Oswald, and Alexander Dickson, *Goethe in England (and America)* in *Publications of the English Goethe Society*, 2 vol. (1909, 1951). Translations include: *Wilhelm Meister's Theatrical Mission* by G.A. Page (1913); *Wilhelm Meister* by Thomas Carlyle (1824; Everyman ed., 1912, reissued 1965). *Faust I and II* by A.G. Latham (1902-06; Everyman ed., 1908); other translations of *Faust* by Bayard Taylor (1870-71; World's Classics, 1932); Philip Wayne (Penguin ed., 1949); Louis MacNeice (1951); Bertram Jessup (1958); John Shawcross (1959); Alice Raphael (1963); Walter Kaufmann (1963); and C.E. Passage (1965). *Egmont* by Michael Hamburger in *The Classic Theatre*, vol. 2 ed. by E. Bentley (1959); by Willard R. Trask (1960); *Götz von Berlichingen*, by C.E. Passage (1965); *Iphigenia in Tauris*, by C.E. Passage (1963); *Italian Journey, 1786–1788*, by W.H. Auden and Elizabeth Mayer (1962); *Torquato Tasso*, by C.E. Passage (1966); *Kindred by Choice* (*Die Wahlverwandtschaften*) by H.M. Waidson (1961); by Elizabeth Mayer and Louise Bogan (1963); *Werther* by William Rose (1929); *Werther* by B.Q. Morgan (1957); *Die Leiden des jungen Werther*, Dual-Language book, ed. by H. Steinhauer (1962); *Werther: The New Melusina. Novelle* by Victor Lange (1949); *Novelle* by C.P. Middleton (1959); *The West-Eastern Divan* by Edward Dowden (1914). *The Briefwechsel* with Schiller and a collection of early and miscellaneous letters are in Bohn (see above); that with Carlyle, ed. by C.E. Norton (1887). *Letters from Goethe*, selected and trans. by M. von Herzfeld and C.M. Sym (1957). *Conversations With Eckermann* by John Oxenford, 2 vol. (1850; vol. 6 of Bohn); by R.O. Moon (1951).

BIBLIOGRAPHY

Works, drawings, diaries, correspondence, and conversations: Goethes Werke (Vollständige Ausgabe letzter Hand), 40 vol. (1827–30), and its continuation, *Nachgelassene Werke*, 20 vol. (1832–42)—both also in a *Taschenausgabe*. The standard critical edition was published in Weimar at first under the auspices of the grand duchess Sophie (hence known as the *Weimarer Ausgabe* or *Sophienausgabe*), 133 vol., including the scientific works, diaries, and letters (1887–1919); a new critical edition by the Deutschen Akademie der Wissen-

schaften zu Berlin is in progress (1952–). Of modern editions the most noteworthy are: the *Jubiläums-Ausgabe*, ed. by EDUARD VON DER HELLEN, 40 vol., plus an invaluable *Registerband*, with introductions and notes (1902–12); the *Gedenkausgabe*, ed. by ERNST BEUTLER, 24 vol., including selections from the scientific works, correspondence, and conversations, with useful indexes (1948–54) plus *Ergänzungsband I: Briefe aus dem Elternhaus* (1960); and the *Hamburger Ausgabe*, ed. by ERICH TRUNZ, 14 vol., the most up-to-date in its introductions and bibliographical surveys (1948–60). There are innumerable collections of selected works, such as *Goethes Amtliche Schriften*, ed. by WILLY FLACH (1950, in progress). Goethe's drawings and sketches are being published in the *Corpus der Goethezeichnungen*, ed. by GERHARD FEMMEL, 6 vol. to date (1958–70); see also LUDWIG MUNZ, *Goethes Zeichnungen und Radierungen* (1949); *Goethes Gespräche*, ed by F.W. VON BIEDERMANN (1st ed., 10 vol., 1889–96; 2nd ed., 5 vol., 1909–11); and JOHANN P. ECKERMANN, *Gespräche mit Goethe . . . 1823–32*, 3 vol. (1836–48, many times reprinted). The only complete collection of Goethe's letters is in the *Weimarer Ausgabe;* see also the letters to Goethe, ed. by RUDOLF C. GOLDSCHMIDT-JENTNER, *Eine Weltschreibt an Goethe* (1937). His correspondence with particular individuals, German and foreign, is conveniently listed in the *Hamburger Ausgabe*, vol. 14; see also D.F.S. SCOTT (ed.), *Some English Correspondents of Goethe* (Matthew ["Monk"] Lewis, T. Holcroft, Sir Walter Scott, P.P. Gillies, Sir John Bowring, Lord Leveson-Gower, Sarah Austin, etc.; 1949).

Bibliography: Only the researcher will consult the vast compilation in KARL GOEDEKE, *Grundriss . . .*, 4 pt. (1910–13 and later editions), superseded by HANS PYRITZ, *Goethe-Bibliographie*, new ed., 2 vol. (1965–68). Vol. 14 of the *Hamburger-Ausgabe* (1960), is adequate for most purposes. The *Goethe-Handbuch*, ed. by ALFRED ZASTRAU (1955, in progress), provides alphabetical references to all topics relating to Goethe. HEINZ KINDERMANN, *Das Goethebild des XX. Jahrhunderts* (1952), is in effect a lively catalogue raisonné of European and American critical literature during the 20th century.

Goethe societies: Wiener Goethe-verein, 1878; Weimarer Goethe Gesellschaft, 1885; English Goethe Society, 1886, with sister societies in the Commonwealth of Nations from 1949; Goethe Society of Maryland and the District of Columbia, 1932, Japanese Goethe Gesellschaft, 1932. All issue publications.

Biography and criticism: The best introduction in English is still GEORGE H. LEWES, *The Life and Works of Goethe*, 2 vol. (1855; 18th ed., 1903; new ed., *The Life of Goethe*, 1965). Others available in English are by ALBERT BIELSCHOWSKY, 3 vol. (1905–08, reprinted 1969); BENEDETTO CROCE (1923, reprinted 1970); and GEORGE M. BRANDES, 2 vol. (1925). See also EMIL LUDWIG, *Goethe, Geschichte eines Menschen*, 2 vol. (1926; Eng. trans., *Goethe: the History of a Man, 1749–1832*, 2 vol., 1928); HENRY W. NEVINSON, *Goethe: Man and Poet* (1931); JOHN G. ROBERTSON, *The Life and Work of Goethe, 1749–1832*, 2nd ed. (1932); LUDWIG LEWISOHN (ed.), *Goethe: The Story of a Man*, 2 vol. (1949); ALBERT SCHWEITZER, *Goethe* (1949); and KARL VIETOR, *Goethe, The Poet* (1949, reprinted 1970). Criticism in English includes BARKER FAIRLEY, *Goethe, as Revealed in His Poetry* (1932) and *A Study of Goethe* (1947); THOMAS MANN, *Three Essays* (Eng. trans. 1932); WILLIAM ROSE (ed.), *Essays on Goethe* (1949); A.R. HOHLFELD, *Fifty Years with Goethe, 1901–1951: Collected Studies* (1953); ELIZABETH M. WILKINSON and LEONARD A. WILLOUGHBY, *Goethe: Poet and Thinker* (1962); RONALD PEACOCK, *Goethe's Major Plays* (1959).

On Faust: Urfaust and Faust, ed. by LEONARD A. WILLOUGHBY (1943); *Faust I and II*, ed. by R.-M.S. HEFFNER, HELMUT REHDER, and W.F. TWADDELL (1954–55); HERMANN G. FIEDLER, *Textual Studies of Goethe's Faust* (1946). Further interpretations of Faust in English are by F.M. STAWELL and G.L. DICKINSON (1928); RONALD D. MILLER (1939); D.J. ENRIGHT (1949); BARKER FAIRLEY (1953); ALEXANDER GILLIES (1957); and STUART ATKINS (1958); see also ELIZA M. BUTLER, *The Fortunes of Faust* (1952); and ADOLF I. FRANTZ, *Half a Hundred Thralls to Faust* (1949).

Science, philosophy, and aesthetics: SIR CHARLES SHERRINGTON, *Goethe on Nature and on Science*, 2nd ed. (1949), unsympathetic; as corrective, AGNES ARBER, *Goethe's Botany*, including trans. of *Die Metamorphose der Pflanzen* (1946), and *The Natural Philosophy of Plant Form* (1950); RENE MICHEA, *Les Travaux scientifiques de Goethe* (1943); RONALD D. GRAY, *Goethe, the Alchemist* (1952); ERICH HELLER, *The Disinherited Mind*, 3rd ed. (1971); OTTO HARNACK, *Goethe in der Epoche seiner Vollendung, 1805–1832*, 3rd ed. (1905); GEORG SIMMEL, *Goethe* (1913); LANCELOT L. WHYTE, *The Next Development in Man* (1944); CARL J. OBENAUER, *Goethe in seinem Verhältnis zur Religion* (1921); PETER DEMETZ, *Goethes "Die Aufgeregten": Zur Frage der politischen Dichtung in Deutschland* (1952); KARL VIETOR, *Goethe: Dichtung, Wissenschaft, Weltbild* (1949; Eng. trans. of pt. 2–3, *Goethe the Thinker*, 1950); HANS M. WOLFF, *Goethes Weg zur Humanität* (1951); FRITZ-JOACHIM VON RINTELEN, *Der Rang des Geistes: Goethes Weltverständnis* (1955); PAUL STOCKLEIN, *Wege zum späten Goethe: Dichtung, Gedanke, Zeichnung*, 2nd ed. (1960); MATTHIJS JOLLES, *Goethes Kunstanschauung* (1957); HERMANN J. WEIGAND (ed. and trans.), *Goethe: Wisdom and Experience* (1949), selections in English; GUNTHER MULLER, *Maximen und Reflexionen* (1943); THOMAS MANN, *The Permanent Goethe* (1948); WOLFGANG LEPPMANN, *The German Image of Goethe* (1961).

Weltliteratur: JEAN M. CARRE, *Goethe en Angleterre* (1920); WILLIAM ROSE, *From Goethe to Byron: The Development of "Weltschmerz" in German Literature* (1924); JOHN G. ROBERTSON, *Goethe and Byron* (1925); STUART ATKINS, *The Testament of "Werther" in Poetry and Drama* (1949); ELIZA M. BUTLER, *Byron and Goethe: Analysis of a Passion* (1956); FREDERICK NORMAN, *Henry Crabb Robinson and Goethe*, 2 vol. (1930–31); JAMES B. ORRICK, *Matthew Arnold and Goethe* (1928); FRITZ STRICH, *Goethe und die Weltliteratur* (1946; Eng. trans., *Goethe and World Literature*, 1949); HUMPHRY TREVELYAN, *Goethe and the Greeks* (1941; a corrective to ELIZA M. BUTLER, *The Tyranny of Greece Over Germany*, 1935); JAMES BOYD, *Goethe's Knowledge of English Literature* (1932); BERTRAM BARNES, *Goethe's Knowledge of French Literature* (1937); HORST OPPEL, *Das Shakespeare-Bild Goethes* (1949); ARNOLD FEDERMANN, *Der junge Goethe und England* (1949).

Cultural background: HERMANN A. KORFF, *Geist der Goethezeit*, 5 vol. (1923–58); ERNST BEUTLER, *Essays um Goethe*, 6th ed., 2 vol. (1962); *Goethe und siene Welt*, with 580 illustrations, ed. by ERNST BEUTLER, HANS WAHL, and ANTON KIPPENBERG (1932); ARNOLD BERGSTRAESSER, *Goethe's Image of Man and Society* (1949); WALTER H. BRUFORD, *Theatre, Drama, and Audience in Goethe's Germany* (1950) and *Culture and Society in Classical Weimar, 1775–1806* (1962).

(E.M.Wn.)

Gogh, Vincent van

Largely on the basis of the works of the last three years of his life, which ended in suicide in 1890, when he was only 37, Vincent van Gogh is generally held to have been the greatest Dutch painter after Rembrandt. Everything in van Gogh's pictures seems to be pulsating with life. His work exerted a powerful influence on the development of much modern painting, in particular on the works of Soutine and the German Expressionists. Yet of the more than 800 oil paintings and 700 drawings which constitute his life's work, he sold only one in his lifetime. Always desperately poor, he was sustained by his faith

By courtesy of Mr. and Mrs. Leigh B. Block, Chicago



"Self-Portrait with Pipe and Bandaged Ear," oil painting by Vincent van Gogh, 1889. In the Leigh B. Block Collection, Chicago.

in the urgency of what he had to communicate and by the generosity of a younger brother, Theo, who believed in him implicitly. The letters that he wrote to Theo from 1872 onward, and to other friends, give such a vivid account of his aims and beliefs, his hopes and disappointments, his fluctuating physical and mental state, that they form a unique and touching biographical record which is also a great human document.

Vincent Willem van Gogh was born on March 30, 1853, the eldest of six children of a Protestant pastor, at Zundert, a small village in the Brabant region of southern Netherlands. His early years in his father's parsonage were happy, and he loved wandering in the countryside. At 16 he was apprenticed to the Hague branch of the art dealers Goupil and Co., of which his uncle was a partner.

Van Gogh's working life can be roughly divided into two periods. The first, from 1873 to 1885, during which he wrestled with temperamental difficulties and sought his true means of self-expression, was a period of repeated apprenticeships, failures, and changes of direction. The second, from 1886 to 1890, was a period of dedication, rapid development, and fulfillment, until it was interrupted by a series of mental crises from 1889 onward.

He worked for the art-dealing firm Goupil in London from 1873 to May of 1875 and in Paris from May 1875 to April 1876.

Daily contact with works of art aroused van Gogh's artistic sensibility, and he soon formed a taste for Rembrandt, Hals, and other Dutch masters, although his preference was for two contemporary French painters, Millet and Corot, whose influence was to last throughout his life. Van Gogh disliked art dealing. Moreover, his approach to life darkened when his love was rejected by a London girl in 1874. His burning desire for human affection thwarted, he became and remained increasingly solitary. He became a language teacher and lay preacher in England and, in 1877, worked for a bookseller in Dordrecht. Impelled by a longing to give himself to his fellowmen, he envisaged entering the ministry and took up theology but abandoned this project for short-term training as an evangelist in Brussels in 1878. A conflict with authority ensued when he disputed an orthodox doctrinal approach. Failing to get an appointment after three months, he left to do missionary work among the impoverished population of the Borinage, a coal-mining region in southwest Belgium. There, in the winter of 1879–80, he experienced the first great spiritual crisis of his life. He was sharing the life of the poor completely but in an impassioned moment gave away all his worldly goods and was thereupon dismissed for a too literal interpretation of Christian teaching.

Penniless and with his faith destroyed, he sank into despair, cut himself off from everyone, and began seriously to draw, thereby discovering in 1880 his true vocation. Van Gogh decided that his mission from then on would be to bring consolation to humanity through art, and this realization of his creative powers restored his self-confidence.

Discovery of his vocation

His artistic career was extremely short, lasting only the ten years from 1880 to 1890. During the first four years of this period, while acquiring technical proficiency, he confined himself almost entirely to drawings and watercolours. First, he went to study drawing at the Brussels Academy; in 1881 he moved to his father's parsonage at Etten, The Netherlands, and began to work from nature.

Van Gogh worked hard and methodically but soon perceived the difficulty of self-training and sought the guidance of more experienced artists. Late in 1881 he settled at The Hague to work with a Dutch landscape painter, Anton Mauve. He made visits to museums and had meetings with other painters. Van Gogh thus extended his technical knowledge and experimented in the summer of 1882 with oil paint. In 1883 the urge to be "alone with nature" and the peasants took him to Drenthe, a desolate part of northern Netherlands frequented by Mauve and other Dutch artists, where he spent three months before returning home, which was now at Nuenen, another village in the Brabant. He remained at Nuenen during most of 1884 and 1885, and during these years his art grew bolder and more assured. He painted three types of subject—still life, landscape, and figure—all interrelated by their reference to the peasants' daily life, to the hardships they endured and the countryside they cultivated. Émile Zola's *Germinal* (1885), a novel about the coal-mining region of France, greatly impressed van Gogh, and sociological criticism is implicit in many of his pictures—*e.g.*, "Weavers" and "The Potato Eaters." Eventually he felt too isolated in Nuenen.

His understanding of the possibilities of painting was evolving rapidly; from studying Hals he saw that academic finish destroys the freshness of a visual impression; Veronese and Delacroix taught him that colour expresses something by itself. This led to enthusiasm for Rubens and a sudden departure for Antwerp, where the greatest number of Rubens' works can be seen. The revelation of Rubens' simple means, of his direct notation, and of his ability to express a mood by a combination of colours proved decisive. Simultaneously, van Gogh discovered Japanese prints and Impressionist painting. His refusal to follow academic principles led to rows at the Antwerp academy, where he was enrolled, and after three months he left precipitately in 1886 to join his brother Theo in Paris. There, still concerned with improving his drawing, van Gogh met Toulouse-Lautrec, Paul Gauguin, and others who were to play historic roles in modern art. They opened his eyes to the latest developments in French painting. At the same time, Theo showed him Impressionist paintings at Goupil's and introduced him to Camille Pissarro, Georges Seurat, and other artists of the group.

Influence of the Impressionists and Post-impressionists

By this time van Gogh was ready for such revelations, and the changes that his painting underwent in Paris between the spring of 1886 and February 1888 led to the creation of his personal idiom and style of brushwork. His palette at last became colourful, his vision less traditional, and his tonalities lighter, as may be seen in his first paintings of Montmartre. Finally, van Gogh's Post-impressionist style crystallized by the beginning of 1888 in masterpieces such as "Portrait of Père Tanguy" and "Self-Portrait in Front of an Easel," as well as in some landscapes of the Parisian suburbs.

After two years, van Gogh was tired of city life, physically exhausted, and longing "to look at nature under a brighter sky." His passion was now for "the Japanese way of feeling and drawing" and for "a full effect of colour." He left Paris in February 1888 for Arles.

In his pictures of the following 12 months—his first great period—he strove to respect natural appearances and yet to convey by emphatic contours and heightened effects of colour the reality of his own feelings about the subject. Van Gogh's pictorial conception was partly Expressionist and partly Symbolist. His procedure was not calculated, however, but spontaneous and instinctive, for he worked with great speed and intensity, determined to capture an effect or a mood while it possessed him. His Arles subjects include blossoming fruit trees, views of the town and surroundings, self-portraits, portraits of Roulin the postman and his family and other friends, interiors and exteriors of the house, a series of sunflowers, and a "starry night."

Van Gogh knew that his approach to painting was individualistic, but he was also deeply aware that some tasks are beyond the power of isolated individuals to accomplish. In Paris he had hoped to form a separate Impressionist group with Gauguin, Toulouse-Lautrec, and others whom he supposed to have similar aims. He rented and decorated "a yellow house" in Arles with the intention of persuading them to join him and found a working community of "Impressionists of the South." Gauguin arrived in October 1888, and for two months they worked together; but while each influenced the other to some extent, their relations rapidly deteriorated because they had opposing ideas and were temperamentally incompatible.

Mental breakdown

On Christmas Eve 1888, van Gogh broke under the strain and cut off part of his left ear. Gauguin left, and van Gogh was taken to a hospital; he returned to the



"Bedroom at Arles," oil painting by van Gogh, 1888 or 1889. In The Art Institute of Chicago.
By courtesy of The Art Institute of Chicago, Helen Birch Bartlett Memorial Collection

"yellow house" a fortnight later and resumed painting: "Self-Portrait with Pipe and Bandaged Ear," still lifes, "La Berceuse." Several weeks later, he again showed symptoms of mental disturbance severe enough to cause him to be sent back to the hospital. At the end of April 1889, fearful of losing his renewed capacity for work, which he regarded as a guarantee of his sanity, he asked to be temporarily shut up in the asylum at Saint-Rémy-de-Provence in order to be under medical supervision.

Vincent stayed there for 12 months, haunted by recurrent attacks, alternating between moods of calm and despair, and working intermittently: "Garden of the Asylum," "Cypresses," "Olive Trees," "Les Alpilles," portraits of doctors, and interpretations of paintings by Rembrandt, Delacroix, and Millet. The keynote of this phase (1889–90) is fear of losing touch with reality and a certain sadness. Confined for long periods to his cell or the asylum garden, having no choice of subjects, and realizing that his inspiration depended on direct observation, van Gogh fought against having to work from memory. At Saint-Rémy he muted the violent colours of the previous summer and tried to make his painting calmer. As he repressed his excitement, however, he involved himself more imaginatively in the drama of the elements, developing a style based on dynamic forms and a vigorous use of line (line often equated with colour). The best of his Saint-Rémy pictures are thus bolder and more visionary than those of Arles.

Van Gogh himself brought this period to an end. Oppressed by homesickness—he painted souvenirs of Holland—and loneliness, he longed to see Theo and the north once more and arrived in Paris in May 1890. Four days later he went to stay with a homeopathic doctor-artist, Dr. Paul-Ferdinand Gachet, a friend of Pissarro and Paul Cézanne, at Auvers-sur-Oise. Back in a village community such as he had not known since Nuenen, four years earlier, van Gogh worked at first enthusiastically; and his choice of subjects such as fields of corn, the river valley, peasants' cottages, the church, and the town hall reflects his spiritual relief. A modification of his style follows: in the northern light he adopted pale, fresh tonalities; the brushwork is broader and more expressive; the vision of nature more lyrical. But this phase was short; it ended in quarrels with Gachet and feelings of guilt at his inescapable dependence on Theo (now married and with a son) and his inability to succeed. In despair of ever overcoming his loneliness or of being cured, he shot himself and died two days later, on July 29, 1890. Coincidentally, Theo died six months later (January 25, 1891) of chronic nephritis.

The name of van Gogh was virtually unknown when he took his life. He had exhibited a few canvases at the Salon des Indépendants in Paris between 1888 and 1890 and in Brussels in 1890; both salons showed small commemorative groups of his work in 1891. One-man shows of his work did not occur until 1892; only one article about him appeared during his lifetime. His fame, made largely by other painters, dates from the early years of the 20th century and his reputation has never ceased to grow.

MAJOR WORKS

"In the Field" (1883; Rijksmuseum Vincent van Gogh, Amsterdam); "The Loom" (1884; Rijksmuseum Kröller-Müller, Otterlo, The Netherlands); "Autumn Landscape with Four Trees" (1885; Rijksmuseum Kröller-Müller, Otterlo); "Peasant Woman in a Red Bonnet" (1885; Rijksmuseum Vincent van Gogh, Amsterdam); "The Potato Eaters" (1885; Rijksmuseum Vincent van Gogh, Amsterdam); "Skull with a Cigarette" (1886; Rijksmuseum Vincent van Gogh, Amsterdam); "Boots" (1887; Museum of Art, Baltimore); "View of Arles with Irises" (1888; Rijksmuseum Vincent van Gogh, Amsterdam); "La Mousmé" (1888; Chester Dale Collection, National Gallery of Art, Washington, D.C.); "The Postman Roulin" (1888; Museum of Fine Arts, Boston); "The Chair and the Pipe," popularly known as "Van Gogh's Chair" (1888–89; Tate Gallery, London); "Self-Portrait in Front of an Easel" (1888; Rijksmuseum Vincent van Gogh, Amsterdam); "The Night Café" (1888; Yale University Art Gallery); "Self-Portrait with Pipe and Bandaged Ear" (1889; Leigh B. Block Collection, Chicago); "Green Corn" (1889; Národní Galerie, Prague); "Pietà" (1889; Rijksmuseum Vincent van Gogh, Amsterdam); "L'Arlésienne" (1890; Museu de Arte, São Paulo, Brazil); "Hospital Corridor at Saint-Rémy" (1889; Museum of Modern Art, New York); "The Starry Night" (1889; Museum of Modern Art, New York); "Old Man in Sorrow," also known as "On the Threshold of Eternity" (1890; Rijksmuseum Kröller-Müller, Otterlo); "Mademoiselle Gachet at the Piano" (1890; Kunstmuseum, Basel, Switzerland); "Stairway at Auvers" (1890; City Art Museum, St. Louis, Missouri); "The Church at Auvers" (1890; Louvre, Paris); "Self-Portrait" (1890; Louvre, Paris).

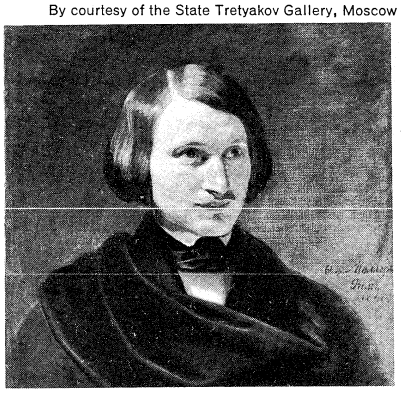
BIBLIOGRAPHY. There is no thorough, detached biography of van Gogh. The best recent contributions are JEAN LEYMARIE, *Van Gogh* (1951), a sound and intelligent study; also the same author's *Qui était Van Gogh?* (1968; Eng. trans., *Who Was Van Gogh?*, 1968); CARL A.J. NORDENFALK, *The Life and Work of Vincent van Gogh* (1953), an elementary work; HENRI PERRUCHOT, *La Vie de van Gogh* (1955), a journalistic account; and MARC E. TRALBAUT, *Vincent van Gogh* (1969), well documented but erratic and melodramatic. The basic reference books are *The Complete Letters of Vin-*

cent van Gogh, 3 vol. (1958)—the translation, transcription, and chronology are erratic and can be corrected against the original texts published in *Verzamelde brieven van Vincent van Gogh*, 4 vol. (1952–54), and acceptable revised datings argued by J. HULSKER in articles in the review *Maatstaaf* (1958–61), summarized in English in numbers of the review *Museumjournaal* (1959–61); and J. BAERT DE LA FAILLE, *L'Oeuvre de Vincent van Gogh*, 4 vol. (1928; Eng. trans., *The Works of Vincent Van Gogh: His Paintings and Drawings*, 3rd ed. rev., 1970), a catalogue raisonné critically revised (though inadequately) by Dutch art historians. The following scholarly studies of specific aspects of van Gogh's work are also specially recommendable: ANNA SZYMANSKA, *Unbekannte Jugendzeichnungen Vincent van Goghs . . . 1870–1880*, 2nd ed. (1968); W. VANBESELAERE, *De Hollandsche periode (1880–85) in het werk van Vincent Van Gogh* (1937); MARK W. ROSKILL, *Van Gogh, Gauguin and the Impressionist Circle* (1970; with a complementary *Catalogue Raisonné of Key Works*, 1971); KURT BADT, *Die Farbenlehre Van Goghs* (1961); DOUGLAS COOPER, *Drawings and Watercolours by Van Gogh* (1955); and NICHOLAS WADLEY, *The Drawings of Van Gogh* (1969).

Gogol, Nikolay

The great tradition of the Russian novel—through which Russia assumed a place among the major literary peoples of the world—was established largely by Nikolay Gogol, whose stories, plays, and novels of Russian life—noted for their satiric humour and their mixture of realistic and fantastic elements—revealed the proclivity of Russian fiction for a medium of expression that was more Realist than Romantic.

By courtesy of the State Tretyakov Gallery, Moscow



Gogol, oil painting by F.A. Moller, 1840. In the State Tretyakov Gallery, Moscow.

Youth and early fame. Nikolay Vasilyevich Gogol was born on March 31 (March 19, old style), 1809, at Sorochintsy near Poltava. The Ukrainian countryside, with its colourful peasantry, its Cossack traditions, and rich folklore, constituted the background of his boyhood. A member of the petty Ukrainian gentry, Gogol was sent at the age of 12 to the high school at Nezhin. There he distinguished himself by his biting tongue, his contributions of prose and poetry to a magazine, and his portrayal of comic old men and women in school theatricals. In 1828 he went to St. Petersburg, hoping to enter the civil service, but soon discovered that without money and connections he would have to fight hard for a living. He even tried to become an actor, but his audition was unsuccessful. In this predicament he remembered a mediocre sentimental–idyllic poem he had written in the high school. Anxious to achieve fame as a poet, he published it at his own expense, but its failure was so disastrous that he burned all the copies and thought of emigrating to the United States. Though he actually took a boat to Lübeck, he soon ran out of money in Germany and returned to St. Petersburg to look again for a job. What he obtained was such an ill-paid government post that after three months he exchanged it for another, which he also left about a year later.

In the meantime he wrote occasionally for periodicals, finding an escape in childhood memories of the Ukraine. He committed to paper what he remembered of the sunny landscapes, peasants, and boisterous village lads and also related tales about devils, witches, and other demonic or fantastic agents that enliven Ukrainian folklore. Romantic stories of the past were thus intermingled with realistic incidents of the present, in which Gogol's sense of fun and at times his feeling of evil came into their own. Such was the origin of his eight narratives, published in two volumes in 1831–32 under the title *Evenings on a Farm near Dikanka* (*Vechera na khutore bliz Dikanki*). Written in a lively and at times colloquial prose, they contributed something fresh and new to Russian literature. In addition to the author's whimsical inflection, they abounded in genuine folk flavour, including many Ukrainian words and phrases, all of which captivated the Russian literary world.

Use of colloquial language and folklore

Mature career. The young author became famous overnight. Among his first admirers were the poets Aleksandr Pushkin and Vasily Zhukovsky, both of whom he had met before. This esteem was soon shared by the writer Sergey Aksakov and the critic Vissarion Belinsky, among others. Having given up his second government post, Gogol was now teaching history in a boarding school for girls. In 1834 he was appointed assistant professor of medieval history at St. Petersburg University, but he felt inadequately equipped for the position and left it after a year. Meanwhile, he prepared energetically for the publication of his next two books, *Mirgorod* and *Arabesques* (*Arabeski*), which appeared in 1835. The four stories comprising *Mirgorod* were a continuation of the *Evenings*, but they revealed a strong gap between Gogol's romantic escapism and his otherwise pessimistic attitude toward life. Such a splendid narrative of the Cossack past as "Taras Bulba" certainly provided an escape from the present. "Story of the Quarrel between Ivan Ivanovich and Ivan Nikiforovich" ("Povest o tom, kak possorilsya Ivan Ivanovich s Ivanom Nikiforovichem") was, for all its humour, full of bitterness about the meanness and vulgarity of existence. Even the idyllic motif of Gogol's "Old-World Landowners" ("Starosvetskiye pomeshchiki") is undermined with satire, for the mutual affection of the aged couple is marred by gluttony, their ceaseless eating for eating's sake.

Aggressive realism of a romantic who can neither adapt himself to the world nor escape from it, and is therefore all the more anxious to expose its vulgarity and evil, predominates in Gogol's Petersburg stories printed (together with some essays) in the second work, *Arabesques*. In one of these stories, "Diary of a Madman" ("Zapiski sumasshedshego"), the hero is an utterly frustrated office drudge who finds compensation in megalomania and ends in a lunatic asylum. In another, "Nevsky Prospect," a tragic romantic dreamer is contrasted to an adventurous vulgarian, while in the revised finale of "The Portrait" ("Portret") the author stresses his conviction that evil is ineradicable in this world. In 1836 Gogol published in Pushkin's *Sovremennik* ("The Contemporary") one of his gayest satirical stories, "The Coach" ("Kolyaska"). In the same periodical also appeared his amusingly caustic surrealist tale, "The Nose" ("Nos"). Gogol's association with Pushkin was of great value because he always trusted his friend's taste and criticism; moreover, he received from Pushkin the themes for his two principal works, *The Government Inspector* (*Revizor*), sometimes translated *The Inspector General*, and *Dead Souls* (*Myortvye dushi*), which were important not only to Russian literature but also to Gogol's further destiny.

Publication of the Petersburg stories

A great comedy, *The Government Inspector* mercilessly lampoons the corrupt bureaucracy under Nicholas I. Having mistaken a well-dressed windbag for the dreaded incognito inspector, the officials of a provincial town bribe and banquet him in order to turn his attention away from the crying evils of their administration. But during the triumph, after the bogus inspector's departure, the arrival of the real inspector is announced—to the horror of those concerned. It was only by a special order of the tsar that the first performance of this comedy of indictment and "laughter through tears" took place on April 19, 1836. Yet the hue and cry raised by the reactionary press and officialdom was such that Gogol left

Russia for Rome, where he remained, with some interruptions, until 1842. The atmosphere he found in Italy appealed to his taste and to his somewhat patriarchal—not to say primitive—religious propensity. The religious painter Aleksandr Ivanov, who worked in Rome, became his close friend. He also met a number of travelling Russian aristocrats and often saw the émigrée Princess Zinaida Volkonsky, a convert to Roman Catholicism, in whose circle religious themes were much discussed. It was in Rome, too, that Gogol wrote most of his masterpiece, *Dead Souls.*

Composition of *Dead Souls*

This novel, or "epic," as the author labelled it, reflects feudal Russia, with its serfdom and bureaucratic iniquities. Chichikov, the hero of the novel, is a polished swindler who, after several reverses of fortune, wants to get rich quick. His bright but criminal idea is to buy from various landowners a number of the recently deceased serfs (or "souls," as they were called in Russia) whose deaths have not yet been registered by the official census and are therefore regarded as still being alive. The landowners are only too happy to rid themselves of the fictitious property on which they continue to pay taxes until the next census. Chichikov intends to pawn the "souls" in a bank, and, with the money thus raised, settle down in a distant region as a respectable gentleman. The provincial townsmen of his first stop are charmed by his polite manners; he approaches several owners in the district who are all willing to sell the "souls" in question, knowing full well the fraudulent nature of the deal. The sad conditions of Russia, in which serfs used to be bought and sold like cattle, are evident throughout the grotesquely humorous transactions. The landowners, one more queer and repellent than the last, have become nicknames known to every Russian reader. When the secret of Chichikov's errands begins to leak out, he hurriedly leaves the town.

Dead Souls appeared in 1842, the same year in which the first edition of Gogol's collected works was published. The edition included, among his other writings, a sprightly comedy titled the *Marriage* (*Zhenitba*) and the story "The Overcoat" ("Shinel"). The latter concerns a humble scribe who, with untold sacrifices, has acquired a smart overcoat; when robbed of it he dies of a broken heart. The tragedy of this insignificant man was worked out with so many significant trifles that, years later, Dostoyevsky was to exclaim that all Russian Realists had come "from under Gogol's greatcoat." The apex of Gogol's fame was, however, *Dead Souls.* The democratic intellectuals of Belinsky's brand saw in this novel a work permeated with the spirit of their own liberal aspirations. Its author was all the more popular because after Pushkin's tragic death he was now looked upon as the head of Russian literature. Gogol, however, began to see his leading role in a perspective of his own. Having witnessed the beneficent results of the laughter caused by his indictments, he was sure that God had given him a great literary talent in order to make him not only castigate abuses through laughter but also to reveal to Russia the righteous way of living in an evil world. He therefore decided to continue *Dead Souls* as a kind of *Divine Comedy* in prose; the already published part would represent the *Inferno* of Russian life, and the second and third parts (with Chichikov's moral regeneration) would be its *Purgatorio* and *Paradiso.*

Creative decline. Unfortunately, having embarked upon such a soul-saving task, Gogol noticed that his former creative capacity was deserting him. He worked on the second part of his novel for over ten years, but with meagre results. In drafts of four chapters and a fragment of the fifth found among his papers, the negative and grotesque characters are drawn with some intensity, whereas the virtuous types he was so anxious to exalt are stilted and devoid of life. This lack of zest was interpreted by Gogol as a sign that, for some reason, God no longer wanted him to be the voice exhorting his countrymen to a more worthy existence. In spite of this he decided to prove that at least as teacher and preacher—if not as artist—he was still able to set forth what was needed for Russia's moral and worldly improvement. This he did in his ill-starred *Selected Passages from Correspondence with My Friends* (*Vybrannye mesta iz perepiski s druzyami,* 1847), a collection of 32 discourses eulogizing not only the conservative official church but also the very powers that he had so mercilessly condemned only a few years before. It is no wonder that the book was fiercely attacked by his one-time admirers, most of all by Belinsky, who in an indignant letter called him "a preacher of the knout, a defender of obscurantism and of darkest oppression." Crushed by it all, Gogol saw in it a further proof that, sinful as he was, he had lost God's favour forever. He increased his prayers and his ascetic practices; in 1848 he even made a pilgrimage to Palestine, but in vain. Despite a few bright moments he began to wander from place to place like a doomed soul. Finally he settled in Moscow where he came under the influence of a fanatical priest, Father Matvey Konstantinovsky, who seems to have practiced on Gogol a kind of spiritual sadism. Ordered by him, Gogol burned the presumably completed manuscript of the second volume of *Dead Souls* on February 24 (February 11, O.S.), 1852. Ten days later, on March 4 (February 21, O.S.), he died, on the verge of semi-madness.

Political conservatism and religious fanaticism

Influence and reputation. Whatever the vagaries of Gogol's mind and life, his part in Russian literature was enormous. Above all, it was from the nature of such works as *The Government Inspector, Dead Souls,* and "The Overcoat" that Belinsky derived the tenets of the "natural school" (as distinct from the "rhetorical," or Romantic, school) that was responsible for the trend of subsequent Russian fiction. Gogol was among the first authors to have revealed Russia to herself. Yet in contrast to the simple classical–realistic prose of Pushkin, adopted by Leo Tolstoy, Ivan Goncharov, and Ivan Turgenev, Gogol's ornate and agitated prose was assumed by Fyodor Dostoyevsky and, later, by the Symbolist poet and novelist Andrey Bely whose style has, in turn, influenced several post-revolutionary writers of the Soviet Union. Gogol's realism of indictment found many followers, among them the great satirist Saltykov-Shchedrin. He was also a champion of the little man as a literary hero. His vexation of spirit, too, was continued (but on a higher level) by both Tolstoy and Dostoyevsky as was his effort to transcend "mere literature."

MAJOR WORKS

NOVEL: *Myortvye dushi* (1842; *Dead Souls,* trans. by Constance Garnett in *The Works of Nikolay Gogol,* 1922–23; by G. Reavey, 1948; and David Magarshack, 1961).

STORIES: *Vechera na khutore bliz Dikanki,* vol. 1 (1831), vol. 2 (1832), including "Sorochinskaya yarmarka," trans. by Rosa Portnova as "Sorochinsky Fair" in *Tales from Gogol* (1945); "Vecher nakanune Ivana Kupala," trans. by Constance Garnett as "St. John's Eve" in *The Works of Nikolay Gogol;* "Mayskaya noch; ili, utoplennitsa," trans. by C. Field as "A May Night" (1915); "Propavshaya gramota," trans. by Constance Garnett as "The Lost Letter" in *The Works of Nikolay Gogol;* "Noch pered Rozhdestvom," trans. by Rosa Portnova as "Christmas Eve" in *Tales from Gogol;* and "Zakoldovannoye mesto," trans. by Constance Garnett as "The Bewitched Spot" in *The Works of Nikolay Gogol; Arabeski* (1835), including "Portret," "Nevsky prospekt," trans. by Rosa Portnova as "The Nevsky Prospect" in *Tales from Gogol,* and "Zapiski sumasshedshego," trans. by C. Field as "Memoirs of a Madman" (1915): the first two were trans. by David Magarshack as "The Portrait" and "Nevsky Avenue" in *Tales of Good and Evil* (1949): the last two were translated separately by B. Scott as "Diary of a Madman" and "Nevski Prospect" (1945); *Mirgorod* (1835), including "Starosvetskiye pomeshchiki," trans. by Constance Garnett as "Old-World Landowners" in *The Works of Nikolay Gogol;* and "Povest o tom, kak possorilsya Ivan Ivanovich s Ivanom Nikiforovichem," trans. by Rosa Portnova as "How the Two Ivans Quarrelled" in *Tales from Gogol;* "Viy," trans. by C. Field (1915); and "Taras Bulba," trans. by David Magarshack in *Tales of Good and Evil; "Nos"* (1836; "The Nose," trans. by C. Field, 1915); "Kolyaska" (1836; trans. by Rosa Portnova as "The Coach" in *Tales from Gogol*); "Shinel" (1842; trans. by C. Field as "The Mantle," 1915; and as "The Overcoat," trans. by David Magarshack in *Tales of Good and Evil*).

PLAYS: *Zhenitba* (1842; *Marriage,* trans. by B. Costello, 1969); *Revizor,* trans. by David Magarshack as *The Government Inspector,* and by E.O. Marsh and J. Brooks, 1968).

OTHER WORKS: *Vybrannye mesta iz perepiski s druzyami* (1847; trans. by Constance Garnett as *Selected Passages from Correspondence with My Friends* in *The Works of Nikolay Gogol*, 1922).

BIBLIOGRAPHY. One of the best of the numerous Russian editions of Gogol's works is that by Н.С. ТИХОНРАВОВ, 7 vol. (1889). *Собрание сочинений*, 7 vol. (1950, 1966–67), is also a good Soviet edition. In English, see *The Works of Nikolay Gogol*, 6 vol., trans. by CONSTANCE GARNETT (1922–23). Selections include *Tales from Gogol*, trans. by ROSA PORTNOVA (1945); and *Tales of Good and Evil*, trans. by DAVID MAGARSHACK (1949). Владимир Иванович Шенрок, *Материалы для биографии Гоголя*, 4 vol. (1892–97), is indispensable for thorough investigation of Gogol's life and work. Виссарион Григорьевич Белинский, *В.Г. Белинский о Гоголе; статьи, рецензии, письма* (1949); and Николай Гаврилович Чернышевский, *Очерки Гоголевского периода русской литературы* (1953), are Soviet re-editions of the important early, for the most part, socio-political essays. Нестор Александрович Котляревский, *Николай Васильевичь Гоголь 1829–1842* (1903), is an excellent early biography. A popular, more recent biography is В.В. Ермилов, *Гоголь (Николай Васильевич)* (1952). In English, see PAUL DEBRECZENY, *N. Gogol and His Contemporary Critics* (1966), a survey; V.V. NABOKOV, *Nikolai Gogol* (1944), an astute and rather subjective monograph; J. LAVRIN, *Gogol* (1951, reprinted 1968), mainly a psychological approach; DAVID MAGARSHACK, *Gogol: A Life* (1957, reprinted 1969), based on contemporary evidence; and VICTOR ERLICH, *Gogol* (1969), an excellent, recent study.

(Ja.La.)

Goiás

An inland state (*estado*) of Brazil, Goiás, the site of the federal district (*distrito federal*) and capital, Brasília, is part of the rapidly developing frontier of the country. It is bounded by the states of Maranhão on the north, Mato Grosso and Pará on the west, Minas Gerais and Mato Grosso on the south, and Maranhão, Bahia, and Minas Gerais on the east. Its area is 247,913 square miles (642,092 square kilometres). Its population early in the 1970s was about 3,000,000. The state capital, since 1937, has been Goiânia.

History. The first European exploration of this interior part of Brazil was carried on by expeditions from São Paulo in the 17th century. Gold was discovered in the stream gravels of a tributary of the Rio Araguaia by the explorer Bartolomé Bueno de Silva in 1682. The settlement he founded there, called Santa Anna, became the colonial town of Goiás, the former state capital. In 1744 the large inland area, much of it still unexplored by Europeans, was made a captaincy general, and in 1822 it became a province of the empire of Brazil. It became a state in 1889. The Brazilian constitution of 1891 specified that the nation's capital should be moved to the Planalto Central (Central Plateau), and in 1956 Goiás was selected as the site for the federal district and capital city, Brasília. The seat of the federal government was officially moved to Brasília in 1960.

Physical geography. Goiás lies wholly within the Brazilian Highlands. In the south it occupies the larger part of the Planalto Central, the vast level surface of which stands between 2,500 and 3,000 feet above sea level. A few rounded ridges stand higher than this, the highest being the Chapada dos Veadeiros (5,505 feet [1,678 metres]). The Planalto Central forms the divide between three of Brazil's largest river systems: to the south Goiás is drained through the Rio Paranaíba, a tributary of the Paraná River; to the east it is drained by tributaries of the São Francisco River; and the greater part of the state is drained northward through the Rio Araguaia and the Rio Tocantins and their tributaries. None of these rivers is navigable except for short distances. The southern part of the state is covered with a woodland savanna known in Brazil as *campo cerrado*. To the north, where the Rio Araguaia and Rio Tocantins have eroded deep valleys, the land is covered with tropical rain forest, or *selva*.

Climate

The climate of the plateau, usually described as moderate, is subtropical. The deeper valley regions are tropical. Average monthly temperatures vary from 78.1° F (25.6° C) in the warmest month to 72° F (22° C) in the coldest month. The year is divided into a rainy season (October–March) and a dry season (April–September). Average annual rainfall is about 67 inches (1,700 millimetres).

Population. The Grande Região Centro-Oeste (Great Central East Region), consisting of Goiás, Mato Grosso, and the federal district, is the fastest growing region of Brazil. The population of Goiás was 1,214,921 in 1950; 1,954,862 in 1960; and 2,997,570 in 1970. Outside the federal district the greater part of Goiás is still very thinly populated. Only a few isolated settlements are scattered throughout the northern two-thirds of the area. The chief concentration of settlement is in the southeast, across the border from Minas Gerais. Population density in the state is 11.9 persons per square mile (4.6 per square kilometre).

The population has been predominantly rural (57.7 percent), the influx of immigrants after 1950 consisting of approximately equal numbers of rural and urban dwellers. Thus, the state has experienced rural and urban expansion at the same time. Areas adjacent to the federal district have grown the most rapidly.

Historically, the inhabitants have been predominantly of mixed European–Indian ancestry. The new settlers have come from all parts of Brazil, including many from nearby areas, and thus represent varying degrees of European, Indian, and Negro origins so that the ethnic composition of the population is becoming more typical of the nation as a whole.

Administration and social conditions. The state government moved from Goiás to Goiânia in 1937 and was officially inaugurated there in 1942. Although Goiás has been a state since 1889, it remained undeveloped in both population and resources so that, like most of the other states of Brazil, it has been unable to carry out the functions of modern government, having to rely on support at the federal level.

The standard of living is low, especially in the rural areas. Deaths from malaria have been reduced, but health services are limited, and life expectancy is still low. More primary and secondary schools have been built and attendance has increased, but growth has not kept up with the increase in population. Higher education is offered at the Universidade Católica de Goiás (founded 1959) and the Universidade Federal de Goiás (1961), both at Goiânia, and the Universidade de Brasília.

Economy. Goiás is a modern frontier area reflecting a spirit of growth and progress. Agriculture and livestock raising continue to be the most important economic activities, serving the growing urban markets. The principal crops are rice, corn (maize), beans, manioc (cassava), and coffee. Livestock raising is expanding, with cattle predominant on the open *campos* and pigs in the settled farming regions. Mineral resources include gold and diamonds (mining, carried on since the 16th century, continues although the output has never been large); tin and titanium ores (of which Goiás is Brazil's leading producer); and rock crystal (quartz crystal), chromium, and mica.

Principal crops

Goods and services for the growing pioneer society have been expanding with the growth of population since 1950. Anápolis, for example, reached by rail from Rio de Janeiro and São Paulo, has become a rapidly growing frontier town, serving the new zone of pioneer settlement in the northwest, the Mato Grosso de Goiás. Farm settlements along the valleys in the southeast have also required goods and services. The construction of Brasília and the formal transfer of the capital there reinforced Goiás' economic growth and assured its continuance.

Transportation. Until Anápolis was connected by rail with São Paulo in 1913, transport to and from the coast was by mule train. A network of feeder roads has been constructed in southern Goiás and a highway extended to Brasília. Direct access to the interior is by air.

Cultural life. Outside of Goiânia, the seat of two young universities, the state has had few cultural institutions. The establishment of the national capital within its boundaries has resulted in the development of major cultural centres there.

(P.E.J.)

Gold Products and Production

Gold possesses several outstanding qualities that have made it exceptionally valuable throughout history. It is attractive in colour and brightness, durable to the point of virtual indestructibility, highly malleable, and usually found in nature in a comparatively pure form. Examples of elaborate gold workmanship survive from ancient Egyptian, Minoan, Assyrian, and Etruscan artists, many in nearly perfect condition. The development of a more complex economic system gave gold a new major function as high-denomination currency and later as backing for paper-currency systems, a function it has retained in many countries to the present. So precious was gold in the Middle Ages that the foundation of chemistry derived largely from the vain experiments of medieval alchemists seeking to convert other metals into gold.

Gold is widely dispersed throughout the Earth's crust. Naturally occurring metallic, or native, gold usually contains variable amounts of silver, copper, platinum, palladium, or certain other elements mixed with it. Gold exists in association with most copper and lead deposits, and although the quantity of gold present is often extremely small, it is readily recovered as a by-product in the refining of those base metals.

Large masses of rock rich enough to be called ores are unusual. Those that do exist are usually quartz lodes or veins. Ores may also be found in deposits which were brought to their sites from veins, such as in river gravels and gold-bearing quartz conglomerate beds (called blankets or reefs) of the Witwatersrand system in the Transvaal and Orange Free State in South Africa. The mineral most commonly associated with gold, other than quartz, is pyrite, FeS_2, or "fool's gold," the yellow disulfide of iron.

The gold mined during great gold strikes in California, Colorado, and Alaska was removed by the placer method (see below *Placer mining*) from alluvial deposits in which metallic gold occurred in the form of gold dust, larger grains, irregularly shaped masses, or occasional nuggets dispersed through the sand or caught in rock crevices.

For the most part, the gold found in rock is invisible to the naked eye, although it does sometimes occur as grains or flakes large enough to be seen, and more rarely as specimen rock. Crystals an inch or more across have been found in alluvial deposits in California. Boulders of pure gold were found in the early days of the Australian gold rush of 1851. Compounds formed with the element tellurium, known as gold tellurides, are found principally in Western Australia and Colorado: the principal ores are calaverite, a bronze-yellow telluride that contains 40 percent gold, and sylvanite, steel gray in colour and containing up to 28 percent gold combined with some silver.

Gold is also found in seawater. Though the total amount in the oceans is estimated at billions of tons, the concentration of it is less than six parts per 1,000,000,000,000 parts of seawater, making it not economically feasible to mine the sea.

Alluvial deposits of native gold found in or along streams were the principal sources of the metal for the ancient civilizations of the Middle East. The washing of gold ores is depicted on Egyptian monuments of the 1st dynasty (*c.* 3100–*c.* 2890 BC). The famous legend of the Golden Fleece was based on an expedition (about 1200 BC) to seize gold that was washed out of the river sands with the aid of sheepskins in the region later known as Armenia. Rich deposits were known in Lydia and the lands of the Aegean, and in Persia (now Iran), India, China, and other lands.

During the Middle Ages the chief sources of gold in Europe were the mines of Saxony and Austria; Spain also produced some gold.

The era of gold production that followed the discovery of the Americas was in all probability the greatest the world had witnessed to that time. The exploitation of mines by slave labour and the looting of palaces, temples, and graves in Central and South America resulted in an influx of gold that literally unbalanced the economic structure of Europe and disturbed its political structure. From the discovery of America by Columbus in 1492 until 1600, more than 8,000,000 ounces (225,000 kilograms) of gold, or 35 percent of world production, came from South America. South American mines—especially in Colombia—continued into the 17th and 18th centuries to account for 61 and 80 percent, respectively, of world production; 48,000,000 ounces (1,350,000 kilograms) were mined in the 18th century.



Miners panning for gold in the Klondike, Yukon Territory, 1897.
The Bettman Archive

Gold in Central and South America

Russia became the leading producer in 1823, and for 14 years contributed the bulk of the world supply.

During the second great era of expanding production (1850–75) more gold was produced in the world than in all the years since 1492, chiefly because of discoveries in California and Australia. A third marked increase in world gold recovery (1890–1915) stemmed from discoveries in Alaska, Yukon Territory, and South Africa. Beginning in the 1920s, gold production was increased by the development of gold fields in Canada. A major factor in the increase of the world supply of gold was the introduction in 1890 of the cyanide process for recovery of gold from low-grade ores and ores containing minute, particle-size gold.

Throughout the years gold production rose until the average yearly production reached almost 52,000,000 ounces (1,500,000 kilograms) in 1970, which was greater than the totals for either 1493–1600 or 1600–1700.

MINING

Gold is recovered by placer mining of alluvial deposits, and by lode or vein mining. It is also recovered as a by-product of base-metal mining. Placer mining, the oldest method, entails exploiting the high density (or heaviness) of gold to separate it from the much lighter siliceous material with which it is found. The alluvial deposits mined by placer methods are the gold-bearing sands and gravel that have been deposited by rapidly moving streams and rivers at places where they widen or for some other reason lose speed. As the current slows, the sediment being carried downstream settles to the bottom.

Placer mining. Although the basic principles of placer mining have not altered since early times, the methods have improved considerably, chiefly in mechanical procedures. In the great American gold strikes in California, Colorado, and Alaska, placer mines were almost exclusively the source of gold, and the panning method was the technique utilized by the individual miners. The miner used a pan or a batea (a pan or basin with radial corrugations) in which he placed a few handfuls of dirt and a large amount of water. Swirling the pan washed the siliceous material over the side and left the denser material in the centre of the pan. After many washings,

Panning

only gold and the other heavy minerals were left. At this point, if the gold particles were large enough, it was comparatively easy to separate them from other materials. If not, further concentration was needed.

The cradle or rocker was an improvement over the pan and the batea. The cradle, named for a child's cradle, which it resembled, could process larger quantities of ore. Gravel was shovelled onto a perforated iron plate, and water was poured on. The finer material dropped onto the apron, which distributed it across riffles, pieces of wood or iron perpendicularly fixed to the bottom and sides of the cradle. The entire apparatus was rocked, and as the material moved through the rocker, the gold was caught by the riffles. When enough gold was accumulated, the riffles were cleaned.

Although the cradle or rocker is largely obsolete, its riffles remain in use in sluice boxes and corduroy tables. The name riffle is applied to any strip, bar, or groove placed at right angles to the flowing stream to provide a protective spot where the gold can settle. The corduroy table consists of wide, sloping plates with shallow sides that hold a coarse corduroy cloth. Periodically, the corduroy is removed and washed by hand in boxes partly filled with water to recover the gold-rich concentrates.

Hydraulic mining. In California, thick beds of gravel on the hillsides were worked by hydraulic mining. Powerful jets of water at pressures of hundreds of pounds per square inch were passed through giant swivel-mounted nozzles to break down the gravel banks and wash the material through lines of sluices. Although great volumes of water and many miles of pipes and flumes were required, the cost of treatment was only a few cents per cubic yard, which made it possible to work even poor ground at a profit. The millions of tons of tailings (discarded residue) that were washed into the Yuba and Feather rivers, however, had such an adverse effect on farming downstream that an injunction was obtained against hydraulic miners in 1880, and the work thereafter was strictly limited.

Dredging. In the early 1900s, dredging became the most important type of placer mining and has become probably the single most prevalent technique. The dredge generally used the world over is similar to that employed to deepen harbours and rivers. Invented in New Zealand, the gold-mining dredge achieved its greatest popularity on the rivers there and in California. Dredging is the major technique used today in the U.S.S.R.

Paddock dredging, a later development in the western United States, makes it possible to work placer deposits even if they are not in or near a riverbed. The paddock dredge floats in a pond that is continuously extended by the digging equipment at one end of the dredge while simultaneously being filled in by the waste or tailings at the other end. In this way, the dredge moves across country taking its pond or reservoir with it. By piling more gravel around the reservoir and increasing the water level, the dredge can be made to work its way uphill. In 1910 there were 72 operating dredges in California. After World War II, however, only a few remained in operation.

Lode mining. Many of the methods used in the underground mining and exploration of gold lode or vein deposits are similar to the shaft- and pit-mining methods used for other metals. Tremendous tonnages of gold ore are treated throughout the world since most gold-mine ores contain an extremely low percentage of gold. For example, in one area three tons of ore are processed for every ounce of gold obtained. Ordinarily, a vertical shaft is sunk to allow the miner to get to the lode. Horizontal working levels then follow the vein. In the Homestake mine in South Dakota (the largest in the U.S.), the working depth lies 6,800 feet (2 kilometres) below the surface, and since 1877, some 200 horizontal miles (320 kilometres) have been worked.

RECOVERY AND REFINING

Gold particles wetted by mercury adhere to each other and to copper plates that have been coated with mercury, a process called amalgamation. It forms the basis for a technique for recovering gold from ore. For many centuries this was the method used in treating massive ores. The percentage of gold recovered varied with the type of ores, and considerable gold was lost. Amalgamation

Cyanide process. The cyanide process, introduced in South Africa in 1890, effected a vast improvement over amalgamation and other earlier methods, and has been extensively used ever since. In this process, the gold in finely ground ore is dissolved by treating it with a very dilute solution of sodium cyanide or, the less expensive calcium cyanide plus lime and oxygen from air. The mixture is held for some hours in large tanks equipped with agitators. The chemical reaction yields a water solution of gold cyanide and sodium cyanoaurite. This solution of gold is treated to remove oxygen, then clarified and mixed with zinc dust to precipitate the gold and the other metals, such as silver and copper, that were dissolved by the cyanide. The precipitate is then treated with dilute sulfuric acid to dissolve residual zinc plus most of the copper. The residue is washed, dried, and melted with fluxes (materials used to promote fusion of the gold and silver and to dissolve the remaining copper). The operation may be repeated to flux off more base metal. The remaining gold and silver alloy, called doré, is then cast into moulds for assay.

The cyanide process may constitute the whole recovery process or may follow amalgamation. At some mines in the United States amalgamation is used to recover about 60 percent of the gold, after which cyanidation achieves an overall recovery of 95 to 96 percent. The cyanide process has been used to treat the tailings from early operations and fortunes have been made in reworking old dumps. The process is employed almost exclusively in recovering gold from the South African mines, the world's largest source in the 1970s.

Some South African mines reach depths of at least two miles. In contrast, open-pit mining has opened in a new region, northeastern Nevada. The Carlin mine there is the second-largest U.S. producer. Deposits there are sedimentary and originally contained some organic material, which prevented satisfactory cyanide extractions. Weathering, however, destroys the organic material and makes the use of cyanide effective. An electrolytic method has been devised for oxidizing the deeper unoxidized deposits and destroying the troublesome organic material so that they ultimately can be worked with the cyanide process.

The dominant position of South Africa in world production of gold (see Table) is a tribute to the ability of modern mining methods to cope with the tremendous rock pressures and high temperatures encountered at great depths.

In addition to the gold from placers and hard rock mines, considerable gold is recovered from copper and copper nickel ores in the course of electrolytic refining, a process in which a metal is refined by being deposited electrolytically much as in electroplating.

Recovery from scrap. Considerable amounts of gold are recovered from scrap. The composition of scrap varies widely and dictates the process to be used. The scrap is largely produced in the manufacture of gold jewelry and electrical contacts, and as electroplating residues. Melting the scrap under oxidizing conditions will volatilize some of the zinc and remove some of the iron in the slag. The metal may then be formed into pellets and if the gold content is about 20 percent it can be treated with nitric acid to remove the base metals. If the pellets are low in lead, treatment with hot sulfuric acid may suffice. Retreatment will be required, however, to yield gold of acceptable purity if the sulfuric acid process is used. These methods are appropriate for scrap that is free of platinum metals. Lower-grade scrap may be melted and cast into anodes that are electrolyzed in a sulfate solution. The gold will remain in the anode mud and can be recovered by first converting it to gold chloride and then precipitating the gold by adding ferrous sulfate.

World Gold Production
(in 000 ounces)

	1965	1967	1969	1970*
South Africa	33,521	33,520	34,314	35,289
U.S.S.R.†	5,518	6,253	6,857	7,131
Canada	3,956	3,250	2,792	2,566
U.S.	1,871	1,738	1,901	1,913
Ghana	828	837	775	772
Australia	963	879	767	677
Philippines	478	539	627	661
Rhodesia	597	549	527†	549†
Japan	291	278	270	281
Colombia	350	283	240	221
Mexico	237	181	198	198
Brazil	170	189	194	198
Zaire (Congo [Kinshasa])	99	169	193	194
Korea, North†	176	176	176	176
Nicaragua	217	195	132	126
Peru	115	105	144	114
India	144	112	120	114
Fiji	120	122	100	114
Total‡	49,651	49,375	51,046	51,956

*Preliminary. †Estimate. ‡Includes other minor producers.
Sources: *Annual Bullion Review* (1968), Samuel Montague & Co.; National Materials Advisory Board Report 254, *Trends in the Usage of Gold* (1969), by permission; U.S. Bureau of Mines, *1970 Minerals Yearbook*.

Wohlwill process

Refining. In the Wohlwill process, doré gold from the usual mining and smelting operations is electrolyzed in a chloride solution using direct plus alternating currents. The gold in the doré anode dissolves and is deposited on the cathode as very pure gold. The silver is converted into chloride of which some clings to the anode and must be removed from time to time. This material also contains considerable gold and must be reworked. Any platinum or palladium in the anode will dissolve and is recovered by treating the electrolyte. This extensively used process yields gold at least 99.95 percent pure.

Because the Wohlwill process is rather slow, and for other reasons, it has largely been replaced by the Miller process in which chlorine is bubbled through the molten doré, converting the base metals into chlorides, some of which volatilize. The silver also is converted to silver chloride, which is molten and can be poured off and recovered. After this treatment, the remaining gold usually has a purity of 99.5 percent or above.

ASSAYING

Fire assay. Assaying, a process for determining the content of gold in ores, scrap, and refinery products demands great skill in collecting a proper sample as well as in carrying out all the subsequent steps. A sample of an ore, for example, must be finely ground, then mixed with a flux consisting of soda, borax, silica (if needed), lead oxide (litharge), and flour to reduce the lead oxide to metallic lead during fusion by heat. The reduced lead picks up the precious metals and settles to the bottom. After solidification the lead button is placed in a cupel (small porous cup) made of bone ash and heated in a special furnace provided with a strongly oxidizing atmosphere. The lead oxidizes and the resulting litharge soaks into the cupel.

When the oxidation is complete the appearance of the remaining metallic bead changes and the cupel is removed from the furnace and allowed to cool slowly. The bead, which consists of gold and silver, is removed, cleaned, hammered, and rolled into a thin strip, which is treated with nitric acid to remove the silver, an operation called parting. For proper parting, the silver content of the bead should be about three times that of the gold.

The gold strip, which should remain intact, is washed, heated, and weighed as gold. If platinum metals are present, they will enter the bead and alter its appearance.

Examining arc spectrum

An examination of the arc spectrum of the bead will give both some clue to its content of platinum metals and guidance in selecting a method for further detailed analysis. If osmium is present initially, it may be lost during oxidation in the cupel and the wet analysis must begin with the lead button. Fire assay is excellent for laboratories that are set up to carry it out on a routine basis, but for occasional use in general laboratories it is quite unsuitable.

Assay of low-grade ores. A method has been devised for the accurate assay of low-grade gold ores. In it, the ore is ground fine enough to pass through a screen woven of 100 wires per inch, heated at 550° C to destroy organic matter, dissolved in a mixture of nitric and hydrochloric acids, then dried to a paste. Slightly diluted hydrochloric acid is then added, followed with a measured amount of 4-methyl-2-pentanone, that extracts the gold chloride. After this product is shaken and allowed to separate, small amounts of the ketone, which carry the gold, are fed into an ethyne air burner and the amount of gold determined by atomic absorption measurements.

Gold compounds. Since gold is not attacked by any single acid and does not react with oxygen, sulfur, or the dry halogens (chlorine, bromine, fluorine), it remains bright after outdoor exposure. It is attacked by wet halogens and particularly by a 3:1 mixture of hydrochloric and nitric acids called aqua regia. It is also attacked when it becomes the positive electrode of an electrolytic cell in a hydrochloric acid solution. This fact is the basis for an electrolytic refining process.

The gold chlorine compounds $AuCl_3$, AuCl, and $HAuCl_4$ are involved in the electrolytic refining of gold. When gold is dissolved in aqua regia, $HAuCl_4$ is produced, and gold can be precipitated, after evaporation and resolution, by adding sulfur dioxide, formic acid, or other reductants. If palladium is present in the chloride solution, however, great care is necessary to insure complete precipitation of the gold to avoid the production of the explosive fulminating gold during the later recovery of palladium.

Potassium cyanoaurate $K[Au(CN)_2]$ is the basis for most gold plating baths (the solution employed when gold is electroplated) and is best made by passing current from a gold anode in a potassium cyanide solution to a cathode within a porous pot filled with a potassium hydroxide solution.

Gold forms a host of other inorganic compounds, most of them of little practical importance. Likewise, a large number of organic gold compounds exist. Only two will be mentioned here. The first forms the base of the "liquid bright gold" and is made by reacting gold chloride with a sulfurized terpene. The resulting compound is dissolved in certain organic solvents, mixed with fluxing materials and a trace of rhodium, and then applied to glass or ceramics. After drying and firing, a bright film of gold will remain, well bonded to the glass.

Organic gold compounds

The second, disodium aurothiomalate, is used in the treatment of rheumatoid arthritis. The compound is administered in small intramuscular injections over a period of 24 weeks. Careful supervision of the patient is required because the amount of gold needed for a useful result is close to the point where undesirable side effects result.

USES OF GOLD

Jewelry remains the largest single use for gold in all parts of the world. Monetary use is rare, but industrial applications are increasing, especially in electronics and related industries. In some countries, considerable amounts are used in dentistry.

Most gold used in jewelry is alloyed with silver, copper, and a little zinc to produce various shades of yellow gold, or with nickel, copper, and zinc to produce white gold. Most of this gold is of 14-karat quality, the karat representing a measure of purity in a scale of 1 to 24; thus 14-karat gold contains 14/24 or 58.35 percent gold. Considerable amount of jewelry is gold clad—that is, coated over with a thin cover of gold and is called "gold filled" or "rolled gold plate." Varying international designations describe the weight of gold cladding in accordance with government or industry regulations.

Electronics. In the electrical and electronics industries, gold is applied as a finish to electrical connectors, and as a thin coating over palladium in telephone relay con-

tacts. It is also used in the manufacture of transistors and microelectronic assemblies, and to protect etched circuit boards during storage. Large quantities of less expensive jewelry are finished with an electrodeposit of nearly pure gold.

Dental use. Dental applications include wrought and cast gold alloys with a gold content of 60 percent or more along with silver, copper, and sometimes platinum; the rest of the alloy consists of palladium. Some of these dental alloys can be hardened by heat treatment; all possess good strength and a high resistance to corrosion. Gold solders also are used in some industrial fields in which strength plus corrosion resistance are required.

Other uses. Small amounts of gold are used in decorating china and glassware and to provide heat-reflecting surfaces for window glass. Extremely thin coatings on glass with a metal cost of about five cents per square foot will reflect most of the sun's heat rays, but will transmit a reasonable amount of visible light. Some gold goes into heat-reflecting shields in aircraft, and small quantities find applications in the medical and chemical fields. Gold leaf is used for decorative purposes.

World use of gold. Gold is used in so many ways and circulates through so many channels that it is difficult to determine amounts and end uses. The mode of use differs sharply in the various countries.

The total of fabricated gold (excluding the U.S.S.R.) about equals the amount of gold mined (about 85 percent in 1968). The total amount of gold released, including newly mined gold and that from banking and other international operations, is about 61,600,000 ounces (1,700,000 kilograms), however. The difference between mine production and gold released amounted to nearly 20,000,000 ounces (600,000 kilograms) in 1968 and comprised net hoarding and speculative buying, as well as errors in estimating consumption in the various countries.

In 1968, the United States consumed about 16 percent of the world's fabricated gold, Europe 41 percent, and India and Pakistan about 14 percent, almost entirely in jewelry. The United States was the largest user for electronics and dentistry.

BIBLIOGRAPHY. Illustrations of the ancient artistic uses of gold and the history of the principal sources are well treated by C.H.V. SUTHERLAND in *Gold: Its Beauty, Power and Allure*, 2nd rev. ed. (1969). The discovery and development of the South African sources is detailed by ERIC ROSENTHAL in *Gold, Gold, Gold* (1970). A new U.S. source is described in *Gold Resources in the Oxidized Ores and Carbonaceous Sediments —Northeastern Nevada*, Bureau of Mines Heavy Metals Program Progress Report (March 1968). The properties of the alloys of gold and their major applications are covered by E.M. WISE in "Gold and its Compounds," *Encyclopedia of Chemical Technology*, 2nd ed., vol. 10 (1966), and in *Gold: Recovery, Properties, and Applications*, ed. by WISE (1964). For information on the uses of gold in the U.S., see *Trends in the Usage of Gold*, Report of the National Materials Advisory Board, Pub. NMAB 254 (1969). An estimate of the mode of consumption of gold throughout the world has been made by D.O. LLOYD-JACOB and P.D. FELLS in *Gold* (1969). A tabulation of the important physical properties of pure gold is given in J. PENTON, *Properties of High Purity Noble Metals* (1965). The chemistry of gold and silver is presented by J.W. LAIST, "Copper, Silver and Gold," in *Comprehensive Inorganic Chemistry*, vol. 2 (1954). Analytical methods are summarized in E.E. BUGBEE, *Textbook of Fire Assaying*, 3rd ed. (1940), and in F.E. BEAMISH, *The Analytical Chemistry of the Noble Metals* (1966).

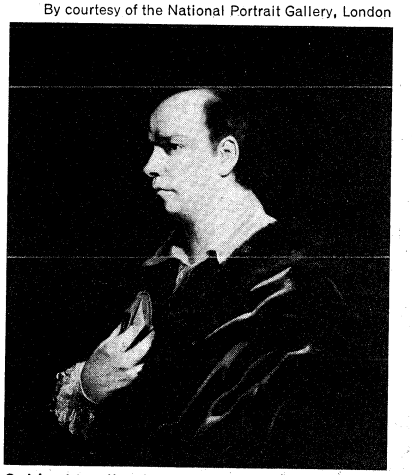
(E.M.W.)

Goldsmith, Oliver

When Oliver Goldsmith died unexpectedly in April 1774 at the approximate age of 43, he had achieved eminence among the writers of his time as an essayist, a poet, and a dramatist. He was one "who left scarcely any kind of writing untouched and who touched nothing that he did not adorn"—such was the judgment expressed by his friend Dr. Johnson in the Latin epitaph on the monument placed in the Poets' Corner of Westminster Abbey by Goldsmith's friends. It is a judgment that over the years has remained undisputed. His contemporaries were as one in their high regard for Goldsmith the writer, but they were of different minds concerning the man himself. He was, they all agreed, one of the oddest personalities of his time. Born in Ireland of established Anglo-Irish stock, he kept his brogue and his provincial manners in the midst of the sophisticated Londoners among whom he moved. His bearing was undistinguished, and he was unattractive physically—ugly, some called him—with ill-proportioned features and a pock-marked face. He was a poor manager of his own affairs and an inveterate gambler, wildly extravagant when in funds, generous sometimes beyond his means to people in distress. The graceful fluency with words that he commanded as a writer deserted him totally when he was in society—his conversational mishaps were memorable things. Instances were cited of his incredible vanity, of his constant desire to be conspicuous in company, of his envy of others' achievements. "Poor Doctor Goldsmith!" wrote Johnson's friend Mrs. Thrale (afterward Mrs. Piozzi), who had no high regard for him: "Lord bless us what an anomalous Character was his. . . ." According to the writer Horace Walpole he was "an inspired idiot." Johnson's biographer, James Boswell, described him as an *étourdi*—a giddy scatterbrain. Yet there were those who held him in deep affection, remembering him for his great personal charm. The painter Sir Joshua Reynolds, one of Goldsmith's closest friends, observed that a genius such as he "could not be a fool or such a weak man as many people thought him." In the end what most impressed Goldsmith's contemporaries was the paradox he presented to the world: on the one hand the assured and polished literary artist, on the other the person notorious for his ineptitudes in and out of society. Again it was Johnson who summed up the common sentiment. "No man," he declared, "was more foolish when he had not a pen in his hand, or more wise when he had."

Life and works. Goldsmith was the son of an Anglo-Irish clergyman, the Rev. Charles Goldsmith, curate in

By courtesy of the National Portrait Gallery, London



Goldsmith, oil painting from the studio of Sir Joshua Reynolds, 1770. In the National Portrait Gallery, London.

charge of Kilkenny West, County Westmeath. At about the time of his birth—November 10, 1730, is the date now generally accepted, though the precise records are lacking—the family moved into a substantial house at nearby Lissoy, where Oliver spent his childhood. Much has been recorded concerning his youth, his unhappy years as an undergraduate at Trinity College, Dublin, where he received the B.A. degree in February 1749, and his many misadventures before he left Ireland in the autumn of 1752 to study in the medical school at Edinburgh. His father was now dead, but several of his relations had undertaken to support him in his pursuit of a medical degree. Later on, in London, he came to be known as Dr. Goldsmith—Doctor being the courtesy title for one who held the Bachelor of Medicine—but he took no degree while at Edinburgh nor, so far as anyone

knows, during the two-year period when, despite his meagre funds, which were eventually exhausted, he somehow managed to make his way through Europe. The first period of his life ended with his arrival in London, bedraggled and penniless, early in 1756.

Rise from obscurity

Goldsmith's rise from total obscurity was a matter of only a few years. An usher (assistant schoolmaster) in a boys' school, then a literary hack writing book reviews for magazines, he became an author in his own right with the publication, in April 1759, of his first book, *An Enquiry into the Present State of Polite Learning in Europe*. He scored his first triumph with an amusing series of essays appearing during 1760–61 in a newspaper, *The Public Ledger*, and originally known as the *Chinese Letters* and later, in collected form, as *The Citizen of the World*. By now he was well on the way to recognition. He had been introduced to Johnson, foremost among the men of letters; he was already on friendly terms with Joshua Reynolds, not yet knighted but firmly established as London's most fashionable portrait painter; and on Christmas Day 1762 he and the young Boswell met for the first time. With the publication of his poem *The Traveller* (December 1764), Goldsmith came fully into his own. He was one of the nine original members of the famous Club founded at this time, and his presence there, his conversational encounters with Johnson and others, his foolishness and his wit have been preserved for all time in Boswell's *Life of Samuel Johnson*. Meanwhile, he was extending his reputation as an author in the fictional field with *The Vicar of Wakefield* (1766); in poetry, with *The Deserted Village* (1770); in comedy for the theatre, with his first play, *The Good Natur'd Man* (first performed 1768), and his dramatic triumph, *She Stoops to Conquer* (1773).

His income as a professional writer was now considerable, but his extravagances kept him in debt and to maintain himself he accepted all sorts of commissions from the publishers for books of a strictly popular nature—biographies, histories, and the work he was finishing at the time of his death, *An History of the Earth, and Animated Nature*. He lived for some time in handsomely furnished rooms at No. 2, Brick Court, in the Temple; and it was here that he died after a short illness on April 4, 1774. He was buried privately in the Temple burying ground; and at the last moment the coffin was opened so that a lock of his hair might be obtained in compliance with the belated request of Mary Horneck, one of the two sisters, still in their youth and well-known in London society for their charm and beauty, to whom Goldsmith had been a devoted friend. He left behind him several works of a minor character, together with a masterly but unfinished set of verses, *Retaliation*, in which he took off some of his famous friends and acquaintances—the statesman and writer Edmund Burke, for one, and the actor David Garrick and Reynolds—in a way that was at once witty, penetrating, and good-natured.

Early death

Reputation. Goldsmith's world—the world of Georgian London during the 1750s and 1760s—was one in which men of achievement rapidly became well-known personalities, to be commented upon then and thereafter in newspapers, magazines, letters, diaries, memoirs, and in biographies by those who held them in high regard, by those who did not, and by the merely curious. About Goldsmith, too much has been told and not enough—too much that is trivial, too little that bears on the essential qualities of his genius. It is, for instance, obvious enough that Goldsmith commanded nothing like Johnson's ordered and encyclopaedic knowledge; but it is not true, as his contemporaries liked to believe, that he was intellectually shallow, a writer whose charm and ease of manner compensated for a lack of ideas. Read today, he seems to be one who had a keener insight into the forces shaping later 18th-century civilization both in England and abroad than did some of his eminent friends. He hated the sentimentalism that had taken root in literature. He looked with apprehension on the increasingly materialistic temper of England. He found nothing to glorify in the military exploits in foreign lands that were laying the foundations of a British Empire. An admirer of the French writers Montesquieu and Voltaire and taking from them something of a sociological approach that was more French than English, he was an observer of European culture, commenting on national differences in manners and taste and seeking for common norms. His writings that most clearly embody these ideas in one fashion or another are the *Enquiry into the Present State of Polite Learning* and the two long poems *The Traveller* and *The Deserted Village*. Goldsmith best deserves to be remembered not only as a poet but as the writer whose comic sense found memorable expression in *The Citizen of the World*, *The Vicar of Wakefield*, and, dramatically, in *She Stoops to Conquer*. Here again he was often mistakenly described as one who worked more or less without a plan, achieving his happiest effects on the spur of the moment. But planning there was—a great deal of it in the calculated action of *She Stoops to Conquer* and more than was generally perceived in *The Vicar of Wakefield*, with its sly and skillfully controlled ambiguities. Goldsmith saw people, human situations, and indeed the human predicament from the comic point of view; he was a realist, something of a satirist, but in his final judgments unfailingly charitable.

Goldsmith's comic sense

MAJOR WORKS

An Enquiry into the Present State of Polite Learning in Europe (1759); *The Bee* (1759), a collection of essays originally published in the periodical of the same name; *The Citizen of the World; or, Letters from a Chinese Philosopher, Residing in London, to his Friends in the East*, 2 vol. (1762); *The Life of Richard Nash, of Bath, Esq.* (1762); *An History of England in a Series of Letters from a Nobleman to His Son*, 2 vol. (1764); *The Traveller, or a Prospect of Society* (1764), verse; *The Vicar of Wakefield: A Tale*, 2 vol. (1766); *The Good Natur'd Man: A Comedy* (1768); *The Deserted Village* (1770), verse; *She Stoops to Conquer: or, The Mistakes of a Night. A Comedy* (1773); *Retaliation. A Poem* (1774); *An History of the Earth, and Animated Nature*, 8 vol. (1774).

BIBLIOGRAPHY. The Goldsmith bibliography in *The New Cambridge Bibliography of English Literature*, vol. 2 (1971), covers both Goldsmith's writings and Goldsmith criticism through 1969. RICARDO QUINTANA, *Oliver Goldsmith: A Georgian Study* (1967), contains bibliographical notes in the appendix. There is a bibliography of Goldsmith first editions in I.A. WILLIAMS, *Seven XVIIIth Century Bibliographies* (1924). Still of interest is TEMPLE SCOTT, *Oliver Goldsmith Bibliographically and Biographically Considered* (1928), based on the material in the W.M. Elkins collection now in the Free Library of Philadelphia. On the accessibility of Goldsmith manuscripts, see K.C. BALDERSTON, *A Census of the Manuscripts of Oliver Goldsmith* (1926).

Major collected editions, editions of separate works, and selections: *The Collected Works of Oliver Goldsmith*, ed. by ARTHUR FRIEDMAN, 5 vol. (1966), supersedes all other collected editions; those by PETER CUNNINGHAM, 4 vol. (1854), and J.W.M. GIBBS, 5 vol. (1884–86), though now outdated, still possess a limited usefulness. *The Collected Letters of Oliver Goldsmith*, ed. by K.C. BALDERSTON (1928), is standard. All of the poetry, together with full notes and extensive commentary, is given in *The Poems of Thomas Gray, William Collins, and Oliver Goldsmith*, ed. by R.H. LONSDALE (1969). *The Complete Poetical Works of Oliver Goldsmith*, ed. by A. DOBSON (1906), is still valuable for its commentary and notes, as is the edition of the plays by A. DOBSON and G.P. BAKER, *The Good Natur'd Man and She Stoops to Conquer* (1905). Of the many editions of *The Vicar of Wakefield* those by OSWALD DOUGHTY (1928) and F.W. HILLES (1951) are outstanding by reason of their critical introductions. Selections are numerous; those of more recent date include *Goldsmith: Selected Works*, by RICHARD GARNETT (1950); *Oliver Goldsmith: The Vicar of Wakefield, and Other Writings*, by F.W. HILLES (1955); *A Goldsmith Selection*, by A.N. JEFFARES (1963); *Selected Poems of Samuel Johnson and Oliver Goldsmith*, by ALAN RUDRUM and PETER DIXON (1965).

Biography and criticism: The authoritative biography is R.M. WARDLE, *Oliver Goldsmith* (1957). Significant older biographical works are SIR JAMES PRIOR, *The Life*, 2 vol. (1837), and JOHN FORSTER, *The Life and Times of Oliver Goldsmith*, 2nd ed., 2 vol. (1854). Of the many treatments of Goldsmith by contemporaries of his, the more important are these: that by BOSWELL in his *Life of Samuel Johnson*, 2 vol. (1791); the prose portrait by SIR JOSHUA REYNOLDS in *Portraits*, ed. by F.W. HILLES (1952); and MRS. THRALE's remarks in her *Anecdotes of Johnson* (1786) and in her diary, *Thrali-*

ana, ed. by K.C. BALDERSTON, 2nd ed. (1951). Recent critical studies include C.M. KIRK, *Oliver Goldsmith* (1967); RICARDO QUINTANA, *Oliver Goldsmith: A Georgian Study* (1967); and R.H. HOPKINS, *The True Genius of Oliver Goldsmith* (1969). More specialized is KARL EICHENBERGER, *Oliver Goldsmith: Das Komische in den Werken seiner Reifeperiode* (1954). A.N. JEFFARES, *Oliver Goldsmith* (1959), is brief but informative. Studies of individual writings include J.H. PITMAN, *Goldsmith's Animated Nature: A Study of Goldsmith* (1924); H.J. SMITH, *Goldsmith's The Citizen of the World: A Study* (1926); and MACDONALD EMSLIE, *Goldsmith: The Vicar of Wakefield* (1963); and SVEN BACKMAN, *This Singular Tale: A Study of the Vicar of Wakefield and Its Literary Background* (1971).

(R.Q.)

Golf

Golf is a cross-country game played by striking a small ball with various clubs from a series of teeing grounds into a like series of holes on a course. The player who holes his ball in the fewest strokes is the winner. The game developed originally in Scotland and has spread from obscure antiquity to worldwide popularity. Played in the beginning on seaside links with their crisp turf and natural hazards and afterward on downs, moorland, and parkland courses of various lengths and individual characteristics, it combines with its open air and exercise an intrinsic fascination. Its devotees—men, women, and children—now number millions around the world. A few examples of the numbers of players include, in the United States, 10,000,000; Australia, about 2,000,000; the British Isles, approximately 600,000; Japan and Canada, 50,000 each. The 10,000th course in the United States was opened in 1970.

This article is divided into the following sections:

I. History
- Origins and early history
  - Continental games
  - Scotland
- Development of golfers' associations
  - Early British societies
  - The United States and Canada
  - Other countries
- Development of equipment
  - Golf balls
  - Golf clubs
- New manufacturing methods
- Players and championships
  - The premier championships
  - Ascendancy of U.S. players
  - The international scene
  - Senior golf

II. The modern game
- Courses and forms of play
  - Match and medal play
  - Handicaps
  - Par golf
- Equipment
  - Golf balls
  - Golf clubs
- The rules of golf
  - Local rules
  - Definition of amateur
- Play of the game
- Variants
  - Par-three golf
  - Driving ranges
  - Putting games
- New trends

I. History

ORIGINS AND EARLY HISTORY

Controversy over origin of the game

The game's origin has long been a subject of controversy. Broadly, the issue has been whether its natal source is Scotland or Holland. Its antiquity in Scotland is beyond question, though the problem of tracing its unknown roots was summed up by the noted Scots essayist Andrew Lang who contended that its exhaustive history would necessitate a thorough study of all Scots Acts of Parliament, Kirk Session records, and many other sources; and it was essential that it be undertaken by a young man as he would be aged at the end of the task. The Scottish claim has at any rate documentary proof of its long lineage; for instance, in the oft-quoted statute of the Parliament of King James II of 1457 that decreed that both "Fute-ball and Golfe be utterly cryed downe" because they interfered with the practice of archery, which was necessary for the defense of the realm. Two other decrees to the same effect were also enacted, the third in 1491 in the reign of James IV, which indicted them as "unproffitable sportis." One obvious implication of these laws is that golf had become a popular obsession in Scotland even before the middle of the 15th century.

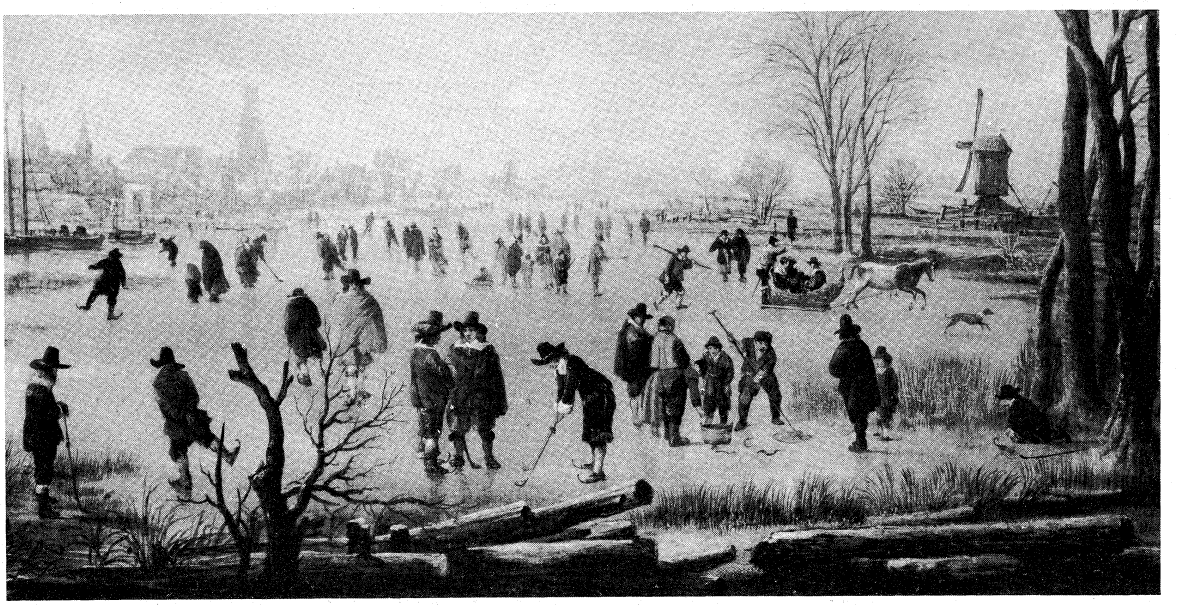
Continental games. Comparisons between ancient continental games and Scottish golf make an interesting study, with the balance of evidence in favour of Scotland. In the time of the Roman domination in Europe, a popular game was *paganica* (from *paganus*, "country man" or "rustic"), a cross-country pastime played with a bent stick and with, it seems, a ball of leather stuffed with feathers, as were early golf balls in Scotland; but it can be disregarded merely as what Scots golf historian Sir Guy Campbell called a link between legend and history. Other old-time games prevalent in the Netherlands have a more striking similarity to golf. Of these, the Dutch game of *kolven* or *kolf* has appealed to the imagination of some as being a source of golf partly because of the name *kolf* for a club and the fact that the clubs seen in early Dutch pictures suggest something like golf. Notwithstanding such evidence, there are obvious and vital differences between that game and golf. *Kolven* is not a cross-country game but is played in a walled court, and its objective is not holing out but the striking of posts either first or in the fewer strokes. According to delightfully scenic 17th-century Dutch paintings, *kolven* was also played on ice, but, in their inclusion of posts and the crowded state of the ice, which would be incompatible with the conditions needed for golf, the illustrations by no means suffice to clinch the Dutch-origin argument; also, the game had been played in Scotland for about two centuries before the pictures were painted.

The games of *kolven* and *chole*

Of the continental games, the venerable game of *chole* (*choulla* or *choulle*) had some resemblance to golf. More Belgian (Flemish) than Dutch, it was a cross-country pastime in which each side, whether one player or several, used one ball and played to a prearranged target—a church door, for example—at a considerable distance from the starting point. At the outset each player (or side) bid the number of strokes it would employ to reach the goal—subject to the proviso that after each three-stroke turn the opposing player (or side) had the right to hit the ball into any available hazard, the more difficult the better. Such interference with the play of an opponent is, of course, utterly alien to golf. Still, certain features of the game noticeable in old pictures—the types of club used (though only one to each player) and the teeing up of the ball—have suggested that golf may have been an offshoot of *chole*.

Scotland. Passing from conjecture to chronicle, Scotland's King James IV, who had embargoed the game in 1491, later became the first player of whom there is authentic record. The Lord High Treasurer's accounts early in the 16th century include payments for the King's equipment—"golf clubbis and ballis." The royal line of the Stuarts gave the game its first woman golfer, too, in the person of Mary, queen of Scots, who was charged with playing in the fields beside Seton only a few days after the murder of her husband, Darnley.

James VI was noted for his zest in playing and promoting the game. When he succeeded to the English throne as James I, he took his courtiers and his clubs south with him. He had in the meantime (1603) appointed William Mayne, a burgess bower (the bow maker also made golf clubs) of Edinburgh, to be the royal club maker for life, and some years later he enacted a protective ban on behalf of Scottish trade that projects an element of confusion into the debate on the origin of golf: he forbade the purchase of balls from Holland (on which Scots were spending "no small quantitie" of gold and silver) and assigned a monopoly of ball making in Scotland to James Melvill for 21 years. The inference of the words "no small quantitie" spent on the Dutch imports confirms



Dutchmen playing *kolven* on the ice. Detail from "Riverscape in Winter," oil painting by Aert van der Neer (b.c. 1603). In the Rijksmuseum, Amsterdam.
By courtesy of the Rijksmuseum, Amsterdam

the popularity of the game with the Scots, but, as some historians have pointed out, not necessarily that the game they were used for in the Netherlands was golf.

DEVELOPMENT OF GOLFERS' ASSOCIATIONS

Early British societies. There is a tradition that James I introduced golf to Blackheath in 1608, which has been given, erroneously according to some authorities, as the foundation year of the historic Royal Blackheath Golf Club. But, though no doubt King James and his courtiers played their golf somewhere in the vicinity, it is doubtful whether any organized societies then existed, and research has set its date nearly two centuries later. In the *Chronicles of Blackheath Golfers*, W.E. Hughes, editor of the *Chronicles*, ascribes the club's foundation to 1787.

The Honourable Company. The oldest club with documentary proof of its origin is the Company of Gentlemen Golfers, now the Honourable Company of Edinburgh Golfers, whose modern home is at Muirfield in East Lothian. Its genesis was a move by a group of players to hold a competition or tournament. In 1744 "several Gentlemen of Honour skillful in the ancient and healthfull exercise of Golf" petitioned the Edinburgh city council to provide a Silver Club for annual competition on the links of Leith, which the magistrates agreed to do. The winner was to be named "Captain of the Golf"; *i.e.*, captain of the Company of Gentlemen Golfers. Intended to be open to all comers, the continued absence of outside challengers resulted in the contest being limited to members only after 1764.

The Silver Club was first played for in 1745, and the first winner, thus the first captain of the company, was an Edinburgh surgeon, John Rattray, who also won in the following year and again in 1751. The earliest known rules of golf are the 13 recorded in the first minute book of the company under Rattray's captaincy. The code, therefore, cannot be more recent than 1751 and more likely was entered in the minute book when the competition for the Silver Club was instituted in 1745.

Earliest rules of golf

The Royal and Ancient Golf Club. The Society of St. Andrews, now the Royal and Ancient Golf Club of St. Andrews, Scotland, was formed in 1754 by a group of 22 golfers who played there. The rules that the society adopted were almost identical with the Edinburgh Gentlemen Golfers' rules. These two clubs played major roles in the development of the game in Scotland. Eventually, the Royal and Ancient Golf Club (R. and A.) became, by common consent, the oracle on rules. In 1919 it accepted the management of the British Open and Amateur championships. The R. and A. thus became the governing body for men's golf in the British Isles and throughout most of the Commonwealth.

Other cradles of history. The Royal Musselburgh Golf Club, a few miles from Edinburgh, derives from the presentation to local golfers in 1774 of a cup for competition, and the trophy is still played for. In the early days, as on other courses, money matches were the chief media of rivalry, and bitter feelings often were aroused among spectators. In 1925 the club moved from the old links to the enclosed Prestongrange Estate nearby.

At the ancient Bruntsfield Links in Edinburgh, within 50 yards of one of the capital's busiest traffic arteries, players from small schoolboys to keen octogenarians can be seen at the "gowff" where it held sway five centuries ago. Formerly a conventional course with holes needing full-length shots, time and circumstances have forced it into short-hole compass. The members of the Royal Burgess Golfing Society were staunch conservators of the citizens' right to play there and for many years successfully resisted the intrusion of proposed quarries, new pathways, the beating of carpets, the training of horses, and the drilling of troops. Conditions did, however, eventually compel the members to leave; and in 1877 they migrated to Musselburgh, as did also the Bruntsfield Links Society and other clubs. With the encroachment of the city, the Leith golfers also had to desert their links. The city corporation, however, purchased the Braid Hills, where the game has since 1890 been extremely popular. The "Golfers' Tavern" on the edge of the links was refreshing golfers in 1717 and is still doing so. Musselburgh in time became congested. The Honourable Company moved down the coast in 1892 to Muirfield, and the Burgess Society and the Bruntsfield Society became close neighbours on the outskirts of Edinburgh.

Early clubs in England

With the birth of the Royal North Devon Club in 1864, the game took a firm foothold in England. It was the first course on seaside links outside Scotland. The Royal Liverpool Club was established in 1869 on a rabbit warren at Hoylake. In its infancy players simply cut holes with their penknives and stuck feathers in them for the guidance of those who were coming behind. The rabbits were the greenkeepers. By 1870 the club was fairly founded, and members played matches against Blackheath, Royal North Devon Club at Westward Ho!, and others. Their influence widened interest in the game. The Royal Liverpool Club hosted Great Britain's first Amateur Championship in 1885 and the first English Amateur Championship in 1925; the first Scotland–England amateur match was organized in 1902; and it was at Hoylake in 1921 that an unofficial contest between Brit-

ish and U.S. players, a curtain raiser to the Amateur Championship, was the genesis of the Walker Cup series.

The United States and Canada. The ancient Dutch game of *het kolven* apparently was played in the New World in the 17th and 18th centuries. The record of a court in Dutch-settled Fort Orange (now Albany, New York) reveals that the sheriff there filed a complaint against three men who had been playing the game on the ice on a Sunday in 1657, and in 1659 the magistrates of Fort Orange issued an ordinance to "forbid all persons to play 'het kolven' in the streets." While these documents are sometimes cited as evidence of the earliest playing of golf in the United States, the game does not seem to have been the linksland pastime handed down by the Scots. There seems little doubt, however, that the following advertisement, which appeared in James Rivington's *Gazette* in New York on April 21, 1779, referred to golf: "To the GOLF PLAYERS: The Season for this pleasant and healthy Exercise now advancing, Gentlemen may be furnished with excellent CLUBS and the veritable Caledonian BALLS, by enquiring at the Printer's."

Early records of golf in America

The next evidence appears in the Carolinas. The *South Carolina and Georgia Almanac* of 1793 published, under the heading "Societies Established in Charleston," the following item: "Golf Club Formed 1786. Dr. Purcell —President. Edward Penman—Vice President. James Gardner—Treasurer and Secretary." The *Charleston City Gazette and Daily Advertiser* of September 18, 1788, reported: "There is lately erected that pleasing and genteel amusement, the KOLF BAAN. Any person wishing to treat for the same at private sale will please apply to Mr. David Denoon in Charleston, or to the subscriber on the spot." Later notices dated 1791 and 1794 referred to the South Carolina Golf Club, which celebrated an anniversary with a dinner on Harleston's Green in the latter year. Although these fragments constitute the earliest clear evidence of golf clubs in the United States, the clubs appear to have been primarily social organizations that did not survive the War of 1812.

Early clubs and courses. The first permanent golf club in the Western Hemisphere was the Royal Montreal Golf Club, established in 1873. The members played on Fletcher's Fields in the city's central area until urban growth compelled a move of some miles to Dixie, a name derived from a group of Southern refugees who arrived there after the U.S. Civil War. The Royal Quebec Golf Club was founded in 1874; the Toronto and Niagara, Ontario, clubs in 1876; and the Brantford, Ontario, club in 1879. In the meantime, golf was played experimentally at many places in the United States without taking permanent root until, in 1885, it was played in Foxburg, Pennsylvania. The Oakhurst Golf Club in West Virginia, which later became the Greenbrier Club, is said to have been formed in 1884; and the Dorset Field Club in Dorset, Vermont, claims to have been organized and to have laid out its course in 1886, although in both instances written records are lacking. The Foxburg Golf Club has provided strong support for the claim that it was organized in 1887 and is the oldest golf club in the United States with a permanent existence. Foxburg also claims the oldest U.S. golf course. The course and club came into existence through the interest and generosity of Joseph Mickle Fox of Philadelphia, a summer resident of Foxburg who is believed to have been introduced to golf and to have acquired his first (left-handed) clubs and balls while in Scotland in 1884. The next oldest course may be that of the Middlesborough (Kentucky) club, which apparently was founded in 1889 by English immigrants. The course is still in existence, but there is a question as to whether play has been continuous on it.

Golf as an organized game. But golf as an organized game in the United States usually is dated from the founding of the St. Andrews Golf Club at Yonkers on the Hudson in 1888. As its name implies, its progenitors were Scots, John Reid and Robert Lockhart, both natives of Dunfermline in Fifeshire. Lockhart, a merchant in Yonkers, after a visit to his hometown, imported some golf clubs and balls and on February 22, 1888, tried them out with his friends Reid and John B. Upham on an improvised three-hole layout. At a meeting that fall, five men formed the club. Reid, who was its first president, has been dubbed "the father of American golf." Their original course was later extended to six holes on a cow pasture, and its hazards were numerous apple trees. In consequence, these pioneers have been known as the "Apple Tree Gang." Each player used six clubs, three woods and three irons, and a gutta-percha ball, and the group was the subject of the earliest known photograph of golf in America. The club moved several times but has been at Hastings-on-Hudson since 1897.

Other early courses included Newport, Rhode Island (1890); Shinnecock Hills on Long Island (1891), laid out by the Scottish professional Willie Dunn of the Royal Montreal Club; and the Chicago Golf Club (1892) at Wheaton. The Tuxedo Golf Club, New York, founded 1889, met the Shinnecock men in 1894 in what has been assumed to be the first interclub match in the United States. The Newport club staged an invitation tournament for amateurs in September 1894, and in October the St. Andrews Club promoted a similar competition. These were announced as championships, but that was questioned because the events were each promoted by a single club and on an invitational basis. It was from the controversy roused by these promotions that the United States Golf Association (USGA) was instituted in 1894. Its aims were to organize the U.S. Amateur and Open championships and to formulate a set of rules for the game. The founding fathers, two from each club, were from St. Andrews, Shinnecock Hills, Chicago, the Country Club at Brookline, and Newport. The following year saw the inaugural U.S. national championships—the Amateur, the Women's Amateur, and the Open.

First championships and the founding of the USGA

Other countries. Before organized clubs had been established in North America, the global extension of the game had already set in as colonies of British settlers, merchants, and civil servants carried their devotion to golf with them.

India can claim to have had the oldest club outside of Great Britain—the Royal Calcutta Golf Club founded in 1829, and the Royal Bombay Golf Club came some dozen years later. The Royal Calcutta initiated an amateur championship for India, and the two clubs paved the way for many in the Far East. The Royal Bangkok Golf Club (1890) was housed to begin with in an ancient temple. Golfers made their debut in China when the Shanghai Golf Club came into being in 1896, until which time the game was apparently unknown outside of Hong Kong. The Japanese were a few years later in their adoption of the game when a course was constructed at Kōbe. The Tokyo Golf Club was founded in 1914. With the boom in the popularity of the game in that country after World War II, addicts can now be numbered in the thousands.

The first club in Australia, the Royal Adelaide Golf Club, was formed in 1870, and it is believed that the game was planted in Melbourne in 1847 but went into abeyance for nearly half a century, the gold rush having taken priority over golf for the settlers.

New Zealand origins have been dated from the formation of the Christchurch Golf Club in 1873. South Africa's first golf shot was, according to some research, hit at the Maritzburg Golf Club, the first in Natal, in 1884; though the Royal Cape Golf Club (1885) has been rated as the country's senior club.

DEVELOPMENT OF EQUIPMENT

Golf balls. How the ball is hit and directed is the essence of the game. The result can be fun or frustration, delight or despair. The changing story of the ball's manufacture falls broadly into well-defined phases after it emerged from the unknown primitive, beginning with the "feathery," which was used for centuries until it was superseded by gutta-percha.

The feather-ball era. Some time in the past balls were made from wood, but in the early years of the 17th century the era of feather balls opened and was no doubt hailed for its advance on anything that was used previously. These balls consisted of boiled feathers that were

Manufacture of the feather ball

compressed through a hole left in pieces of stitched leather that composed the cover. For stuffing in the feathers a wooden tool was first used, after which the stuffing iron had to complete the job. When the leather case was crammed beyond increase, the hole was stitched up, the ball hammered and made as round as possible, and then painted white. The whole process was so slow that the maker did well to finish four in a day, so that they cost as much as five shillings each. Although the "feathery" could be hit a surprisingly long way, it became sodden and disabled in wet weather and was destroyed by hacks from iron clubs, so that the balls were short-lived as well as expensive.

The gutta-percha era. Gutta-percha is the evaporated milky juice or latex of various South American and South Pacific islands trees. It is soft and impressible at the temperature of boiling water, but becomes hard, non-brittle, and retains its shape when cooled. It is not affected by water except at boiling temperature. The emergence of the gutta-percha ball about 1848 brought a revolutionary change in the game. There have been differing views as to where and by whom it was invented. A substantial account narrates the arrival for R.A. Paterson, a professor in the University of St. Andrews, of a marble statue of the Hindu god Viṣṇu (Vishnu), which was packed in gutta-percha. A son of the consignee began experimenting with the gutta wrapping and made it into a rough ball, and, though his first trials with it were discouraging, he persisted with the experiment and sought the cooperation of his brother near Edinburgh, who was quick to size up its potential and improved it into a marketable product. But when it was displayed in London nobody wanted it. The professionals had divided views: St. Andrews' Allan Robertson, a leading manufacturer of "featheries," would have nothing to do with it; but "Old Tom" Morris, who was then his assistant, wisely foresaw the possibilities of the "gutty," and on this issue the two actually parted company in 1852, Morris going into business on his own. The ball was heartily welcomed by the golf community, not least for its economy—cost, one shilling each—and its coming immediately swelled the golfing ranks.

The first examples were smooth and ducked suddenly and went to the ground after a short flight, until they had been hacked and marked in the course of play. Consequently, the ball makers learned to nick them and then to mold them with raised or indented surface patterns so they would fly properly. The "feathery" was dead and gone.

Invention of the rubber ball

The rubber ball. The beginning of the 20th century introduced a new ball and a new era. It was ushered in with the American patent of the rubber ball, the invention of Coburn Haskell, a golfer of Cleveland, and Bertram G. Work of the B.F. Goodrich Company. It had a tension-wound rubber thread around a solid rubber core. Its resilience was a joyous advent after the "gutty," which was a hard, irresponsive ball that required skill and strength to get it into the air after the tee shots. The rubber ball was more congenial to hit and gave the striker great pleasure and a sense of power. Its impetus created a tidal wave in the game's popularity. Men who were getting on in years found it easier to play, and hosts of women and children were drawn into the fold. A number of varieties of ball appeared in the wake of the Haskell, such as the Kempshall, the Zodiac, the Spalding, the Dunlop, and the Silver King. The makers' main objective was to cater to the golfer's desire to hit the ball farther. Length was the lure, but the trade race in this connection was upsetting the design of courses, and championship and other tees had to be sited farther back. The development set the game's legislators a nagging problem, and shortly after World War I the Royal and Ancient Club enacted what is called "the 1-62 formula"—that the ball should have a maximum weight of 1.62 ounces (45.93 grams) and a minimum diameter of 1.62 inches (4.11 centimetres). The USGA tried for two years a ball 1.55 ounces (43.94 grams) in weight and 1.68 inches (4.27 centimetres) in diameter, but in 1932 it reverted to a weight of 1.62 ounces while retaining a diameter of 1.68 inches. The figures on both sides of the Atlantic remain as stated and are the sole exception to a uniform code. Various attempts were made to produce satisfactory solid rubber or artificial-rubber balls over the years, and in the early 1970s solid balls had been improved enough to challenge tension-wound balls in popularity.

Golf clubs. *Early clubs.* Reference has already been made to the royal warrant to make clubs granted to Mayne in 1603, but no product of that time exists, and the oldest clubs known were discovered in a house in Hull along with a newspaper carrying a date in 1741. All of the clubs were made of wood, and antiquarian theory is that they were used in the time of the Stuart kings.

Introduction of iron heads

In the Royal and Ancient Club's museum there are specimens of ancient clubs including two woods and an especially notable putting cleek—*i.e.*, a putter having an iron head on a wooden shaft—made in the second half of the 18th century by Simon Cossar of Leith, club maker to the Company of Gentlemen Golfers. During that period, when Robertson saw that his craft would not be injured by the gutta-percha ball, he realized the value of iron clubs for approaching and made a cleek for steadier putting. Other developments included "Young Tom" Morris' idea for the cup-faced niblick for playing the shorter approaches. He could not have better proved the value of iron heads for clubs than by his victory in the series of matches against Arthur Molesworth, the leading Westward Ho! amateur. In a famous round played in wintry conditions at St. Andrews, a circle was cleared in the snow on the greens, and, whereas Molesworth's pitches invariably ran too far, Morris' pitches with the iron-headed niblick were hit right up to the pin and stopped with backspin.

The club makers of outstanding repute in the early years of the 19th century were Hugh Philp at St. Andrews and the McEwan brothers of Musselburgh, notably Douglas, whose clubs were described as models of symmetry and shape. They were artists at a time when clubs were passing from "rude and clumsy bludgeons" to a new and handsome look. The older, thorn-tree cuts were a thing of the past, and apple and afterward beech, which was a superior driving wood, were used and made the clubs more stylish. Philp, who was the master craftsman, specialized in making wooden putters with a long head of pear-shaped design, and he is reputed to have made no more than 100. A genuine Philp is now a prized collector's piece.

The clubs of a well-equipped player

In his illustrated *Golfing* (1887), the famous encyclopaedia publisher Robert Chambers lists the clubs the well-equipped player used: play club, for driving; long spoon, for a hanging or rough lie; short spoon, for the vital shot within about 100 yards (90 metres) of the green; brassie, brass soled to protect the clubhead; sand iron, for bunkers or other hazards and lofting balls over stymies; cleek, for long shots (note: originally referring to any iron-headed club, cleek had become a Scots name for a no. 1 or driving iron); niblick, or track iron, was small with a heavy iron head for getting the ball out of cart ruts or tough whins.

The hickory shafts of the woods—the play club (modern driver), the spoons, and the brassie—were spliced to heads of apple or beech faced with horn.

Modern clubs. Golf was being overtaken by the Industrial Revolution when the rubber ball came into the game at the beginning of the 20th century. These two factors wrought major changes in the clubs and the production methods, as craftsmanship moved out of the individual professional's shop and into the factory.

NEW MANUFACTURING METHODS

The harder rubber ball brought about the use of persimmon and, later, laminated club heads. Hard insets appeared in the faces. Increased demand led to the adaptation of shoe-last machine tools for the fashioning of wooden club heads. Sockets were bored in the club heads, and shafts were inserted rather than spliced. Drop-forging completely replaced hand forging in the

fashioning of iron clubs, and faces were deepened to accommodate the livelier ball and were machine lined to increase the spin on the ball in flight. Stainless steel replaced carbon steels. Seamless steel shafts took the place of hickory. Composition materials were developed as an alternative to leather in grips, and the grip foundations were molded in so many ways that they were regulated in 1947. Inventive minds created novel clubs, not only centre-shafted and aluminum putters and the sand wedge but also types that were such radical departures from the traditional form and make that they could not be approved by the USGA or the R. and A.

Rules related to club manufacture

In promulgating its revised code of rules in September 1908, the R. and A. appended the notation that it would not sanction any substantial departure from the traditional and accepted form and make of golf clubs. This principle has been invoked many times in an effort to preserve the original form of the game. When Jock Hutchison won the British Open in 1921 with deeply slotted faces on his pitching clubs, the R. and A. immediately banned such faces, and the USGA concurred with a regulation governing markings that became effective in 1924. After Horton Smith, a prominent tournament professional in the 1930s, had effectively used a sand wedge with a concave face designed by E.M. MacClain of Houston, Texas, the principle of concavity was banned in 1931. The U.S. golfer Gene Sarazen, however, developed a straight-faced sand wedge and used it so well in winning the British and USGA opens in 1932 that he completed the revolution of bunker play. Experiments with steel shafts went through several phases. In 1924 the Union Hardware Company of Torrington, Connecticut, drew a seamless shaft of high-carbon steel that could be heat-treated and tempered. Approved by the R. and A. in 1929, it substantially replaced hickory in the early 1930s. Later, shafts of fibre glass and of aluminum were introduced.

Improvement of the shaft was accompanied by the general introduction of numbered, rather than named, clubs and by the merchandising of matched sets rather than individual clubs; clubs had become more numerous and more finely graduated than the names that traditionally had been applied to them, and shafts could be manufactured to specifications for flexibility and point of flex. Whereas formerly a golfer seeking new clubs went through a rack of mashies until he found one that "felt right" and then tried to find other clubs of similar feel, he later bought a whole set manufactured to impart the same feel. The merchandising opportunities inherent in the numbered and matched sets were carried to an extreme, and in 1938 the USGA limited the number of clubs a player might use in a round to 14. The R. and A. concurred in a similar edict the next year.

PLAYERS AND CHAMPIONSHIPS

Whatever uncertainties may remain about the origins of golf, there is no doubt that its development as an organized sport was distinctly British, and Britain produced the first great players of the game. As the early golfing associations, or clubs, became established in Scotland and then England, there emerged a group of professionals who made golf balls, fashioned and repaired clubs, laid out and maintained courses, and gave lessons. Many of them were outstanding golfers and would take on all comers in the popular stakes (money) matches of their day. St. Andrews' Robertson, for example, is regarded as the greatest golfer of his time, and, according to legend, he was never beaten in a stakes match played on even terms (that is, without giving his opponent a handicap). The British professionals and their amateur counterparts represented the best golf in the world from the second half of the 19th century, when the sport began to gain some world prominence, up to about the 1920s, when U.S. players began to excel. With the tremendous increases in financial rewards to be gained in golf after about 1950, especially on the U.S. professional tour, and with the great mobility provided by jet-airplane transportation, golfers from other countries (Australia, New Zealand, South Africa, Japan, Argentina, for example) began to appear in the top tournaments.

The premier championships. The most prestigious championships are the Open and Amateur championships of the British Isles and the U.S. Open, Amateur, Professional Golfers Association (PGA), and Masters.

Great Britain. The Open Championship of the British Isles, which the British like to call The Open to emphasize the tradition and priority of the event begun in 1860, was the concept of the Prestwick Club in Scotland, whose minutes recorded a proposal that all clubs should contribute to raise a fund for a trophy for professional competition. Their hope, however, was badly disappointed, and the offer of support was so meagre that Prestwick decided to go it alone and spent 30 guineas on the ornamental challenge belt to be awarded to the champion. That first Open Championship was won by Willie Park, who repeated in 1863 and 1866. "Old Tom" Morris won in 1861, 1862, 1864, and 1867, and Andrew Strath won in 1865. After Morris' last victory his son "Young Tom" Morris retired the belt, which became his property by virtue of his three successive wins (1868, 1869, and 1870). In the absence of a prize, there was no championship in 1871; but the next year a cup, which has been in competition ever since, was put up, and "Young Tom" won it, for his fourth successive victory. Jamie Anderson, St. Andrews, and Bob Ferguson, Musselburgh, were the idols of their towns in the late 1870s and early 1880s, each winning the Open three times in consecutive years.

The Open Championship

In the late 1880s the championship consisted of four rounds of the nine-hole Musselburgh course in a single day, and to end it before dark bribes were given to competitors who obviously had no chance to induce them to leave the course. Even so, scores had to be checked by candlelight. Willie Park, Jr., who coined the aphorism that "the man who can putt is a match for anyone," was champion in 1887 and 1889.

At the end of the 19th century, England was producing great players. John Henry Taylor and Harry Vardon, together with James Braid, a Scotsman, between them won the Open Championship 16 times between 1894 and 1914. Vardon, the greatest player that the world had seen up to that time, won the title six times. These three supreme golfers were known as "the great triumvirate" and were primarily responsible for the formation of the Professional Golfers Association in 1901. This body is responsible for professional tournaments in Great Britain and for the biennial Ryder Cup match (for professionals) when it is played there.

The early 1900s, however, did not see a complete British monopoly of the Open. In 1907 the title was won by Arnaud Massy of France, the first player from outside Great Britain to achieve this distinction.

The British Amateur and Women's championships

The British Amateur Championship was started in 1885 after the Royal Liverpool Golf Club at Hoylake had proposed a tournament "open to all amateur golfers." The winner was Alan F. MacFie, a Scottish member. The tournament attracted nearly all the best amateurs of the time, but it was not immediately recognized as the championship. The following year the Royal Liverpool suggested to the R. and A. that the tournament be established as the amateur championship, and 24 clubs joined together to purchase a trophy and manage the event. Among British players who won the amateur championship at least two times before the series was interrupted by World War I were H.G. Hutchinson (1886 and 1887), John Ball (1888, 1890, 1892, 1894, 1899, 1907, 1910, and 1912), J.E. Laidlay (1889 and 1891), and H.H. Hilton (1900, 1901, 1911, and 1913). Ball also won the Open in 1890, the first English golfer to do so. Hilton won the Open in 1892 and again in 1897.

Outstanding British amateurs between World Wars I and II were Cyril Tolley, who was Amateur champion in 1920 and 1929 and a regular Walker Cup choice; Roger Wethered, who tied for the 1921 Open and was Amateur champion in 1923; and Scots Hector Thomson, Jack McLean, and A.T. Kyle. Gifted players after World War II included the Irish golfer Joe Carr, who won the title in

1953, 1958, and 1960, and Michael Bonallack, who won in 1961, 1965, 1968, 1969, and 1970.

The tendency for a talented amateur to turn professional is less marked in Great Britain and Ireland than it is in the United States, but a handful of young players did so in the late 1960s, and the tendency increases. As in the United States, the decision is commonly prompted by the amount of tournament prize money on offer and hardly ever by the desire to serve as a club professional. In both countries the professional tends to settle down to a club appointment when his touring days are over. The better he performs on the circuit or in national championships, the more lucrative the eventual appointment.

The Ladies Golf Union was formed rather late, in 1893. The first British Ladies Amateur Championship was held that year on the old St. Anne's course in England and won by Lady Margaret Scott, as were the next two championships. One of the first outstanding woman golfers was Dorothy Campbell, who won the Ladies Amateur Championship in 1909 and 1911 and was runner-up in 1908. She won the U.S. Women's Amateur Championship in 1909, 1910, and 1924 and the Canadian championship in 1910, 1911, and 1912. In all, she won more than 750 prizes in golf.

Of the many women who played championship golf between the wars, Joyce Wethered, sister of Roger Wethered, was outstanding. She won the British Ladies Amateur title in 1922, 1924, 1925, and 1929. Later Lady Heathcoat-Amory, she still played occasionally, delighting spectators with a style and ability that Bobby Jones considered supreme among women golfers. Her great rival was four times British Ladies Amateur champion Cecil Leitch, who had something like Braid's "Divine Fury" in her swing. Enid Wilson and Frances Stephens were among the many fine players who succeeded them.

The United States championships

United States. The first official U.S. Open, Amateur, and Women's Amateur championships were held in 1895; the respective winners were Horace Rawlins, Charles B. Macdonald, and Mrs. Brown. An unofficial open, at match play, in 1894 was won by Willie Dunn.

Walter J. Travis was the first great U.S. golfer. He was born in Australia, but his golf was wholly learned in the United States. Of striking appearance, with jet-black beard and impeccable garb, he was unpopular with fellow golfers because of his austere, taciturn demeanour. But he proved his ability as a golfer by winning the U.S. Amateur in 1900, 1901, and 1903, by reaching the semifinals in five other years, and by winning the qualifying medal in 1900, 1901, 1902, 1906, 1907, and 1908. He won the British Amateur title the only year he entered this event—1904.

Jerome D. Travers learned his golf at Nassau Country Club, Long Island, under the tutelage of Alex Smith, a famous Scots professional who went to the U.S. in 1898. Travers was a player with indomitable courage, an ability to outgame an opponent at match play, and nerve that rarely failed him in a crisis. He won the U.S. Amateur Championship in 1907, 1908, 1912, and 1913 and was finalist in 1914; he won the U.S. Open title in 1915; and he won a long list of sectional championships.

Among the women golfers, after Mrs. Brown won the first amateur championship in 1895, Beatrix Hoyt won the next three in a row. Other early multiple winners were Genevieve Hecker (1901 and 1902) and the outstanding Margaret Curtis (1907, 1911, and 1912).

Ascendancy of U.S. players. After World War I the influence of the many Scottish golfers who had emigrated to the United States became evident. U.S. golfers (principally Walter Hagen and Robert T. [Bobby] Jones, Jr., who achieved the unparalleled performance of winning the Open and Amateur championships of Great Britain and the U.S. in the same year, 1930) monopolized the British Open Championship until Henry Cotton won in 1934, a feat he repeated in 1937 and 1948. Cotton had determined to break the U.S. monopoly and did so after long practice and severe self-denial.

By the early 1930s U.S. dominance of the international scene was growing. Since 1933 the only victories by British teams in the biennial matches against the United States have been in 1938 and 1971, when the amateurs won the Walker Cup (the match was tied in 1965); in 1952 and 1956, when the women won the Curtis Cup, which they retained in 1958 with a tie; and in 1957, when the professionals won the Ryder Cup.

Early amateur golf champions

Francis D. Ouimet became a national hero in 1913 when, unknown as a golfer except around Boston, he tied Vardon and Ted Ray, two of the best British professionals, at 304 strokes for 72 holes in the U.S. Open, held at Brookline, and defeated them in a play-off. The following year Ouimet won the U.S. Amateur, and he repeated 17 years later, in 1931. He was runner-up in 1920 and a semifinalist in 1923, 1924, 1926, and 1927. He was a member of the United States team against Great Britain for the Walker Cup from the first of these international matches in 1922 to 1949, serving as captain in 1932, 1934, 1936, 1938, 1947, and 1949.

Charles ("Chick") Evans, Jr., first showed promise as a golfer around Chicago in the period 1906–10. He was runner-up in the U.S. Amateur of 1912 and the U.S. Open of 1914, winning the Western Amateur title in both those years and also in 1909 and in 1915, 1920, 1921, 1922, and 1923. In 1916 he became the first golfer to win the U.S. Amateur and Open in the same year; his score of 286 in the Open was unbeaten for 20 years, until Tony Manero scored 282 in the 1936 Open. In 1944, at the age of 54, he won the Chicago city championship, for the fourth time. Evans competed in 50 USGA Amateur championships and won 57 individual matches through 1962.

Bobby Jones was regarded as the greatest amateur golfer of modern times. His career was brilliant from his debut in national competition in the U.S. Amateur of 1916 until his unparalleled performance in 1930 of winning all four of the world's most difficult titles—the British Amateur, the British Open, the U.S. Amateur, and the U.S. Open. This feat became known as Jones's "grand slam." During his golfing career Jones won the British Open three times, the British Amateur once, the U.S. Open four times, and the U.S. Amateur five times. He played for the U.S. against Britain in the Walker Cup team matches in 1922, 1924, 1926, 1928, and 1930.

W. Lawson Little, Jr., in 1934 won both the United States and British amateur titles and the following year repeated his victories in both events. Little turned professional in 1936 and won the Canadian Open that year and the U.S. Open in 1940.

Jack Nicklaus, a very long hitter, won the U.S. Amateur twice, in 1959 and 1961, by the time he had reached the age of 21; but he turned professional later that same year and compiled an even more impressive record by winning the PGA Championship and the British Open twice each, the U.S. Open three times, and the Masters four times.

Arnold Palmer, also a golfer of exceptional strength, first won the U.S. Amateur in 1954 and after turning professional compiled another great modern record in winning the U.S. Open once, the British Open twice, and the Masters four times.

The popular appeal of the U.S. Amateur Championship has been seriously weakened by these departures to the professional ranks, and it has become exceptional for an amateur champion to resist the lure of tournament money. It is equally unlikely that he will have made his name without having had the experience of playing collegiate and intercollegiate golf, in which standards of instruction and play are generally high.

Professional golfers have contributed richly to the history of the game in the United States, and, while it was well into the 20th century before there was a native-born champion among them, the overall record of the professionals was remarkable—Walter Hagen and Gene Sarazen were particularly outstanding.

Walter Hagen and Gene Sarazen

Hagen first appeared on the national scene in the 1913 U.S. Open at Brookline, where he gave an excellent account of himself, considering his competitive inexperience, and tied for fourth against an expert field. The following year at Chicago he won the event. His golf was unorthodox; he had no desire to copy the smoother

swings of his fellow professionals, and he was entirely self-taught. He scorned to practice by the hour. He dressed immaculately and lived luxuriously, and to him, more than to any other golfer, goes the credit for breaking down the social barriers between amateurs and professionals. Between 1914 and 1936 Hagen won the U.S. Open twice, the British Open four times (a feat matched after World War I only by A.D. ("Bobby") Locke of South Africa and Peter Thomson of Australia), the PGA Championship five times (including four in a row—1924, 1925, 1926, 1927), the Canadian, French, and Belgian opens once each, and at least 45 other events. In all, he participated in not less than 200 open tournaments and was rarely out of the money. In addition, he played in about 1,500 exhibition matches in the U.S. and other countries, many of them for high fees or stakes. He is said to have earned around $1,000,000 during the 22 years he was rated as a topflight golfer.

Sarazen, a man of small stature, reached golfing fame in 1922 by winning the U.S. Open at Chicago and proved he was a golfer of more than passing ability by adding the professional title that same year at Oakmont, Pennsylvania, and the following year at Pelham, New York. No further titles of importance came his way for ten years, but during this period he was a constant competitive threat. In 1932 he reappeared as a champion, winning the British Open with a brilliant 283 and the U.S. Open with an equally brilliant 286, which tied the then low-scoring record established in 1916 by "Chick" Evans. In 1933 he won the PGA title for the third time and in 1934 came within one stroke of winning the British Open. In 1940 he tied Lawson Little for the U.S. Open Championship with a score of 287, but lost the play-off with 73 to Little's 70. He was a persistent opponent to slow play, which has blemished numerous championships, particularly after World War II. (The practice, unhappily, spread to club golf as well.)

In the late 1930s the so-called pro circuit, underwritten by civic and club organizations throughout the country, began putting up major prize money for the experts. Robert E. Harlow developed this circuit and was the first tournament manager of the PGA. Fred Corcoran succeeded him in 1936. That year, aggregate prize monies totalled $100,000; when Corcoran left the field, in 1947, they totalled $650,000. In the 1970s the PGA circuit offered in excess of $7,000,000 in prizes annually.

Other leading professionals

Other names in U.S. professional golf history include the English-born James Barnes, who won the professional title in 1916 and 1919, the U.S. Open in 1921, and the British Open in 1925; Leo Diegel, who won the PGA title in 1928 and 1929; Denny Shute, who won in 1936 and 1937; and Ralph Guldahl, who won the U.S. Open title in 1937 and 1938. After 1940 Byron Nelson, Ben Hogan, and Sam Snead, a gifted natural athlete, shared the major portion of prize monies. Following World War II Nelson, an extraordinarily accurate iron player, retired from serious tournament participation. After winning the U.S. Open and the PGA title in 1948, Hogan, following his recovery from serious injuries suffered in an automobile accident, returned to the golfing circuit in 1950 and won the U.S. Open in that year and in 1951 and 1953, the year he also won the British Open—to duplicate the feats of only two other Americans at that time, Bobby Jones and Gene Sarazen, each of whom won both opens in one year, as did Lee Trevino in 1971. Hogan also scored a record 14-under-par 274 to win the Masters.

The Masters tournament, an open, invitational event played each spring at Augusta, Georgia, was originated in 1934. Hogan's record fell to 25-year-old Jack Nicklaus (1963 winner) in 1965 when he scored a 17-under-par 271, including one round of 64, tieing the record set by Lloyd Mangrum in 1940. Nicklaus won by 9 strokes over Palmer and Gary Player of South Africa (winner in 1961); Jones called it "the greatest performance in golfing history."

The years succeeding World War II saw many other outstanding players approach and often attain these high standards. Cary Middlecoff, a very deliberate player who had turned professional shortly before, won the U.S. Open in 1949 and again in 1956. Julius Boros, who captured the title in 1952, regained it 11 years later. Bill Casper, an immensely consistent golfer, was U.S. Open champion in 1959 and 1966. In 1964 Tony Lema, who had come up through the caddie ranks, at his first attempt on the British Open took the title after giving himself almost no time to practice. Lema was the first eminent golfer of Portuguese ancestry and a formidable contender on the professional tour.

During this era the professional talent in particular became more and more a product of the Southern and Western states and less characteristic of the Eastern Seaboard, as it had been in Hagen's and Sarazen's day. This was especially true of San Diego, where the encouragement offered to juniors later brought onto the national scene not only Casper but also Gene Littler, Phil Rodgers, and—in ladies' golf—Mickey Wright. Littler was U.S. Open champion in 1961 and won numerous tournaments with a swing of great elegance. Rodgers, a capable tournament player, tied for third in the British Open of 1962 and a year later tied for first (he was defeated by Bob Charles of New Zealand in the play-off). Miss Wright's record will be referred to later.

The year 1971 saw an extraordinary accomplishment by Lee Trevino. Partly of Mexican birth, he had developed his game in the U.S. Army, principally in the Far East, and came to prominence as U.S. Open champion in 1968. Three years afterward he recaptured the Open title, won the Canadian Open, and then triumphed in the British Open, all within the record period of a month.

As the pace quickened and increasing numbers of promising players set their minds on a tournament career, the amateurs found themselves outclassed in national opens. There was now no Bobby Jones or Chick Evans to draw more spectators than were following the professionals. Once they had turned professional, the leading amateurs, for the most part, were quite successful, but no national title holder emerged from the amateur ranks after Doug Sanders won the Canadian Open in 1956. Subsequently, he too turned professional.

On the British circuit no player had a longer or more remunerative career than Dai Rees, though he continually failed to win the British Open, just as Sam Snead never won the U.S. title. In many ways their careers—at the beginning of the 1970s almost 40 years in duration—ran in parallel. Both began to play golf in economically depressed areas, Snead in Virginia, Rees in South Wales; both were Ryder Cup selections over a long period (and were honoured with captaincies); and both won numerous events abroad as well as in their own countries.

It was not until 1970 that a British professional made the first serious dent into American supremacy. Tony Jacklin, after playing successfully on the American circuit, became U.S. Open champion by a seven-shot margin. This was the first time a Briton had won America's most important championship since Ted Ray had done so half a century before.

In women's golf, Alexa Stirling of Atlanta, Georgia, won her first U.S. Women's Amateur while quite young in 1916 and repeated in 1919 and 1920. She was runner-up in 1921, 1923, and in 1925, as Alexa Fraser. She also won the Canadian championship in 1920 and 1934.

Outstanding women golfers

Another great U.S. woman golfer, Glenna Collett Vare, won her first Women's Amateur title in 1922 and repeated on five occasions—1925, 1928, 1929, 1930, and 1935. She made four attempts at the British championship but was turned back—on two occasions in the final round. Mrs. Vare's extended hold on women's golf was broken when Virginia van Wie of Chicago replaced her as champion during the three seasons of 1932, 1933, and 1934. Miss van Wie forfeited her title without contest in 1935, when Mrs. Vare regained the title by defeating Patty Berg. Betty Jameson won the title twice in a row, in 1939 and 1940. And in 1960 Mrs. Vare won the U.S. women's Senior championship.

The greatest names in women's golf after World War II included Mrs. Mildred ("Babe") Didrikson Zaharias, Patty Berg, Louise Suggs, Betsy Rawls, and Mickey Wright, all of whom played professionally. Mrs. Zahari-

as, an Olympic winner in track and field in 1932, picked up a golf club that year at the invitation of the sports writer Grantland Rice and from that time played the game with astounding success. She regained her amateur status long enough to win the British Ladies Amateur title in 1947—the first American to do so. Thereafter she turned professional again and embarked on a series of successful golf tours. The women follow a professional circuit similar to that of their male counterparts. From 1946, when the women started their own open championship, Mrs. Zaharias continued to be the leading woman player until her death in 1956.

Outstanding amateurs of the 1950s and 1960s included JoAnne Gunderson Carner, five times winner of the U.S. Women's Amateur championship, and Anne Quast Decker Welts, the winner three times. Among the professionals, Mickey Wright and Betsy Rawls each won the U.S. Women's Open four times during those years; Donna Caponi won the Open both in 1969 and 1970.

The international scene. *International matches and tournaments.* The first organized series of regular international matches were between Great Britain and the United States. The amateur team match between the two countries for the Walker Cup was inaugurated in 1922 and the professional team match (Ryder Cup) in 1927. The women's amateur team match (Curtis Cup) began in 1932. Although the competition in all of these contests has often been close, the U.S. players have managed to win the cups, with rare exceptions.

More truly international matches appeared on the scene in the 1950s. The World Cup (formerly the Canada Cup), played for by two-man teams of professionals, was first offered in 1953, when it was won by Argentina. Only seven nations took part in that first contest; there were 25 in the second; and when Canada, the original host nation won for the first time in 1968, there were 40 teams of two players. More than 50 countries are represented on the World Amateur Golf Council, which was organized by officials of the USGA and the R. and A. in 1958. Its first tournament for the Eisenhower Trophy, held that year, was won by the Australian team. In women's golf, the World Women's Amateur Team Championship for the Espirito Santo Trophy was instituted in 1964. The first winner was France.

International circuits. The coming of speed and sophistication in jet transport stimulated competition. Ocean hopping has become routine, enabling outstanding players from South America, Australia, New Zealand, and South Africa to compete in the premier championships in Great Britain and the United States and on the rich U.S. PGA tour. The first successful invader from the Southern Hemisphere was Bobby Locke of South Africa, a magnificent putter, who won the British Open in 1949, 1950, 1952, and 1957. Peter Thomson of Australia won it five times, the first in 1954 and the last in 1965. In the 1960s Kel Nagle (Australia), left-hander Bob Charles (New Zealand), and Roberto de Vicenzo (Argentina) each won the Open once. But South African Gary Player made perhaps the biggest mark on the international scene, winning the British Open in 1959 and 1968, the U.S. Open in 1965, the U.S. PGA in 1962, and the Masters in 1961. He also did well on the PGA tour through the 1960s and into the 1970s and was the leading official money winner in 1961.

Players from the Southern Hemisphere

With more overseas players entering, and sometimes winning, tournaments in the United States and Great Britain, more U.S. and British players, especially British professionals during the winter months, began to take part in various circuit itineraries in other parts of the world—in Australia, New Zealand, South Africa, and the Far East. The Far Eastern circuit takes in promotions in the Philippines, Singapore, Malaysia, Thailand, Hong Kong, Taiwan, Tokyo, and India. The Asian players have been especially noted for their short-game precision. In Japan the game has become an endemic fever, and it was a sign of the times when Japan won the World Cup in 1957. (For records of the premier championships and international events see also SPORTING RECORD in the *Ready Reference and Index*.)

Finance. Professionals with expertise and crowd-pulling attributes have become a new affluent society thanks to the co-sponsorship, by the PGA and local promoters, of lucrative weekly tournaments with television rights to negotiate. The modern professional also has many sources of peripheral income. The great Harry Vardon received £50 with his medal for winning his record sixth British Open in 1914. By 1971 the winner's prize was £5,500; and in 1972 for the first time all professionals in the championship proper received prize money, including a minimum of £125 for those completing the 72 holes. The emoluments in America far overshadow the escalation of those in Britain. Men who get to the top in America reach the $200,000 bracket, fly their own private planes, and have their lucrative, outside business interests expertly managed for them.

The emoluments of professional golf

The traditional Open Championship in Great Britain has an attached bonus of munificent value. It is one of the four events the winners of which compete in the 36-hole World Series of Golf in the United States, with its first prize of $50,000. The other events that count for this award are the U.S. Open, the Masters, and the U.S. PGA Championship. The U.S. professional match-play championship—revived in 1971—had a total prize payout of $200,000. The largess regularly brings in recruits. There were 250 applicants in 1970 for the 18 qualifying schools held by the PGA to determine approved tournament players. More than $7,000,000 was disbursed in tournament prize money in 1970, but rigid qualifying systems, even after the aspirant has become an approved tournament player, make the circuit life an arduous and frequently disappointing one. To give the less successful professionals some encouragement, events known as satellite tournaments, played concurrently with the main promotion but offering less prize money, now appear on the tournament calendar.

The U.S. PGA is fortunate in their lavish sponsorships. In Great Britain finance is not so easy, and the Royal and Ancient Club, being without any commercial sponsorship, has to rely for meeting its expenses for the Open Championship and international and other commitments, such as the Walker Cup and Eisenhower Trophy tournaments, largely on the Open Championship gate money. The costs are high, with the new standards of spectator accommodation and other amenities, and more than one championship has left a deficit. Promotional costs climbed by five times in about a dozen years up to 1970. Happily, the club has come out on the profit side at other times and has reserves to call upon. The championship proceeds are kept entirely separate from the club's domestic funds and have never been used to subsidize these.

Senior golf. One of the most significant developments in golf in the 20th century occurred in senior golf after 1905, when Horace L. Hotchkiss arranged the first seniors' tournament, for players aged 55 and older, at the Apawamis Club, Rye, New York. Hotchkiss, who was more than 60 at the time, attempted to prove that golf was not exclusively a young man's game. The tournament was such a success that within ten years the number of contestants had passed the 300 mark. The United States Seniors Golf Association was organized on January 17, 1917, in New York with a membership of 400, which within six months increased to 500 and subsequently to 1,000. While the United States seniors' tournament is the leading event of its kind, many membership and invitation events of the same type developed—the American Seniors' Golf Association, the Western Seniors' Golf Association, the North and South Senior Tournament, and others. There are at least 50 senior golfing organizations in the United States alone. Members of the United States Senior Women's Golf Association play annually. In 1918 the Governor General of Canada presented a trophy to be played for annually by the United States Seniors Golf Association and the newly formed Canadian Seniors' Golf Association. Another match was initiated with the Senior Golfers' Society of Great Britain. The annual World Senior Amateur Championship was first held in 1960; the biennial World Amateur Team Championship was established in 1967.

II. The modern game

COURSES AND FORMS OF PLAY

The game consists in playing the ball from a teeing ground into a hole by successive strokes in accordance with the rules. The stipulated round consists of 18 holes, and most golf courses have 18. Standard 18-hole courses measure from 6,500 to 7,000 yards (5,900 to 6,400 metres); individual holes, from 100 to 600 yards (90 to 550 metres). Some courses, however, have only nine holes, and these are played twice in a stipulated round. The clubs are designed for the various positions in which the ball may come to rest and for the various distances to the hole. The objective is to hole the ball in the fewest strokes.

The number of holes on a course

When the game entered the 19th century, there was no agreement on the number of holes on a course; localities differed widely in the matter. When the popularity of Leith, with its five holes, waned and St. Andrews became the hub, the round of 18 holes became the established order. Originally the St. Andrews holes filed straight out alongside the shore and were simply played in reverse for the return journey—11 holes each way. In 1764 the round was modified to 18 holes. The infinite variety of courses gives golf an intrinsic charm. Gleneagles in Scotland and the Augusta National in Georgia are cited as representing only two of the many outstanding examples of environmental beauty of courses around the world.

Course architecture has been greatly modernized. Planning at the beginning of the 20th century was on simple, even primitive, lines. The bunker, or sand trap, that by constructional canons stretched an ugly line across the fairway to trap topped balls was fashionable. The strategic, as against the frankly penal, type of hazard was developed, and bunkers and other hazards that can be artificially created have been a key role in the design of interesting and challenging holes on the best courses. The modern green, even on flat ground, is usually banked, and teeing grounds have been lengthened and—in a few instances—given a crescent shape, in order to vary the distance or the angle to the hole.

Match and medal play. There are two distinct forms of play: match play and stroke (medal) play. In match play the player and his opponent are playing together and competing only against each other, while in stroke play each competitor is competing against every other player in the tournament. In match play the game is played by holes, and each hole is won by the player who holes his ball in the fewer strokes. If both players score the same number of strokes, the hole is halved. When a player has won one more hole than his opponents, he is said to be 1 up. The match is won by the player who is leading by a number of holes greater than the number of holes remaining to be played, as, for example, 3 up and 2 to play. In stroke play the competitor who holes the stipulated round or rounds in the fewest total strokes is the winner. Amateur championships once were all at match play and open championships and most professional events at stroke play, over 72 holes. Some amateur events have been changed to stroke play (the U.S. Amateur event was at stroke play from 1965 to 1973), as has the U.S. PGA Championship. An additional PGA championship, in a modified form of match play, was initiated in 1971; players won or lost on their 18-hole scores, instead of hole by hole.

Scoring match and medal play

Stroke play requires a greater degree of consistency in a player, for one hole where he lapses into a high figure can ruin his total and cost him victory. The same high score on a hole in match play means only the loss of that hole. In both match and stroke play, players can compete as individuals or as partners. When two players compete as partners, each playing his own ball, the better ball on each hole is their score for that hole; this is a four-ball or best ball match. Two players may compete as partners with two others, each pair playing alternate strokes on a single ball; this is a match foursome.

Handicaps. Players of varying abilities compete against each other by using handicaps. A handicap is the number of strokes a player receives to adjust his score to a common level. The better the player, the smaller his handicap, and the best players have handicaps of zero (scratch players). A scratch player whose average score is 70 can have an even match with a player whose average score is 80 by giving him a handicap of 10 strokes. Handicap golf is limited to amateur competitions, and championship tournaments are played without handicaps.

Par golf. An expert golfer plays most of the holes on the course in 4 strokes, a drive of 225 to 250 yards (200 to 225 metres), a shot to the green, and 2 putts. Every course, however, contains a few short holes on which the expert might be expected to drive onto the putting green and a few long holes on which the expert might require a drive and 2 more strokes to reach the putting green. On the former, he would be expected to make 3 and on the latter 5, since 2 putts on each green is the standard.

Every course has a par, which is defined as the score an expert (*i.e.*, a scratch player) would be expected to make, and many courses also have a bogey, which is defined as the score that a moderately good golfer would be expected to make. Both par and bogey are defined as errorless play without flukes and under ordinary weather conditions, allowing 2 strokes on the putting green. Par is essentially a U.S. term that came into use in the early 1900s as a base for computing handicaps under the system devised by L. Calkins of Plainfield, New Jersey. Bogey is essentially a British term that came into use in England in 1891 and was derived from a mythical Colonel Bogey, who was described as uniformly steady but never overbrilliant. Colloquially in the United States, "bogey" is used to indicate a score 1 stroke above par.

The British Standard Scratch Score

The basic Standard Scratch Score (sss) used in Great Britain for a normal course of 6,300 yards (5,800 metres) is 70, but a stroke is added or deducted for each 200 yards (180 metres) by which the total length of the course varies from that figure. An addition may be made to the sss to allow for courses with exceptional difficulties other than length. Handicaps are based on the sss.

EQUIPMENT

Golf balls. Regulation golf balls have a maximum weight of 1.62 ounces (45.93 grams). The minimum size of the United States ball is 1.68 inches (4.27 centimetres) in diameter; that of the British ball 1.62 inches (4.11 centimetres) in diameter. The velocity of the U.S. ball may not be greater than 250 feet (75 metres) per second when measured under prescribed conditions on an apparatus maintained by the USGA, but there is no velocity specification for the British ball.

Golf clubs. In the average good player's set there are either three or four wood clubs and nine or ten irons (no more than 14 clubs may be carried during a round). No two clubs in a set are the same. There are differences in length and suppleness of shaft, weight, size and shape of head, the angle at which the shaft ends and the head begins (the lie), and the angle of the face of the club from the vertical (the loft).

The various clubs are known both by number and by name. The names have come down from the early days of golf; the numbers are a U.S. innovation dating from the early 1920s, when matched sets came into use (see above *New manufacturing methods*). The most widely used clubs are identified below.

Wood clubs

Number 1 (driver), used from the tee for maximum distance; has a large head and a deep, almost vertical, face. The face has an angle of loft of about 10°.

Number 2 (brassie), so called because the sole of the club originally was covered with a brass plate. Used mostly for long shots from good fairway lies, the club has a slightly smaller and shallower face than a driver but with more loft.

Number 3 (spoon), shorter shaft and shallower face than driver or brassie, but face has considerably more loft. The club is used to play the ball from lies too poor for a brassie and also for strokes when the use of a driver or brassie would send the ball beyond the green.

Number 4 (baffy), smaller head, shallower face, and more loft than a spoon. It will hit a ball about as far as a number 1 iron.

Number 5, a great favourite with players who have an

aversion to iron clubs, it replaces the number 3 or 4 iron. It has a small and compact head.

Irons

Number 1 (driving iron or cleek), a long shaft and an angle of loft of about 20°. Used for tee shots and full shots from lies too "heavy" for a wood. A difficult club to use, it produces a long, low shot.

Number 2 (midiron), slightly more loft, for shots of shorter distance than number 1 iron.

Number 3 (mid-mashie), more loft; for shorter distances than number 2.

Number 4 (mashie iron), more loft; for shorter distances than number 3.

Number 5 (mashie), more loft; for shorter distances than number 4. The ball pitches high and stops quickly after hitting the ground. This club is also used for pitch-and-run shots to the green; the ball travels part of the way through the air, then rolls the rest of the way.

Number 6 (spade mashie), more loft; for shorter distances than number 5 and also for playing the ball from high grass or difficult lies, when getting out is more important than distance.

Number 7 (mashie niblick), resembles the spade mashie but has still more loft and head weight; puts considerable backspin on the ball.

Number 8 (pitching niblick), still more loft; for shorter distances than number 7.

Number 9 (niblick), face has a loft of about 47°, and the head is heavy, to carry it through long, tough grass or heavy sand. A ball, properly hit, rises almost vertically and upon hitting the ground may jump backward as a result of the backspin this club imparts.

Number 10 (wedge), face has a loft of more than 50°, but the club has a broad flange on the sole. There are two types—the sand wedge to use in bunkers and the pitching wedge for pitch shots.

The putter, a club with a short, stiff shaft and a straight, or nearly straight, face, for rolling the ball on the green. There are many styles of putters.

THE RULES OF GOLF

The rule-making bodies for golf are the R. and A. and the USGA. They attempt to perpetuate the uniformity in rules by exchanging views on interpretations and on recommendations for revision. The rules govern play all over the world. While the basic principle of the rules is simple, the code itself has become complex over the years.

The first rules

The present code makes an amazing contrast with the first documentary rules in history, 13 in number, that were framed by the Honourable Company. The first of them ordained that the ball had to be teed within a club length of the previous hole and the tee had to be on the ground. Tee and green were one. The ball struck from the tee was not to be changed, and the player could (Rule 5) take his ball out of water or "watery filth" to play it and allow his opponent a stroke. The St. Andrews golfers, in founding the Royal and Ancient Club, adopted an almost carbon copy of the Leith rules. There were periodical adjustments and abortive efforts at reform before the rules committee of the R. and A. was formed in 1897 to become the final authority.

The rules committee has co-opted representatives from the Commonwealth, the European Golf Federation, the United States, and the British Unions Advisory Committee. Britain and America have had separate codes at various times, but a uniform code went into effect in 1967. The sole exception to the arrangement is in the specifications for the ball (see above).

Local rules. Rule 27 of the current code permits clubs to make local rules to cope with abnormal conditions on their courses, but any penalty imposed by the code must not be waived by a local rule. There have been many curious instances of local rules—the Duke of Windsor, for example, used a concession at Jinja on an African tour to lift his ball from a hippopotamus' footprint.

The World Cup competition has its own regulations, which, in effect, come into the category of local rules; and, because the U.S. PGA tournaments are played under varying conditions and in different climates, their rules, too, are not identical with those of the USGA and the R. and A.

Definition of amateur. The rules of golf define an amateur golfer as "one who plays the game solely as a non-remunerative and non-profit-making sport." But the elasticity of this definition perturbs the game's legislators for what it does not define. The whole question of status in its various aspects engages the attention jointly of the R. and A. and the USGA.

Under a recent change, an amateur remains so until and unless he takes specific action toward becoming a professional, even though he might have indicated his intention of becoming a professional in the future. The expense of becoming a top-class amateur is formidable if not impossible unless a player is financially well situated, and the difficulty may drive him into one of two channels —a profitable post obtained through his golf reputation or some form of sponsorship, which is prohibited.

The official veto against an amateur accepting golf merchandise without paying the market price for it has been enforced in the case of nationally known players who received unsolicited golf balls from a firm that aimed at publicity through the gift. The players were temporarily suspended. The USGA committee on status warned commercial interests against offering prizes to amateur golfers as a disservice to both the game itself and the players. It is illegal to accept a prize or prizes in any one competition of a retail value of £50, or $200 in the United States. Prizes of symbolic value such as metal trophies are permissible.

Applications by amateurs for reinstatement are decided on their merits, but a player who has been in the professional ranks for five years is ineligible for it. In the United States over 100 players had their amateur status restored in 1970.

PLAY OF THE GAME

Driving, playing the fairway, and hazards

The starting place for each hole to be played is the teeing ground. The front is indicated by two markers, and the teeing ground is the rectangular space two club lengths in depth directly behind the line indicated by the markers. The player tees his ball anywhere within this space, usually setting it up on a small wooden or plastic peg, and strikes it toward the hole. The stroke from the teeing ground is called the drive. For this, the player usually employs a number 1 wood club, or driver, although to avoid a hazard or to attempt to place his ball in a favourable position for his second shot (for example, on a long hole with a sharp bend, or dogleg) he may prefer one of the other woods, or an iron. On short, par-three holes most players use an iron.

The preferred line to the hole is generally a clear, mowed route called the fairway. The fairway was historically bordered by unmowed vegetation—heather, grasses, weeds, bushes—called rough. Today, however, most courses in the United States are not characterized by deep and tangled rough and when inland make effective use of trees. At strategic places along the preferred line to the hole and guarding the putting green are obstacles called bunkers, depressions in the ground filled with sand (sand traps). Some holes require the player to skirt or cross streams or ponds. Both bunkers and bodies of water are termed "hazards." If a player's drive lands on the fairway, on a long hole he will normally use a number 2 or higher wood or one of the lower numbered irons to attempt to reach the green. If his ball has landed in the rough, he usually will have to use an iron to hit it out—and consider himself fortunate if he happens to get good direction and distance as well. A ball in the water, unless it is very shallow, usually is unplayable, and the player must take a penalty; if the ball is in the loose sand of a bunker, he will try to blast it out, usually with a number 10 iron, or wedge, although if the bunker is at a distance from the green and the ball lies well perched on the sand he may be able to use a longer iron.

Approaching and putting

When the player has come within close range of the green, two methods of play are open to him. He may pitch the ball all the way and depend on backspin to stop his ball near the pin, or he may play a chip shot, in which the ball flies part way through the air, as to the edge of the close-clipped surface of the green, and then rolls the remaining distance.

The hole itself measures 4¼ inches (10.8 centimetres)

in diameter, is at least 4 inches (10 centimetres) deep, and is set in an area of turf especially prepared and maintained and closely mowed for putting. When the player putts he uses a straight-faced club and rolls the ball across the putting green toward and eventually into the hole. The player must hit the ball along a line that allows for very little margin of error and with enough force to roll the ball to the hole but not too far beyond in case he misses it. And, since most greens are not level but have numerous pitches and minor slopes, great care must be taken to select the proper line, which may be quite far to one side or the other of the hole. There are many styles of putters and many styles of putting: the style in which the ball is straddled in the "croquet" stance was banned under R. and A. and USGA rules in 1968.

VARIANTS

Par-three golf. Par-three golf courses, on which each hole measures 100 yards (90 metres) more or less and plays at par three, were developed as a result of the shortage of available open land in congested urban areas. A classic example is Edinburgh's Bruntsfield Links (see above *History*), now a short course. Whereas a regulation 18-hole course may stretch to over 7,000 yards, about four miles (6½ kilometres), an 18-hole par-three, or short-hole, course can be laid out in about 1,800 yards (1.6 kilometres), and a nine-hole short course in less than 1,000 yards (0.9 kilometres). Short courses have appealed to golfers in areas where public courses are overcrowded and private clubs have waiting lists or high membership fees, and they have proved popular among inexperienced players. Their limited area is an important consideration when urban land has greatly increased in value, and they would seem to have a prosperous future. Some par-three courses are flood-lighted for night play.

Driving ranges. Driving ranges were developed as commercial establishments at which golfers and aspiring golfers could, for a small fee, practice their swings and hitting the ball for distance or accuracy or both. They, too, have appealed to golfers in areas in which courses are overcrowded or hard to reach. Driving ranges have proved to be especially popular in Japan, where crowded conditions prevail. By the early 1970s there were over 2,000 of them in that country, with some 300 in the immediate vicinity of Tokyo.

Putting games. Practice putting greens have long been a feature of most courses for the use of even the most expert players in refining the difficult and delicate skills of rolling the ball over slanting surfaces into the hole. With the great growth of interest in golf, more putting greens have been established for recreational purposes.

In the 1930s a putting game known as miniature golf produced a brief craze in the United States. In this game a putter was used to hit a golf ball across a smooth surface (usually crushed cottonseed hulls) and through, over, or around various baffles, or hazards—abrupt dips or rises, sharp turns, curved sections of tin pipe, and so on—and into a series of holes, nine or 18, laid out as a replica of a golf course. Briefly extremely popular, the game survived as a minor attraction in some amusement parks. Other putting games have been invented, and numerous practice putting devices have been marketed, primarily for indoor practice by the individual golfer.

NEW TRENDS

In the last third of the 20th century the game appeared to be moving still further from its Scottish infancy. If the golfer's object was unchanged and his method of playing the conventional one, a number of minor revolutions were taking place in other departments of the game.

Course maintenance

Because the earliest golf courses were links, the natural qualities of seaside turf made greenkeeping much less of a science than it is now. Courses that have been laid out over less helpful terrain—such as desert country—may look artificial in comparison, but they demand a sophisticated knowledge of maintenance, likely to be acquired through university courses in agronomy and augmented by the results of turf-institute research. The modern course superintendent needs to be an expert in the complexities of tractors and machinery, fertilizers, insecticides, weedkillers, and irrigation systems.

Caddies, golf carts, and buggies

As for the traditional caddie (the word derives from the French *cadet*), he is a dying breed in the English-speaking countries, mainly because work of this kind no longer appeals to boys seeking pocket money or to restless, unskilled men. The few remaining caddies who really understand their duties are likely to be nomadic and highly individual, if not downright eccentric, in character. Their employers are normally the amateur champions and the touring professionals, and in the PGA events a caddie can receive a handsome bonus when his man wins.

Partly because of the shortage of caddies, the two-wheel bag carrier—called a golf cart or, in Britain, a trolley—has been widely used since World War II. More recently, the electrically driven golf car or buggy (a two-seater) has proved a blessing, if an expensive one, to players with limited time or who may be physically unable to walk long distances. It is also a useful vehicle to tournament supervisors in the field.

Men's apparel has changed considerably. No one plays in a jacket any longer, and plus fours, or knickers, are all but obsolete. Women still play in skirts, but slacks or shorts are widely accepted.

From the 1960s onward there was a marked increase in the number of courses designed for vacation or retirement living. In some parts of the world—notably the Caribbean islands and the Mediterranean countries—this gave golf a new internationalism. In such states as Florida, Arizona, and California, the golf course was often the best inducement, except for the climate itself, that a land developer could offer his clients.

BIBLIOGRAPHY. BERNARD DARWIN *et al.*, *History of Golf in Britain* (1952), a survey of the early history and development including women's golf; ROBERT TYRE (BOBBY) JONES, *Golf Is My Game* (1960), the great American golfer's own story, with an instructional section; HENRY COTTON, *My Golfing Album* (1959), a chatty roundup of reminiscent bits and pieces about tournaments, travels, and personalities; JAMES K. ROBERTSON, *St. Andrews Home of Golf*, 2nd ed. (1968), an outline of the unique role St. Andrews has had in shaping the game; CHARLES PRICE, *The World of Golf: A Panorama of Six Centuries of the Game's History* (1962), with chapters on pioneers, shotmakers, and masters; WILL GRIMSLEY, *Golf: Its History, People and Events* (1966), with a special section on golf architecture by ROBERT TRENT JONES; ROBERT BROWNING, *A History of Golf* (1955), 34 chapters on various aspects of the game, including the start of the championships and the beginning of the game in the United States; HORACE G. HUTCHINSON, *Golf* (1890), includes contributions by Sir W. Simpson, Bt. Rt. Hon. A.J. Balfour, Andrew Lang, and others, discussing the history, celebrated links, and celebrated players; BERNARD DARWIN, *Golf* (1954), essays in the author's delightful style with extracts from various sources on great matches, temperament, and other subjects; J.B. SALMOND, *The Story of the R. and A.* (1956), an intimate narrative of the growth, functions, and administration of the world's most famous golf club, the Royal and Ancient Golf Club of St. Andrews; HENRY LONGHURST, *Only on Sundays* (1964), a collection of 70 articles by this popular and far-travelled writer that makes an entertaining miscellany of reportage, comment, and anecdote; GEOFFREY COUSINS, *Golfers at Law* (1958), an informative story about the rules of golf and the game itself (but not a book of rules) attractively written and based on research; HERBERT WARREN WIND (ed.), *The Complete Golfer* (1954), containing short stories, cartoons, and sections on great players, historic moments, etc.; DONALD STEEL (ed.), *The Golfer's Bedside Book* (1965), a collection of essays, experiences, reflections, and humorous anecdotes especially contributed by the best contemporary golfing writers of the day; *Golfer's Handbook* (annual), a mine of information, championship and tournaments results over the years, records, curiosities, brief biographies, and many other features; ROBERT SCHARFF (ed.), *Golf Magazine's Encyclopedia of Golf* (1970), the history of golf, results of major tournaments and championships, principles and rules of the game, championship golf courses, and other information.

Official rules of the game are published both by the R. and A. and the USGA.

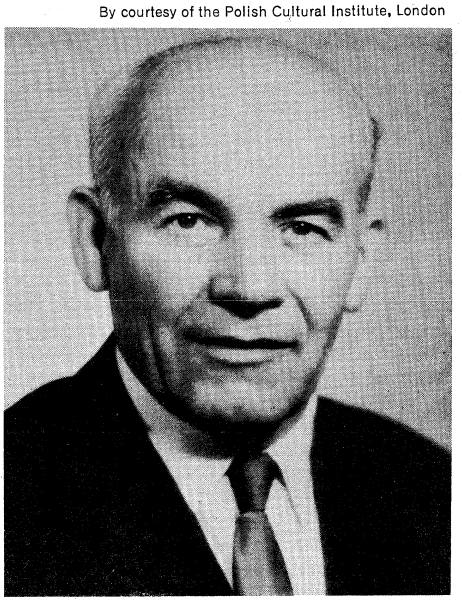
(F.Mo.)

Gomułka, Władysław

Władysław Gomułka stands as the central figure in Polish politics in the mid-20th century. Frail in appearance and

colourless in personality, with a modest, almost severe life style, Gomułka has had an uneventful private life. His only passion has been politics, but his public career has been full of dramatic turns. Emerging as the Communist leader of Poland at the end of World War II, he was later denounced and imprisoned by Premier Joseph Stalin of the U.S.S.R. Restored to power amid popular acclaim in 1956, for the next 14 years, until his ouster in late 1970, he attempted to steer a middle course between the liberals and neo-Stalinists. In victory as well as defeat he has left a deep personal imprint on Poland's recent history.

By courtesy of the Polish Cultural Institute, London



Gomułka.

Gomułka was born Feb. 6, 1905, in Białobrzegi, near Krosno in southern Poland. Before his birth his parents had emigrated to the United States but had returned disillusioned. His father, Jan, was a Socialist and worked in the oil fields. Gomułka completed primary school in 1917 and afterward was trained as a locksmith. From 1922 to 1926 he worked in an oil refinery. At the age of 16 he joined the youth Socialist movement. In 1926 he entered the clandestine Communist Party of Poland and in the same year was first arrested for revolutionary activity.

Professional revolutionary

At this time Gomułka became a professional trade union organizer, and in 1930 he was elected a national secretary of the Chemical Workers' Union. Thereafter he organized workers' strikes throughout the country. During the textile strike at Łódź in 1932, he was seriously wounded in the leg by the police and was left with a permanent limp. He was arrested and sentenced to four years' imprisonment but was released for reasons of health in 1934. In 1934–35 Gomułka studied at the International Lenin School in Moscow. After his return to Poland he continued revolutionary activity in Silesia, and in 1936 he was re-arrested and sentenced to seven years' imprisonment. When the Polish Communist Party was dissolved on Stalin's orders in 1938 and most of its leaders were exterminated in the Soviet Union, Gomułka stayed in prison in Poland. At the outbreak of World War II he was released. After participating in the defense of Warsaw, he moved to the Soviet-occupied eastern part of the country, where he worked as a minor official in a paper mill in Lwów.

Wartime leader

With the outbreak of war between Germany and the Soviet Union in 1941, Gomułka resumed his political activities. At first he returned to his native region of Krosno and organized the Communist underground there. In July 1942 he moved to Warsaw, where he became district secretary and a member of the Central Committee of the newly founded Polish Workers' Party (PPR). There he organized some daring attacks by the underground on the Nazi occupiers. In November 1943, after the arrest of his predecessor, Gomułka became secretary general of the Polish Workers' Party. He is credited with writing the Party's ideological manifesto and helping to establish the Home National Council in cooperation with some other leftist groups. When the Soviet troops entered Poland in July 1944, Gomułka moved to Lublin, where the Communist-dominated provisional government had been set up. In January 1945 he was appointed deputy premier, and in June he also assumed the portfolio of the Recovered Territories, with responsibility for the administration of all the lands taken over by Poland from Germany. In December 1945, at the First Congress of the Polish Workers' Party in Warsaw, Gomułka was elected a member of the Politburo and secretary general of the Central Committee.

Gomułka was ruthless in eliminating all opposition to the Communist rule. He personally led the struggle to crush the Polish Peasant Party (PSL), and he was a strong advocate of the merger, on Communist terms, of the Polish Socialist Party (PPS) and the Polish Workers' Party. At the same time, however, he came out against forcible collectivization of agriculture and spoke favourably of the socialist tradition. In opposing the formation of the Cominform in September 1947, he was even critical of the Soviet line. This led to his political eclipse.

Persecution by Stalin

On Stalin's orders, Gomułka was accused of "nationalist deviation," and in September 1948 he was replaced as secretary general of the Polish Workers' Party by Bolesław Bierut. After the Communist and Socialist parties merged into the Polish United Worker's Party (PZPR) in December 1948, Gomułka was also dropped from the Politburo. In January 1949 he was relieved of his government posts, and in November of the same year he was stripped of his membership in the Communist Party. Finally, in July 1951 he was arrested. Throughout his persecution—even when imprisoned, his life clearly in peril—Gomułka acted in a dignified and courageous manner and refused to admit his guilt.

Toward the end of 1954, after Stalin's death, Gomułka was released, and in 1956, after Premier Nikita Khrushchev had launched the de-Stalinization campaign in February and Bierut had died in March, he was politically rehabilitated. In April the new party secretary, Edward Ochab, reiterated the charges of "nationalist deviation" against Gomułka but admitted he should not have been arrested. After the Poznań workers rioted against the Communist government in June, Gomułka's political fortunes started to rise once again. His persecution by Stalin had turned Gomułka into a popular figure among the Poles, and they now demanded that he be restored to power.

Restoration to power

In the tense atmosphere prevailing in the country, the Communist leaders acceded to the popular wishes. In August 1956 Gomułka was re-admitted to the party and in October was re-elected to the Politburo and to the position of first secretary of the Central Committee. Soon he was also elected a member of Poland's collective presidency, the Council of State. His return to power was a moment of great personal triumph for Gomułka. Hoping that he would undertake substantial reforms, the people gave him their almost universal support.

The reforms adopted by Gomułka were half-hearted. The most oppressive Stalinist features were eliminated: the rule of terror was curbed, the persecution of the Catholic Church was ended, and the collectivization of agriculture was abandoned. Several objectionable features of the older system were, however, preserved: intellectual freedom remained restricted, and no major economic reform was carried out. His retrogressive course led to disillusionment with Gomułka among the Poles, but in the late 1950s many people still believed his policies resulted from pressure from Moscow.

In 1961, after Khrushchev launched his second de-Stalinization campaign, Gomułka failed to exploit this opportunity to undertake further reforms and the situation in Poland remained stagnant. From then on, Gomułka's popular support declined rapidly. The ferment among the people steadily gathered strength until it

Waning of popularity

culminated in March 1968 in the open defiance of the Gomułka regime by the intellectuals and in students' riots in Warsaw and several other Polish cities.

Gomułka survived the crisis and at the Fifth Party Congress in November 1968 was re-elected first secretary, but his political influence was clearly on the wane. He was discredited among a large segment of the people and was challenged by powerful rivals within the party leadership. Gomułka tried to stave off defeat by belatedly adopting some new policies. In 1969 he changed Poland's policy vis-à-vis West Germany, leading to the signing early in December 1970 of a Polish–West German treaty normalizing relations between the two countries and sanctioning the Polish western boundary. At the same time, he launched substantial economic reforms, but by now the Polish economy was so run down that drastic measures were needed. The announcement of increased food prices on the eve of the Christmas holiday led to workers' riots in the coastal cities of Gdańsk, Gdynia, and Szczecin. As in 1956, the new ferment in the country resulted in a change in the top party leadership. On Dec. 20, 1970, Gomułka was ousted as first secretary and was replaced in that post by Edward Gierek.

BIBLIOGRAPHY. NICHOLAS BETHELL, *Gomulka: His Poland, His Communism* (1969), a comprehensive biography of Gomułka from his youth until his political eclipse in the late 1960s; ADAM BROMKE, *Poland's Politics: Idealism vs. Realism* (1967), an attempt at a systematic analysis of Polish postwar politics with a detailed review of Gomułka's programs and policies up to the mid-1960s; M.K. DZIEWANOWSKI, *The Communist Party of Poland: An Outline of History* (1959), a scholarly history of the Polish Communist Party from its inception to the mid-1950s with considerable attention paid to Gomułka's role in it; HANSJAKOB STEHLE, *The Independent Satellite* (Eng. trans. 1965), an extremely informative journalistic presentation of the political situation in Poland in the postwar period, with special focus on the late 1950s and early 1960s.

(A.Br.)

Göring, Hermann

Nominated by Adolf Hitler in 1939 to be his eventual successor, Hermann Göring throughout all of his service to the Nazi Party ranked high among its leadership. At the International Military Tribunal in Nürnberg in 1945–46, he was universally recognized as the principal defendant and spokesman for the Nazi regime. Before Hitler attained the German chancellorship in 1933, Göring had been a prime mover in achieving support for the party among the German industrialists; and, as president of the Reichstag (the lower house of the German parliament) from August 1932, he had served the party's interests. Once the Nazi regime was established, he successfully developed the Nazi police state in its initial form in Prussia, created Germany's new air force (the Luftwaffe), supervised Germany's war economy and rearmament, and played a leading part in promoting Germany's expansion into Austria and eastern Europe.

Göring was born in Bavaria on January 12, 1893, the second son by the second wife of Heinrich Ernst Göring, at the time German consul general in Haiti. The family was reunited in Germany on the father's retirement in 1896. Göring, as a child, was brought up near Nürnberg, in the small castle of Veldenstein, whose owner was Hermann, Ritter von Epenstein, a Jew who was, until 1913, the lover of Göring's mother and the godfather of her children. Trained for an army career, Göring received his commission in 1912 and served with distinction during World War I, joining the embryonic air force. In 1918 he became commander of the celebrated squadron in which the great German aviator Manfred, Freiherr von Richthofen, served. Göring so deeply resented the treatment given army officers by the civilian population during the troubled period after Germany's capitulation that he left the country. After a period as a commercial pilot in Denmark and Sweden, he met the Swedish Baroness Carin von Rosen, who divorced her husband and married Göring in Munich on February 3, 1922.

Göring had met Hitler the previous year and had joined the small National Socialist German Workers' (Nazi) Party late in 1922. As a former officer, he had been given command of Hitler's Storm Troopers (the SA, Sturmabteilung). Göring took part in the abortive Munich *Putsch* of November 1923 in which Hitler tried to seize power prematurely. During the *Putsch*, Göring was badly wounded in the groin. His arrest was ordered, but he escaped with his wife into Austria. Given morphine to deaden the pain from his wounds, he became so severely addicted that he twice underwent treatment in 1925–26 at Långbro mental hospital in Sweden.



Göring, as commander of the Sturmabteilung, 1933.
Heinrich Hoffmann, Munich

Göring's drug addiction

He returned to Germany in 1927, where his contacts in German industry proved useful, and he was taken back into the party leadership. He occupied one of the 12 Reichstag seats that the Nazi Party won in the 1928 election. Thereafter, Göring became the acknowledged party leader in the lower house, and, when the Nazis won 230 seats in the election of July 1932, he was elected president of the Reichstag.

Göring's sole concern in the Reichstag was to stultify the democratic system, which the Reichstag ostensibly represented up to March 1933. He had the ear of 84-year-old president Paul von Hindenburg and used his position to outmanoeuvre the successive chancellors, particularly Kurt von Schleicher and Franz von Papen, until Hindenburg was finally forced to invite Hitler to become chancellor on January 30, 1933. The battle for dictatorial power, however, was still not won; between January 30 and March 23, when an enabling bill giving Hitler his dictatorial powers was passed, Göring was tirelessly active. He used his new position as minister of the interior in Prussia, Germany's largest and most influential state, to Nazify the Prussian police and establish the Gestapo, or secret political police. He also established concentration camps for the "corrective treatment" of difficult opponents. The Reichstag fire of February 27, 1933, which the Nazis most probably instigated, made it possible for Göring to accuse the Communists of intending a coup d'état. The wholesale arrest of Communist and even some Social Democrat deputies succeeded in removing any effective opposition to the passage of the enabling bill the following month.

Göring's position as Hitler's most loyal supporter remained unassailable for the rest of the decade. He collected offices of state almost at will. He was *Reich* commissioner for aviation and head of the newly developed Luftwaffe, which was disguised as a civilian enterprise until March 1935. In 1933 he became master of the German hunt and of the German forests. In June 1934 he took a leading part in the party's purge of the SA leader Ernst Röhm but in the same year ceded his position as security chief to Heinrich Himmler, thus ridding himself of responsibility for the Gestapo and the concentration

camps. In 1937 he displaced Hjalmar Schacht, after 1934 Hitler's minister for economic affairs. Later, in 1936, without consulting Schacht, Hitler made Göring commissioner for his Four-Year Plan for the war economy. He was also constantly employed as Hitler's roving ambassador.

Popularity among the Nazi leaders

Göring was the most popular of the Nazi leaders, not only with the German people but also with the ambassadors and diplomats of foreign powers. He used his impregnable position to enrich himself. The more ruthless aspect of his nature showed in the recorded telephone conversation by means of which he blackmailed the surrender of Austria before the *Anschluss* (political union) with Germany in 1938. It was Göring who led the economic despoliation of the Jews in Germany and in the various territories that fell under Hitler's power.

Göring's first wife had died in 1931, and on April 10, 1935, he married the actress Emmy Sonnemann. Göring was devoted in turn to each of his wives. His hunting interests enabled him to obtain a vast forest estate in the Schorfheide, north of Berlin, where from 1933 he developed a great baronial establishment on a scale commensurate with his ambitions. This he called Carinhall in honour of his first wife. It was at Carinhall that he kept the greater part of his enormous art collection. On June 2, 1938, Emmy bore him a daughter, his only child, Edda.

Although Göring was probably sincere in his desire to avert or postpone war—as his abortive negotiations in 1939 with the Swedish industrialist Birger Dahlerus indicate—it was his Luftwaffe that conducted the blitzkrieg that smashed Polish resistance and softened up country after country as Hitler's campaigns developed. But Göring's self-indulgent nature was too weak to sustain the rigours of war or oppose Hitler's blind prejudice in favour of the production of bombers rather than fighter planes. The Luftwaffe's capacity for defense declined as Hitler's battlefronts extended from northern Europe to the Mediterranean and North Africa, and Göring lost face when the Luftwaffe failed to win the Battle of Britain or prevent the Allied bombing of Germany. On the plea of ill health, Göring retired as much as Hitler would let him into private life, enjoying the luxuries of Carinhall, where he continued to amass his art collection (further enriched with spoils from the Jewish collections in the occupied countries) and receive many gifts from those who sought his favour. His colossal girth was more the result of glandular defect than of gluttony, but his excessive resort to paracodeine tablets (a mild derivative from morphine) poisoned his system and made recurrent treatment for drug addiction necessary. His addiction helped to make him alternately elated or depressed; he was egocentric and bombastic, delighting in flamboyant clothes and uniforms, decorations, and exhibitionist jewelry.

Decline of the Luftwaffe

In spite of Göring's faults, Hitler felt he could not afford to discard a man so closely identified with the regime. In 1939 he had declared him his successor and in 1940 had given him the special rank of *Reichsmarschall des Grossdeutschen Reiches*. The other Nazi leaders both resented his favoured position and despised his self-indulgence, but Hitler did not displace him until the last days of the war, when, in accordance with the decrees of 1939, Göring attempted to assume the Führer's powers, believing him to be encircled and helpless in Berlin. Nevertheless, Göring expected to be treated as a plenipotentiary when, after Hitler's suicide, he surrendered himself to the Americans.

Cured finally of his drug addiction during his period of captivity awaiting trial as a war criminal, he defended himself ably before the International Military Tribunal at Nürnberg. He saw himself as the star defendant, a historical figure; he denied any complicity in the more hideous activities of the regime, which he claimed to be the secret work of Himmler. When after his condemnation his plea to be shot and not hanged was refused, he took poison and died in his cell at Nürnberg on October 15, 1946, the night his execution was ordered. Only in 1967 was it revealed that he had left a note explaining that the poison capsule had been secreted all the while in a container of pomade.

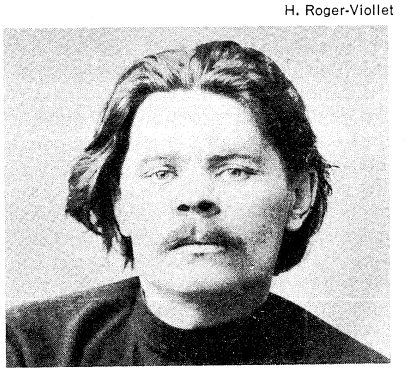
BIBLIOGRAPHY. The fullest life of Göring to date is R. MANVELL and H. FRAENKEL, *Hermann Göring* (1962), with a full bibliography. First-hand accounts of Göring appear in many books, notably in ADOLF GALLAND, *Die Ersten und die Letzten* (1953; Eng. trans., *The First and the Last*, 1955); G.M. GILBERT, *Nuremberg Diary* (1947); NEVILE M. HENDERSON, *Failure of a Mission* (1940); ERNST HANFSTAENGL, *Hitler: The Missing Years* (1957); FRANZ VON PAPEN, *Der Wahrheit eine Gasse* (1952; Eng. trans., *Memoirs*, 1952); HJALMAR SCHACHT, *Abrechnung mit Hitler* (1948; Eng. trans., *Account Settled*, 1949), and *76 Jahre meines Lebens* (1953; Eng. trans., *My First Seventy-Six Years*, 1955); ALBERT SPEER, *Inside the Third Reich* (1970); and in the sentimental recollections of Göring's sister-in-law, FANNY VON WILAMOWITZ-MOELLENDORFF, *Carin Göring* (1934). Göring's own full statements appear in the official record of the International Military Tribunal, Nürnberg (1947–49).

(R.M./H.Fra.)

Gorky, Maksim

Maksim Gorky is the pen name of the Russian author Aleksey Maksimovich Peshkov, whose plays, novels, and memoirs of working class life brought him an international reputation. He was the only Soviet writer whose work embraced the prerevolutionary and postrevolutionary period so exhaustively, and though he by no means stands with Chekhov, Tolstoy, and others in the front rank of Russian writers, he remains one of the most important literary figures of his age.

H. Roger-Viollet



Gorky.

LIFE

He was born on March 28 (March 16, old style), 1868, at Nizhny Novgorod, since renamed Gorky in his honour; and his earliest years were spent in Astrakhan, where his father, a former upholsterer, became a shipping agent. When the boy was five his father died; Gorky returned to Nizhny Novgorod to live with his maternal grandparents, who brought him up after his mother remarried. The grandfather was a dyer whose business deteriorated and who treated Gorky harshly. It was from his grandmother that he received most of what little kindness he experienced as a child. The bitterness of these early experiences later led him to choose the word *gorky* ("bitter") as his pseudonym. Technically of lower middle class origin, he lived in such poverty as a child and young man that he is often considered the greatest "proletarian" in Russian literature. This circumstance, coinciding with the rise of working class movements all over the world, helped to give Gorky an immense literary reputation, which his works do not wholly merit.

He knew the Russian working class background intimately, for his grandfather afforded him only a few months of formal schooling, sending him out into the world to earn his living at the age of eight. His jobs included, among many others, work as assistant in a shoemaker's shop, as errand boy for an icon painter, and as dishwasher on a Volga steamer, where the cook introduced him to reading—soon to become his main passion in life. Frequently beaten by his employers, nearly always hungry and ill clothed, he came to know the seamy

side of Russian life as few other Russian authors before or since. His late adolescence and early manhood were spent in Kazan, where he worked as a baker, docker, and night watchman. It was here that he made his first contact with Russian revolutionary ideas, meeting representatives of the Populist movement, whose tendency to idealize the Russian peasant he came to reject. During this period, oppressed by the misery of his surroundings, he attempted suicide by shooting himself. Leaving Kazan at the age of 21, he became a tramp, doing odd jobs of all kinds during extensive wanderings through south Russia.

First stories

It was in Tbilisi (Tiflis) that Gorky began to publish stories in the provincial press, of which the first was "Makar Chudra" (1892), followed by a series of similar wild Romantic legends and allegories, which have now only a documentary interest. But with the publication of the story "Chelkash" (1895; first Eng. trans., 1902) in a leading St. Petersburg journal, a success story began as spectacular as any other in the history of Russian literature. "Chelkash" itself remains one of his outstanding works and is the story of a colourful harbour thief in which elements of Romanticism and realism are mingled. It began Gorky's celebrated "tramp period," during which he described the social dregs of Russia. He expressed sympathy and self-identification with the strength and determination of the individual hobo or criminal, thus tapping a new vein in Russian literature, for such characters had previously been described more from the outside. Also to the tramp period belong "Malva" (1897), "Byvshye lyudi" (1897; *Creatures That Once Were Men*), and "Dvadtsat shest i odna" (1899; Eng. trans., "Twenty-Six Men and a Girl," 1902). The last, describing the sweated labour conditions in a bakery, is often regarded as his best short story. So great was the success of these works that Gorky's reputation quickly soared, and he began to be spoken of almost as an equal of Tolstoy and Chekhov.

Plays and novels

With the turn of the century Gorky embarked on a series of plays and novels, all less excellent as works of art than his best earlier stories. The first novel, *Foma Gordeyev* (1899), continues to illustrate his admiration for strength of body and will in the person of the masterful barge owner and rising capitalist Ignat Gordeyev, who is contrasted with his relatively feeble and intellectual son Foma, a "seeker after the meaning of life," as are many of Gorky's other characters. From this point, the rise of Russian capitalism became one of Gorky's main fictional interests. Other novels of the period are *Troye* (1900; *Three of Them*, 1905), *Ispoved* (1908; *A Confession*, 1910), *Gorodok Okurov* (1909; "Okurov City"), and *Zhizn Matveya Kozhemyakina* (1910; "The Life of Matvey Kozhemyakin"). These are all to some extent failures because of Gorky's inability to sustain a powerful narrative such as he was well able to begin, and also because of a tendency to overload his work with irrelevant discussions about the meaning of life. *Mat* (1906; Eng. trans., *Mother*, 1911) is probably the least successful of the novels, yet has considerable interest as Gorky's only long work devoted to the Russian revolutionary movement. It was made into a notable silent film by Vsevolod Pudovkin in 1926 and dramatized by Bertolt Brecht in *Die Mutter*, 1930–31. Simultaneously, Gorky was producing a series of plays—*Na dne* (1902; Eng. trans., *The Lower Depths*, 1906), *Vassa Zheleznova* (1910), *Dachniki* (1905; "Summer Residents"), *Vragi* (1906; "The Enemies"), and others. The most famous of these is *The Lower Depths*, which still enjoys great success abroad and in the U.S.S.R., putting on the stage the kind of flophouse character that Gorky had already described so extensively in narrative fiction.

Between 1899 and 1906 Gorky was living mainly in St. Petersburg (Leningrad), where he became a Marxist. He accordingly gave much enthusiastic support to Russian Marxism as represented by the Social Democratic Party. After the split in that party in 1903, Gorky supported its Bolshevik wing. But he was often at odds with the Bolshevik leader Lenin. Nor did Gorky ever, formally, become a member of Lenin's party, though his enormous earnings, which he largely gave to party funds, were for a time one of that organization's main sources of income. In 1901 the Marxist review *Zhizn* ("Life") was suppressed for publishing a short revolutionary poem by Gorky, "Pesnya o burevestnike" ("Song of the Stormy Petrel"). Gorky was arrested but released shortly afterward and went to the Crimea, having developed tuberculosis. In 1902 he was elected a member of the Russian Academy of Sciences, but his election was soon withdrawn for political reasons, an event that led to the resignations of Chekhov and Korolenko from the academy. At about this time Gorky founded a publishing business called *Znaniye* ("Knowledge"), which led to the emergence of a movement sometimes called the *Znaniye* school of fiction. *Znaniye* aimed to give a forum, insofar as censorship conditions permitted, to young writers with revolutionary proclivities whose work was "tendentious," a word that has commonly been used as a term of praise by Russian critics and readers. Gorky took a prominent part in the 1905 revolution and was arrested in the following year, being again quickly released, partly as the result of protests from abroad. He toured America in the company of his mistress, an event that led to his partial ostracism there and to a consequent reaction on his part against the United States expressed in stories about New York, *The City of the Yellow Devil* (1906).

Years of exile

On leaving Russia in 1906, Gorky spent seven years as a political exile, living mainly in his villa on Capri, which became an intellectual centre for politically disaffected Russians. Meanwhile, although his writings continued to enjoy the favour of ordinary Russian readers, he had lost much of his former popularity with the intelligentsia, who were becoming conscious of Gorky's defects. Though still essentially in alliance with Lenin's movement, he was somewhat out of favour even there because of his espousal of a religio-philosophical trend called *bogostroitelstvo* ("God-building"), preached in his novel *Confession* and regarded as a heresy by more orthodox Marxists. Politically, Gorky was a nuisance to his fellow Marxists because of his insistence on remaining independent, but his great influence was a powerful asset, which from their point of view outweighed such minor defects. During World War I he agreed with the Bolsheviks in opposing Russia's involvement in the struggle. But he opposed the Bolshevik seizure of power in November 1917, and went on to attack the victorious Lenin's dictatorial methods on the pages of his newspaper *Novaya zhizn* ("New Life") until July 1918, when his protests were silenced by the censorship on Lenin's orders. From 1919 onward Gorky cooperated with Lenin's government, also helping to relieve the miseries that his fellow writers shared with the rest of the population during the early postrevolutionary years. He was often able to see that they received payment, if only for work such as translating. He also did much during that time to ensure the preservation of works of art from destruction.

Autobiography

It is to the decade ending in 1923 that the production of Gorky's greatest masterpiece belongs. This is the autobiographical trilogy *Detstvo* (1913–14; Eng. trans., *My Childhood*, 1915), *V lyudyakh* (1915–16; Eng. trans., *In the World*, 1917), and *Moi universitety* (1923; Eng. trans., *My Universities*, 1952)—the last title being sardonic because Gorky's only university had been that of life, and his wish to study at Kazan University had been frustrated. This long work is one of the finest autobiographies in Russian. It deals only with the years of Gorky's childhood and early manhood and shows his strength as a now relatively extroverted writer, who had to some extent turned his back on the excessive philosophizing of his early period. It reveals him as an acute observer of detail with a flair for describing people—his own family, his numerous employers, and a panorama of minor but memorable figures who flit across his pages. In a way it is hardly the story of Gorky himself, for seldom was an autobiography more reticent (even his attempted suicide receives only a line or two). The book is permeated with a wonder at the mystery, cruelty, and colour of life, which, it might seem, Gorky was now less earnestly eager to understand or interpret, being content

to portray. But it does contain numerous messages, which he now tends to imply rather than to preach openly: notably protests against motiveless cruelty, continued emphasis on the importance of toughness and self-reliance ("I very early realized that a man is made by his resistance to the milieu which surrounds him"), and musings on the value of hard work often couched in his characteristic rhetorical style, as where he speaks of the dockers' "drunken joy" in unloading a Volga barge, a joy "than which only the embraces of a woman are sweeter."

My Universities was finished in Italy, to which Gorky emigrated. His villa in Sorrento became his base in the period 1921–28, when he made excursions to Germany and elsewhere, but he did not return to Russia. One reason for his retirement was poor health, but a disillusionment with the Soviet Union in the first years after the Revolution seems to have played a part in his decision. In 1928 he yielded to pressures to return, and the lavish official celebration there of his 60th birthday in that year was beyond anything he could have expected.

Return to the U.S.S.R.

In the following year he returned to the U.S.S.R. permanently, and he lived there until his death in 1936. His return coincided with the establishment of Stalin's total ascendancy, and Gorky became a prop of Stalinist political orthodoxy. He was now more than ever the undisputed leader of Soviet writers, and when the Soviet Writers' Union was founded in 1934 he became its first president. At the same time he helped to found the literary method of Socialist Realism, which was now imposed on all Soviet writers, and which obliged them—in effect—to become outright political propagandists.

Gorky remained active as a writer. Despite his close association with official Stalinist literary doctrine, almost all his later fiction is concerned with the period before the Revolution. In *Delo Artamonovykh* (1925; Eng. trans., *The Artamonov Business*, 1948), one of his best novels, he showed his continued interest in the rise and fall of prerevolutionary Russian capitalism. The immense and more ambitious *Zhizn Klima Samgina* (1927–36; "Life of Klim Samgin") is a tetralogy that attempts a portrait of the Russian intelligentsia between 1870 and 1924. There were more plays—*Yegor Bulychov i drugiye* (1932); *Dostigayev i drugiye* (1933; "Dostigaev and Others"; Eng. trans., *The Last Plays*, 1937)—but the most generally admired work to follow his autobiography is a volume of reminiscences of Russian writers (Eng. trans., *Reminiscences of Tolstoy, Chekhov and Andreyev*, 1949). Here the memoir of Tolstoy is so lively and free from the hagiographical approach traditional in Russian studies of their leading authors that it has sometimes been acclaimed as Gorky's masterpiece. Almost equally impressive is Gorky's study of Chekhov. At the other end of the scale come some of his pamphlets devoted to topical events and problems, such as his "Belomorkanal" (1934; Eng. trans., "The White Sea Canal," 1935), in which, as elsewhere, he glorified the most brutal aspects of Stalinism—in this case the construction of the canal by the forced labour of political prisoners.

Studies of writers

Some mystery attaches to Gorky's death, which occurred suddenly in 1936 while he was under medical treatment. Whether his death was natural or not is unknown, but it came to figure in the trial of Bukharin and others in 1938, at which it was claimed that Gorky had been the victim of an anti-Soviet plot by the "Bloc of Rightists and Trotskyites." The former police chief Yagoda, who was among the defendants, confessed to having ordered his death. Some Western authorities have suggested that Gorky was done to death on Stalin's orders, having finally become sickened by the excesses of Stalinist Russia, but there is little evidence of this except that it was characteristic of Stalin to frame others on the charge of accomplishing his own misdeeds.

ASSESSMENT

After his death Gorky was canonized as the patron saint of Soviet letters, the formula "Gorky said . . ." often being used to clinch a literary argument. His reputation abroad has also remained high, but it is doubtful whether posterity will deal with him so kindly. His success was partly due, both in the Soviet Union and to a lesser extent abroad, to political accident. His style, though gradually improving through the years, retained its original defects of excessive striving for effect, of working on the reader's nerves by the piling up of emotive adjectives, and by tending to overstate. Chekhov, his opposite in these matters, had given him good advice in the correspondence that followed Gorky's appeal for literary advice in the 1890s, but this counsel was only partly effective. Among Gorky's other defects, in addition to his weakness for philosophical digressions mentioned above, are his lack of any sense of humour and a certain coarseness of emotional grain. But his eye for physical detail, his talent for making his characters live, and his unrivalled knowledge of the Russian "lower depths" are weighty items on the credit side.

MAJOR WORKS

NOVELS: *Foma Gordeyev* (1899; *Foma Gordyeeff*, 1901; *The Man Who Was Afraid*, 1905; *Foma Gordeyev*, 1956); *Troye* (1900; *Three of Them*, 1905; *The Three*, 1959); *Mat* (1906; *Comrades*, 1907; *Mother*, 1911, 1921, 1950); *Zhizn nenuzhnago cheloveka* (1907–08; *The Spy*, 1908, 1969); *Ispoved* (1908; *A Confession*, 1910); *Leto* (1909); *Gorodok Okurov* (1909); *Zhizn Matveya Kozhemyakina* (1910–11); *Khozyain* (1913); *Delo Artamonovykh* (1925; *Decadence*, 1927; *The Artamonov Business*, 1948; *The Artamonovs*, 1952); *Zhizn Klima Samgina*, 4 pt. (1927–36; the parts being severally entitled in English: *Bystander*, 1930, *The Magnet*, 1931, *Other Fires*, 1933, and *The Specter*, 1938).

SHORT STORIES: The collection of sketches and stories *Ocherki i rasskazy*, 2 vol. (1898), included reissues of Gorky's first publication, "Makar Chudra" (1892), and of his first recognized success, "Chelkash" (1895; *Tchelkash*, 1902). Other notable stories are "Goremyka Pavel" (1894; *Orphan Paul*, 1946); "Starukha Izergil" (1895; *The Old Woman Izergil*, 1906); "Byvshye lyudi" (1897; *Creatures That Once Were Men*, 1905); "Suprugi Orlovy" and "Malva" (both also 1897; *The Orloff Couple* and *Malva*, 1901); "Dvadtsat shest i odna" (1899; *Twenty-Six Men and a Girl*, 1902, 1915); and the two series known as *Skazki ob Italii* (1911–13), and *Russkiye skazki* (1912).

PLAYS: *Meshchane* (1901); *Na dne* (1902; *The Lower Depths*, 1906, 1912, 1923; *At the Bottom*, 1930; *Down and Out*, 1933); *Dachniki* (1905); *Deti solntsa* (1905); *Varvary* (1905); *Vragi* (1906); *Posledniye* (1908); *Vassa Zheleznova* (first version 1910, second 1935); *Deti* (1910); *Chudaki* (1910); *Zykovy* (1913); *Somov i drugiye* (1931); *Yegor Bulychov i drugiye* (1932; *Yegor Bulichoff and Others*, 1937); *Dostigayev i drugiye* (1933; *Dostigaeff and the Others*, 1937).

AUTOBIOGRAPHY AND REMINISCENCES: *Detstvo* (1913–14; *My Childhood*, 1915; *Childhood*, 1950), *V lyudyakh* (1915–16; *In the World*, 1917), and *Moi universitety* (1923; *Reminiscences of My Youth*, 1924; *My Universities*, 1952), together constitute the work republished in English as *The Autobiography of Maxim Gorky* (1953). Supplementary to them are *Vremya Korolenko* (1923), *O pervoy lyubvi* (1923), and *Zametki iz dnevnika* (1924; *Fragments from My Diary*, 1924). Personal acquaintance is the basis of Gorky's *A.P. Chekhov: otryvki iz vospominanii* (1905; *Anton Tchekhov: Fragments of Recollections*, 1921), *Vospominaniya o Tolstom* (1919; *Reminiscences of Leo Nicolayevitch Tolstoi*, 1920), and *Vladimir Ilich Lenin* (1924 with supplements to 1931; *V. I. Lenin*, 1931; *Days with Lenin*, 1933, 1944; *Lenin: A Biographical Essay*, 1969).

POETRY: *Devushka i Smert* (written 1892 but not printed till 1917); *Pesnya o Burevestnike* (1901); *Chelovek* (1902).

OTHER WORKS: Some of Gorky's anti-American satires of 1906 are still remembered: *Odin iz koroley Respubliki; V Amerike;* and the series of New York sketches *Gorod zheltago dyabola*. In his lifetime collections were made of his numerous articles on current affairs, *Publitsisticheskiye stati* (1933), and of his pronouncements on literary questions, *O literature: stati i rechi 1928–35 gg.* (1935; expanded edition, . . . 1928–36 gg., 1937). His complete works, in 30 volumes, were published 1949-55. (R.F.Hi.)

BIBLIOGRAPHY. Major bibliographies on Gorky's life and work and on Gorky criticism are BORISS A. KALEPS (comp.), *Maxim Gorky (1868–1936): A Bibliography of Publications from and on Gorky in English, French, German, Italian, Spanish, and Latvian Languages* (1963); К.Д. МУРАТОВА, *Семинарий по Горькому* (1956), a comprehensive bibliography of almost exclusively Russian and Soviet works, arranged by topics; and С.Д. БАЛУХАТИЙ, *Критика о М.*

Горьком (1934), splendidly comprehensive within its chronological limitations (1893–1932). A very useful selective bibliography in English may be found in Weil (see below). The vast majority of manuscripts, personal papers, and other archival material relating to Gorky is kept at the Gorky Institute of World Literature in Moscow, unquestionably the major world centre of research on Gorky and his works.

Editions and translations: The best and most comprehensive collected edition of Gorky's work, including correspondence, is *Собрание сочинений в тридцати томах,* 30 vol. (1949–55). This is in most respects an admirably scholarly academic edition, though with notable, if inevitable, omissions—*e.g.*, Gorky's diatribes against Lenin and the Bolsheviks and some anti-Soviet pieces. Gorky's works have been widely translated into all the major languages of the world. Only insignificant fragments of his fiction remain untranslated into English. There are, however, no collected editions of his works in English translation and no major translator whom one would single out from the many. A useful short list of English translations of Gorky's works may be found in the bibliographical appendix to Wolfe (see below).

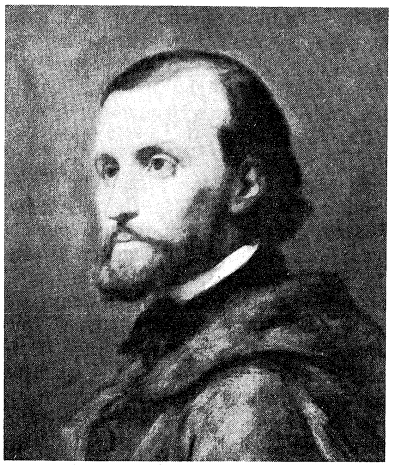
Biographical and critical studies: АКАДЕМИЯ НАУК СССР (Институт мировой литературы), *Летопись жизни и творчества А.М. Горького,* 4 vol. (1958–60), an impressive and valuable collection of biographical information, though highly selective—omits all biographical detail that might be interpreted as contradicting the official Soviet view of Gorky; И.А. ГРУЗДЕВ, *Горький и его время* (1938), an important biography by an outstanding Soviet biographer; F.M. BORRAS, *Maxim Gorky the Writer* (1967), a generally sound critical interpretation, though the material is rather weakly organized; NINA GOURFINKEL, *Gorki par lui-même* (1957; Eng. trans., *Gorky,* 1960), a short, fragmentary work, though with some valuable insights; RICHARD HARE, *Maxim Gorky, Romantic Realist and Conservative Revolutionary* (1962), marred by some rather unbalanced critical attitudes and a general sense of authorial antipathy toward his subject; A.S. KAUN, *Maxim Gorky and His Russia* (1931), a fascinating and sensitive account of Gorky and his career by an American who knew him—contains an outstanding section on Gorky's often stormy relationship with Lenin and the Bolsheviks and Gorky's ideological deviations; В.Т. ХОДАСЕВИЧ, «Горький (воспоминания),» *Современие записки,* 63:131–156 (1939), a profoundly personal—and yet, one feels, very accurate—portrait of a close friend by one of the great literary figures of the Russian émigré world; DAN LEVIN, *Stormy Petrel: The Life and Work of Maxim Gorky* (1965), a generally well-researched and sensitive account of Gorky the man, his life and career—much stronger on the biographical than the critical side and particularly vitiated by infelicities of style; IRWIN WEIL, *Gorky: His Literary Development and Influence on Soviet Intellectual Life* (1966), a good account of Gorky's outstanding features as a writer and an extremely interesting, if somewhat speculative, attempt to trace his influence on the general development of Soviet literature as well as individual Soviet writers; B.D. WOLFE, *The Bridge and the Abyss: The Troubled Friendship of Maxim Gorky and V.I. Lenin* (1967), a fascinating account of a complex relationship that revealed vital aspects of personality in both men.

Gounod, Charles

Among the group of French composers who upheld and developed the tradition of opera in France in the third quarter of the 19th century, Charles François Gounod is one of the most significant. He declared that "The composer who would achieve a successful career must create it through writing operas." Although his many other operatic works no longer enjoy their former popularity, his *Faust* maintains its worldwide reputation.

Gounod was born in Paris on June 17, 1818. His father, François Gounod, was a painter of some distinction. His mother, a woman of wide education, was a capable pianist and gave Gounod his early training in music. He was educated at the Lycée Saint-Louis, where he remained until 1835. After taking his degree in philosophy, he began to study music with the Bohemian composer Anton Reicha. On Reicha's death Gounod entered the Paris Conservatoire, where he studied under Fromental Halévy and Jean-François Lesueur. Three years later his cantata *Fernand* won him the Prix de Rome for music, an award that entailed a three-year stay in Rome at the Villa Medici.

In Italy he devoted much of his attention to the works of Giovanni da Palestrina, an Italian Renaissance composer, and was so deeply influenced by him that he wrote a mass in imitation of Palestrina's style; it was among his earliest important compositions. From Rome he proceeded to Vienna, where a mass and requiem, composed in Italy, were performed in 1842 and 1843. Returning to Paris, he passed through Prague, Dresden, and Berlin. He also spent four days in Leipzig with Felix Mendelssohn and heard a performance of Mendelssohn's *Scottish Symphony* as well as a Mendelssohn recital of organ works by Bach in the Thomaskirche.



Gounod, oil portrait by Ary Scheffer. In the Musée National de Versailles.
By courtesy of the Musee National de Versailles

Influence of Palestrina

In Paris, Gounod became organist and choirmaster at the Church of the Missions Étrangères, and for two years he mainly studied theology. In 1846 he entered the seminary of Saint-Sulpice but in 1847 decided against taking Holy Orders. A requiem and a *Te Deum* he had started writing the previous year remained unfinished, and he turned to composing for the operatic stage.

The reception of his earliest operas, *Sapho* (1851) and *La Nonne sanglante* (1854), was not very enthusiastic, in spite of favourable reviews by the composer Berlioz. In his *Messe de Sainte-Cécile* (1855) he attempted to blend the sacred with a more secular style of composition. An excursion into comic opera followed with *Le Médecin malgré lui* (1858; English title, *The Mock Doctor*), based on Molière's comedy. From 1852 Gounod worked on *Faust*, the production of which, on March 19, 1859, marked a new phase in the development of French opera. This work has continued to overshadow all of Gounod's subsequent stage works, including the fairly successful *Mireille*, based on a Provençal poem by Frédéric Mistral (1864), *Roméo et Juliette* (1867), and his later oratorios.

In 1852 Gounod had become conductor of the Orphéon Choral Society in Paris, for which he wrote a number of choral works, including two masses. From 1870 he spent five years in London, formed a choir to which he gave his name (and which later on became the Royal Choral Society), and devoted himself almost entirely to the writing of oratorios. *Gallia*, a lamentation for solo soprano, chorus, and orchestra, inspired by the French military defeat of 1870, was first performed at the Albert Hall, London, at the opening of the International Exhibition in London on May 1, 1871, and the oratorios *La Rédemption* and *Mors et Vita* (*Life and Death*), at the Birmingham festival, in 1882 and 1885.

He was made a *grand officier* of the Légion d'Honneur in 1888, and died at Saint-Cloud, near Paris, on October 18, 1893.

Assessment

Gounod's melodic vein is unmistakably original, though often oversentimental. He knew how to write for the voice and for the orchestra; but in his operas his sense of musical characterization, though rarely devoid of charm, is often excessively facile, and the religiosity displayed in his sacred music too often superficial. His *Meditation* (*Ave Maria*) superimposed on Bach's *Prelude in C Major* (from *The Well-Tempered Clavier*, Book I) illustrates both his inventiveness and ease as a melodist and his naïveté in matters of style. Yet, in the 1920s Gounod's

music owed to this very naïveté its prestige among the Neoclassicist musicians of those days, who admired Gounod for having been a "pure" (non-philosophical and non-literary) composer. Even Stravinsky, in his *Poétique musicale*, praises not only *Faust* but also *Le Médecin malgré lui*, *La Colombe*, and *Philémon et Baucis*.

MAJOR WORKS

OPERAS: 13 operas, including: *Sapho* (1851); *La Nonne sanglante* (1854; libretto by Eugène Scribe and Casimir Delavigne based on M.G. Lewis' *Monk*); *Le Médecin malgré lui* (1858); *Faust* (1859); *Philémon et Baucis* (1860); *La Colombe* (1860); *La Reine de Saba* (1862); *Mireille* (1864); *Roméo et Juliette* (1867).

CHORAL MUSIC: Many oratorios, including *La Rédemption* (1882), *Mors et Vita* (1885); mass settings, including *Messe de Sainte-Cécile* (1855) and *Messe à la mémoire de Jeanne d'Arc* (1887); other settings of the liturgy; occasional pieces.

OTHER WORKS: Incidental music for Ernest Legouvé's *Les Deux Reines* (first produced 1872) and Jules Barbier's *Jeanne d'Arc* (1873); orchestral, chamber, and piano pieces, mostly now forgotten; more than 100 songs.

BIBLIOGRAPHY. CHARLES GOUNOD, *Autobiography* (1885), and *Mémoires d'un artiste* (Eng. trans. by A.E. CROCKER, 1896); C. BELLAIGUE, *Gounod* (1910), the standard biography; J.G. PROD'HOMME and A. DANDELOT, *Gounod: sa vie et ses oeuvres*, 2 vol. (1911), a study of his progress as a composer; P. LANDORMY, *Gounod* (1942), the most recent authoritative biography, and *Faust de Gounod: étude et analyse* (1944), a full study of the composer's most important work.

(F.Go.)

Goya, Francisco de

Francisco de Goya, trained in the foreign traditions fashionable in 18th-century Spain, became one of the most characteristically Spanish artists of all times and a major figure in the history of European art. His enormous and varied production of paintings, drawings, and engravings, relating to nearly every aspect of contemporary life, reflects the period of political and social upheavals in which he lived. He had no immediate followers, but his many original achievements profoundly impressed later 19th-century French artists—Eugène Delacroix was one of his great admirers—who were the leaders of new European movements, from Romanticism and Realism to Impressionism; and his works continued to be admired and studied by the Expressionists and Surrealists in the 20th century.

Early training and career. Francisco José de Goya y Lucientes was born on March 30, 1746, at Fuendetodos, near Saragossa. He began his studies in Saragossa under José Luzán y Martínez, a local artist trained in Naples, and was later a pupil, in Madrid, of the court painter Francisco Bayeu, whose sister he married in 1773. He went to Italy to continue his studies and was in Rome in 1771. In the same year he returned to Saragossa, where he obtained his first important commission for frescoes in the cathedral, which he executed at intervals during the next ten years. These and other early religious paintings made in Saragossa are in the Baroque–Rococo style then current in Spain and are influenced in particular by the great Venetian painter Giovanni Battista Tiepolo, who spent the last years of his life in Madrid (1762–70), where he was invited to paint ceilings in the royal palace.

By courtesy of the Biblioteca Nacional, Madrid



Francisco de Goya, self-portrait from "Los caprichos" series, etching, *c.* 1798.

Goya's career at court began in 1775, when he painted the first of a series of over 60 cartoons (preparatory paintings; mostly preserved in the Prado, Madrid), on which he was engaged until 1792, for the Royal Tapestry Factory of Santa Bárbara. These paintings of scenes of contemporary life, of aristocratic and popular pastimes, were begun under the direction of the German artist Anton Raphael Mengs, a great exponent of Neoclassicism who, after Tiepolo's death, had become undisputed art dictator at the Spanish court. In Goya's early cartoons the influence of Tiepolo's decorative style is modified by the teachings of Mengs, particularly his insistence on simplicity. The later cartoons reflect his growing independence of foreign traditions and the development of an individual style, which began to emerge through his study of the paintings of the 17th-century court painter Diego Velázquez in the royal collection, many of which he copied in etchings (*c.* 1778). Later in life he is said to have acknowledged three masters: Velázquez, Rembrandt, and, above all, nature. Rembrandt's etchings were doubtless a source of inspiration for his later drawings and engravings, while the paintings of Velázquez directed him to the study of nature and taught him the language of realism.

First cartoons

In 1780 Goya was elected a member of the Royal Academy of San Fernando, Madrid, his admission piece being a "Christ on the Cross," a conventional composition in the manner of Mengs but painted in the naturalistic style of Velázquez' "Christ on the Cross," which he doubtless knew. In 1785 he was appointed deputy director of painting at the Academy and in the following year painter to the king, Charles III. To this decade belong his earliest known portraits of court officials and members of the aristocracy, whom he represented in conventional 18th-century poses. The stiff elegance of the figures in full-length portraits of society ladies, such as "The Marquesa de Pontejos," and the fluent painting of their elaborate costumes also relates them to Velázquez' court portraits; and his representation of "Charles III as Huntsman" (private collection) is based directly on Velázquez' royal huntsmen.

Period under Charles IV. The death of Charles III in 1788, a few months before the outbreak of the French Revolution, brought to an end the period of comparative prosperity and enlightenment in which Goya reached maturity. The rule of reaction and political and social corruption that followed—under the weak and stupid Charles IV and his clever, unscrupulous queen, Maria Luisa—ended with the Napoleonic invasion of Spain. It was under the patronage of the new king, who raised him at once to the rank of court painter, that Goya became the most successful and fashionable artist in Spain; he was made director of the Academy in 1795 (but resigned two years later for reasons of health) and first court painter in 1799. Though he welcomed official honours and worldly success with undisguised enthusiasm, the record that he left of his patrons and of the society in which he lived is ruthlessly penetrating. After an illness in 1792 that left him permanently deaf, his art began to take on a new character, which gave free expression to the observations of his searching eye and critical mind and to his newly developed faculty of imagination. During his convalescence, he painted a set of cabinet pictures said to represent "national diversions," which he submitted to the Vice Protector of the Academy with a

covering letter (1794), saying, "I have succeeded in making observations for which there is normally no opportunity in commissioned works, which give no scope for fantasy and invention." The set was completed by "The Madhouse" in 1794, a scene that Goya had witnessed in Saragossa, painted in a broad, sketchy manner, with an effect of exaggerated realism that borders on caricature. For his more purposeful and serious satires, however, he now began to use the more intimate mediums of drawing and engraving. In "Los caprichos," a series of 80 etchings published in 1799, he attacked political, social, and religious abuses, adopting the popular imagery of caricature, which he enriched with highly original qualities of invention. Goya's masterly use of the recently developed technique of aquatint for tonal effects gives "Los caprichos" astonishing dramatic vitality and makes them a major achievement in the history of engraving. Despite the veiled language of designs and captions and Goya's announcement that his themes were from the "extravagances and follies common to all society," they were probably recognized as references to well-known persons and were withdrawn from sale after a few days. A few months later, however, Goya was made first court painter. Later he was apparently threatened by the Inquisition, and in 1803 he presented the plates of "Los caprichos" to the King in return for a pension for his son.

Etchings attacking abuses of society

While uncommissioned works gave full scope for "observations," "fantasy," and "invention," in his commissioned paintings Goya continued to use conventional formulas. His decoration of the church of San Antonio de la Florida, Madrid (1798), is still in the tradition of Tiepolo; but the bold, free execution and the expressive realism of the popular types used for religious and secular figures are unprecedented. In his numerous portraits of friends and officials a broader technique is combined with a new emphasis on characterization. The faces of his sitters reveal his lively discernment of personality, which is sometimes appreciative, particularly in his portraits of women, such as that of "Doña Isabel de Porcel," but which is often far from flattering, as in his royal portraits. In the group of "The Family of Charles IV," Goya, despite his position as court painter, has portrayed the ugliness and vulgarity of the principal figures so vividly as to produce the effect of caricature.

The Napoleonic invasion and period after the restoration. In 1808 Goya was at the height of his official career when Charles IV and his son Ferdinand were forced to abdicate in quick succession, Napoleon's armies entered Spain, and Napoleon's brother Joseph was placed on the throne. Goya, retained his position as court painter, but in the course of the war he portrayed Spanish as well as French generals, and in 1812 he painted a portrait of "The Duke of Wellington." It was, however, in a series of etchings, "Los desastres de la guerra" ("The Disasters of War"; first published 1863), for which he made drawings during the war, that he recorded his reactions to the invasion and to the horrors and disastrous consequences of the war. The violent and tragic events, which he doubtless witnessed, are represented not with documentary realism but in dramatic compositions—in line and aquatint—with brutal details that create a vivid effect of authenticity.

Depiction of the horrors of war

On the restoration of Ferdinand VII in 1814, after the expulsion of the invaders, Goya was pardoned for having served the French king and reinstated as first court painter. "The 2nd of May 1808: The Charge of the Mamelukes" and "The 3rd of May 1808: The Execution of the Defenders of Madrid" were painted to commemorate the popular insurrection in Madrid. Like "Los desastres," they are compositions of dramatic realism, and their monumental scale makes them even more moving. The Impressionistic style in which they are painted foreshadowed and influenced later 19th-century French artists, particularly Manet, who was also inspired by the composition of "The 3rd of May." In several portraits of Ferdinand VII, painted after his restoration, Goya evoked—more forcefully than any description—the personality of the cruel tyrant, whose oppressive rule drove most of his friends and eventually Goya himself into exile. He painted few other official portraits, but those of his friends and relations and his "Self-Portraits" (1815) are equally subjective. Some of his religious compositions of this period, the "Agony in the Garden" and "The Last Communion of St. Joseph of Calasanz" (1819), are more suggestive of sincere devotion than any of his earlier church paintings. The enigmatic "black paintings," with which he decorated the walls of his country house, the "Quinta del Sordo" (1820–23, now in the Prado) and "Los proverbios" or "Los disparates," a series of etchings made at about the same time (though not published until 1864), are, on the other hand, nightmare visions in expressionist language that seem to reflect cynicism, pessimism, and despair.

Later religious paintings

Last years. In 1824, when the failure of an attempt to establish a liberal government had led to renewed persecution, Goya applied for permission to go to France for reasons of health. After visiting Paris he settled in voluntary exile in Bordeaux, where he remained, apart from a brief trip to Madrid, until his death. There, in spite of old age and infirmity, he continued to record his impressions of the world around him in paintings, drawings, and the new technique of lithography, which he had begun to use in Spain. His last paintings include genre subjects and several portraits of friends in exile: "Don Juan Bautista de Muguiro," "Leandro Fernández de Moratín," and "Don José Pío de Molina," which show the final development of his style toward a synthesis of form and character in terms of light and shade, without outline or detail and with a minimum of colour. He died on April 16, 1828. Though there is little evidence for the legends of Goya's rebellious character and violent actions, he was undoubtedly a revolutionary artist.

MAJOR WORKS

PAINTINGS (RELIGIOUS): "The Adoration of the Name of God" (1772; Cathedral of Nuestra Señora del Pilar, Saragossa); "Christ on the Cross" (1780; Prado, Madrid); "The Queen of Martyrs" (1780–81; Cathedral of Nuestra Señora del Pilar, Saragossa); "St. Bernardino of Siena" (1782–83; San Francisco el Grande, Madrid); "The Annunciation" (1785; private collection, Spain); "The Taking of Christ" (1798; Cathedral, Toledo); "A Miracle of St. Anthony of Padua" and other scenes (1798; San Antonio de la Florida, Madrid); "SS. Justa and Rufina" (1817; Cathedral, Seville); "The Last Communion of St. Joseph of Calasanz" (1819; Escuelas Pías de San Antón, Madrid); "Agony in the Garden" (1819; Escuelas Pías de San Antón, Madrid). (PORTRAITS): "The Count of Floridablanca and Goya" (1783; Banco Urquijo, Madrid); "The Marquesa de Pontejos" (*c.* 1786; National Gallery of Art, Washington, D.C.); "Manuel Osorio de Zúñiga" (1788; Metropolitan Museum of Art, New York); "Family of the Duke of Osuna" (1788; Prado); "The Marquesa de la Solana" (*c.* 1794–95; Louvre, Paris); "The Duchess of Alba" (1797; Hispanic Society of America, New York); "Ferdinand Guillemardet" (1798; Louvre); "La Tirana" (1799; Academy of San Fernando, Madrid); "Queen Maria Luisa, on Horseback" (1799; Prado); "The Family of Charles IV" (1800; Prado); "The Naked Maja" and "The Clothed Maja" (*c.* 1800–05; Prado); "Doña Isabel de Porcel" (*c.* 1806; National Gallery, London); "General Manuel Romero" (*c.* 1810; private collection, Chicago); "The Duke of Wellington" (1812; National Gallery, London); "Mariano Goya" (*c.* 1812–14; private collection, Madrid); "Ferdinand VII in an Encampment" (*c.* 1814; Prado); "Self-Portrait" (1815; Academy of San Fernando, Madrid); "Portrait of Don Juan Antonio Cuervo" (1819; Cleveland Museum of Art); "Self-Portrait with Doctor Arrieta" (1820; Minneapolis Institute of Arts); "Leandro Fernández de Moratín" (1824; Museo de Bellas Artes, Bilboa); "The Milkmaid of Bordeaux" (1825–27; Prado); "Don Juan Bautista de Muguiro" (1827; Prado); "Don José Pío de Molina" (1827–28; Reinhart Collection, Winterthur). (HISTORY, ALLEGORY, AND GENRE): "Tapestry Cartoons" (1775–92; Prado); "The Madhouse" (1794; Virginia Meadows Museum and Elizabeth Meadows Sculpture Court, Dallas); "Allegory of the City of Madrid" (1810; Casa del Ayuntamiento, Madrid); "The Colossus" ("The Panic," *c.* 1810–12; Prado); "Time and the Old Women" (*c.* 1810–12; Musée des Beaux-Arts, Lille); "The Majas on the Balcony" (*c.* 1812; Metropolitan Museum of Art, New York); "Young Women with a Letter" (*c.* 1814–18; Musée des Beaux-Arts, Lille); "The 2nd of May 1808 in Madrid: The Charge of the Mamelukes" (1814; Prado); "The 3rd of May 1808: The Execution of the Defenders of Madrid" (1814; Prado); "The Forge" (*c.* 1819; Frick Collection,

New York); "The Black Paintings from the Quinta del Sordo" (1820–23; Prado).

DRAWINGS, ENGRAVINGS, AND LITHOGRAPHS: The largest and most important collection of drawings of all periods is in the Prado, Madrid.

ETCHINGS: "Los caprichos" (1797–98); "Los desastres de la guerra" (1810–14); "La tauromaquia" (1815–16); "Los disparates" or "Los proverbios" (*c.* 1820–24).

LITHOGRAPHS: "The Bulls of Bordeaux" (1824–25).

BIBLIOGRAPHY

General works: LAURENT MATHERON, *Goya* (1858), the first monograph (in French) on Goya, dedicated to Delacroix—contains interesting data on the artist's last years in Bordeaux; FRANCISCO ZAPATER Y GOMEZ, *Goya. Noticias biográficas* (1868; new ed., 1924), important biographical notes by the son of Goya's intimate friend, Martín Zapater, with extracts from his correspondence; CONDE DE LA VINAZA, *Goya: su tiempo, su vida, sus obras* (1887), a study of Goya and his times, incorporating many documents, with a catalogue raisonné; F.J. SANCHEZ CANTON, *Vida y obras de Goya* (1951; Eng. trans., *The Life and Works of Goya*, 1964), a well-documented general survey of Goya's life and principal works; F.D. KLINGENDER, *Goya in the Democratic Tradition*, 2nd ed. (1968), a stimulating, if partisan, study of the artist in relation to his political and social background, which is treated in considerable detail; PIERRE GASSIER and JULIET WILSON, *Vie et oeuvre de Francisco Goya* (1970; Eng. trans., *The Life and Complete Work of Francisco Goya*, ed. by FRANCOIS LACHENAL, 1971), of major importance as the most comprehensive monograph on Goya, providing an authoritative account of his life and career, with a catalog and reproductions of all his known paintings, drawings, and engravings and detailed bibliographical references.

Paintings: VALENTIN DE SAMBRICIO, *Tapices de Goya* (1946), the definitive work on Goya's tapestry cartoons, with all the relevant documents, and illustration of all the paintings; ENRIQUE LAFUENTE FERRARI, *The Frescos in San Antonio de la Florida in Madrid* (Eng. trans. 1955), a historical and critical study of the frescoes, illustrated in colour; F.J. SANCHEZ CANTON, *Goya and the Black Paintings*, with an appendix by XAVIER DE SALAS (1964), a well-documented account of Goya's career with a detailed study of the "black paintings" and a history of the "Quinta del Sordo," with large and detailed colour illustrations; JOSEP GUDIOL I RICART, *Goya: Biography, Analytical Study, and Catalogue of His Paintings* (Eng. trans., 1971), a detailed study of Goya's life and work accompanied by a critical catalog of his paintings (fully illustrated).

Drawings and engravings: F.J. SANCHEZ CANTON, *Los dibujos de Goya*, 2 vol. (1954), a catalog with reproductions of Goya's drawings in the Prado; and *Los Caprichos de Goya y sus dibujos preparatorios* (1949), a well-illustrated study of the etchings and their preparatory drawings; JOSE LOPEZ-REY, *A Cycle of Goya's Drawings: The Expression of Truth and Liberty* (1956), an interpretation of the drawings in the light of the political and social background during the liberal struggle in Spain following the defeat of Napoleon; and *Goya's Caprichos: Beauty, Reason and Caricature*, 2 vol. (1953), an interpretive study of the "Caprichos" with a catalog of the etchings and preparatory drawings, all illustrated; ENRIQUE LAFUENTE FERRARI, *Los desastres de la guerra de Goya y sus dibujos preparatorios* (1952); and JOSE CAMON AZNAR, *"Los Disparates" de Goya y sus dibujos preparatorios* (1951), two well-illustrated studies of the etchings and their preparatory drawings; TOMAS HARRIS, *Goya: Engravings and Lithographs*, 2 vol. (1964), the most important and most comprehensive work on the subject, including a study of Goya's techniques and methods of production and a complete illustrated catalog with critical and descriptive analysis of every work from the preparatory drawings and working proofs to all the published impressions.

(E.Ha.)

Gracchi, The

The brothers Tiberius Sempronius and Gaius Sempronius Gracchus, by exploiting the tribunate of the people and the legislative power of the People's Assembly of the Roman Republic, initiated what came to be called the Roman Revolution. Their techniques enabled later political leaders to secure power independently of the Senate, an aristocratic council of state dominated by elder statesmen, which had previously seldom been challenged in its control of public affairs.

Early years. The story of the Gracchi is overlain by partisan bias and invention, both friendly and hostile. The narratives of the historians Appian and Plutarch, composed two centuries after the events, tend to obscure both the motives of the actors and the details of the actions. Tiberius Sempronius Gracchus and his brother Gaius Sempronius, born about 163 and 153 BC, respectively, were the sons of a Roman aristocrat whose family had regularly held the highest offices of state for the past century. They could expect to secure election to the annual praetorships and consulships, the plums of public life, without any special exertions. But, maturing in an age of social and political crisis, they were the first to face its problems, and, by advocating radical changes in the face of entrenched conservatism, they undermined their own position and failed to establish a permanent substitute in populist support. They were born to a nexus of political connections with other leading families that would ensure the regular promotion of their careers—most notably with the Cornelii Scipiones, the most continuously successful of the great Roman houses—through their mother, Cornelia, daughter of the conqueror of Hannibal, and through their sister Sempronia, wife of Scipio Aemilianus, the destroyer of Carthage. They were equally associated with the great rivals of the Scipios, the Claudii Pulchri, through Tiberius' wife, Claudia, daughter of Appius Claudius Pulcher, the contemporary head of the house.

The brothers were educated in the new Greek enlightenment that had been adopted by the more liberal families after the Roman conquest of the Hellenistic kingdoms, and this gave form and clarity to their natural talent for public speaking. The new learning was based on literature, oratory, and philosophy. The Stoic teacher Blossius had special influence with Tiberius, but the central Stoic doctrine of duty merely enhanced his natural determination and obstinacy. Other Greek associates may have instructed him in the democratic theory that later coloured his speeches.

Military service

Career of Tiberius Gracchus. As a Roman aristocrat, Tiberius began a normal military career, serving as a junior officer with distinction under Scipio Aemilianus in the war with Carthage (147–146 BC), and in due course went as quaestor, or paymaster, with the consul Mancinus to the protracted colonial warfare in Spain (137 BC). There his personal integrity and family reputation enabled him to save a Roman army from total destruction at Numantia by an honourable compact with the Spanish tribesmen. But, at the insistence of Aemilianus, the agreement was disavowed by the Senate at Rome, and Mancinus, the defeated consul, though not his staff and his troops, was returned to his captors. This setback alienated Tiberius from the Scipionic faction in the Senate and drew him closer to his Claudian friends.

Distribution of public lands. His military experience had shown him the latent weakness of Rome. Its manpower was stretched to the limit to maintain its hegemony over the Mediterranean world, while its sources in Italy were beginning to contract. The primitive subsistence economy that in past centuries had nourished a large population of poor peasants was being eroded by new factors, notably the development of large estates owned by magnates enriched in the imperialist wars and devoted to cash crops worked by slaves and day labourers. The landowning peasantry, who alone were thought useful for military service, were declining in numbers, while the landless citizenry were increasing. Tiberius sought a solution of the manpower problem in a large-scale revival of the traditional Roman policy, abandoned only in the last 30 years, of settling landless men on the extensive public lands acquired by the Roman state during the former conquest of Italy. Much of this land had fallen irregularly but effectively into the hands of the Roman gentry, who regarded it wrongly as their private property. Tiberius, with the support of a small but powerful group of consular senators, primarily of the Claudian faction, who shared his concern and also looked for political advantage from sponsoring such a scheme, concocted a bill for the redistribution of the public lands to landless labourers in plots of viable size. The novelty lay only in the scale of the scheme, which was not limited to a defined area of land or number of persons, and in the

institution of a permanent executive of land commissioners. Opposition from vested interests was certain, but Tiberius hoped to pacify it by a generous provision allowing the great occupiers of public land to retain from 300 to 600 acres in private ownership, according to the size of their families.

Tribunate. To implement this measure Tiberius secured the legislative office of tribune, for 133 BC, which was not an essential part of a senatorial career. Tribunes at this period normally legislated in the People's Assembly on the advice of the Senate, but more than once in recent years tribunes had passed reformist measures without senatorial approval. Tiberius in 133 had the support of the sole consul in Rome—Publius Mucius Scaevola, who had helped to draft the agrarian bill—and of several other leading senators, mostly of the Claudian faction, whose authority could be expected to deflate opposition while hordes of peasants flocked to Rome to use their votes. When, after lengthy public debate, the bill was presented to the voters, the tribune Octavius used his right of veto to stop the proceedings in the interest of the great occupiers. When he refused to give way, Tiberius vainly sought belated approval from the Senate. That should have been the end of the matter, but Tiberius, obstinately convinced of the necessity of his bill, devised a novel method of bypassing the veto: a vote of the Assembly removed Octavius from office, contrary to all precedent. The bill itself was then passed. But the deposition of Octavius alienated many of Tiberius' supporters, who saw that it undermined the authority of the tribunate itself; they rejected the unfamiliar justification, devised by Tiberius, that tribunes who resisted the will of the people ceased to be tribunes.

Fresh complications arose from the lack of financial provision in the agrarian law for the equipment of the new landholders. Tiberius expected the Senate to make the traditional allocation of funds, but Scipio Nasica, an elderly senator from the Scipionic faction, succeeded in limiting these to a derisory sum. Tiberius countered by a second outrageous proposal, of which he failed to see the implication. The King of Pergamum, a city in Anatolia, on his death in 134 had bequeathed his fortune and his kingdom to the Roman state. Tiberius by a fresh bill claimed these monies in the name of the people and assigned them to the land commissioners, thus interfering with the Senate's traditional control of public finance and foreign affairs. The storm over Tiberius' methods continued to rage. He was threatened with prosecution after the end of his tribunate, when he would have no formal means of protecting his law and would be liable to prosecution before the Centuriate Assembly, in which the wealthier classes had a voting advantage. The charge would have been violation of the immunity of the tribune Octavius. Lacking the self-assurance to realize that the people were unlikely either to repeal the agrarian law or to pass sentence against its champion, Tiberius sought refuge in yet another impropriety. He proposed to stand for election to a second tribunate in 132, but re-election had not been practiced for 300 years and was widely believed to have been barred by an ambiguous statute. In the Senate the embittered opposition, again led by Nasica, tried to induce the consul Scaevola to stop the elections by force. Scaevola replied evasively that he would see that nothing illegal was done. Meanwhile, in the Assembly, Tiberius and the other tribunes were at loggerheads over the conduct of the election. An abortive vote had shown that the success of Tiberius was assured if only the election could be completed. He expected no violence and made no preparations against it. Enraged by the attitude of the Consul, Nasica and his associates stormed out of the Senate, equally unarmed. Seizing sticks and staves they precipitated a riot. It may well have begun as an attempt to disperse the electoral meeting, but it ended with the clubbing to death of Tiberius and the indiscriminate killing of some scores of citizens.

Threat of impeachment

The political fault lay with Tiberius. After presentation of the agrarian bill, he failed to act in prudent collaboration with his senatorial supporters, and he added to his troubles by dubious initiatives that were bound to offend the bulk of senatorial opinion. So Scaevola and the others abandoned him and effected a compromise. The Senate recommended that the land commission should continue, and, though in 132 it set up a political court that punished many of the lesser followers of Tiberius, it also encouraged Nasica, who barely escaped prosecution, to leave Italy.

Career of Gaius Gracchus. Gaius Gracchus was not long deterred from politics by his brother's disaster. Though barely 22 years old, he joined in the immediate outcry against Nasica, and he acted energetically as land commissioner in executing his brother's law. He became quaestor, a magistrate usually concerned with finance, in 126 at the normal age, after lengthy military service. When in 124 an intrigue against him at Rome delayed his already overdue recall from Sardinia, he asserted his independence by returning unsummoned and counterattacked his critics by underlining the honesty of his administration: "The purses that I took out to my province full of silver I have brought back empty, while others take out casks of wine and bring them back stuffed with loot."

First tribunate. The contentious tone forecast a vigorous politician, and his candidacy for the tribunate of 123 brought out great crowds of voters, though the opposition of family enemies prevented him from receiving the highest number of votes. He soon showed himself bent on exploiting his legislative power to the maximum. Cooler and more astute than his brother, he planned a complex program, of a sort never before proposed at Rome, and the means to carry it out.

Gaius realized that, by fostering sectional advantages, the influence of the wealthy upper class of landowners and businessmen outside the Senate—later known as Roman knights because of their liability to cavalry service—could be largely detached from its traditional support of the senatorial aristocracy and combined with the votes of the poorer citizens to carry radical reforms that no single group could manage by itself. Unlike his brother, he acted without significant senatorial support. Of the former family connections only Marcus Fulvius Flaccus, consul in 125, stood with Gaius. His other associates were dim and unknown figures. The scope of his proposals is the more remarkable, and the credit is his alone. He diagnosed the weaknesses of the Roman state at that time and proposed effective remedies. But, while the purpose of his particular measures is fairly clear, his general intentions remain obscure. Each of his proposals could be and was interpreted narrowly as a demagogic device for securing a personal predominance in the Assembly or as a direct attack on senatorial authority. Gaius certainly intended to establish an independent popular force in the Assembly, free from the traditional control of the great families. But his purpose was not democratic, for none of his measures intended the permanent replacement of the Senate and the annual officers of state by the popular Assembly. He used the Assembly not as an administrative body but as the source of legislative reform. This is seen clearly in his regulation for the annual assignment of the consular provinces, the most important policy-making moment in the Roman year. By securing passage of this law he ensured that the provinces would be allocated before the consuls were elected. This left the Senate as the policy-making body for war and foreign affairs but corrected an abusive procedure by requiring the Senate to make its decisions for each campaigning season before, instead of after, the election of the consuls who were to conduct affairs, thereby diminishing the opportunity for corruption. As an aristocrat he had no intention, unlike some of his successors, of subordinating the magistrates to the detailed control of the Assembly. He spoke in public as an enlightened, somewhat sardonic aristocrat who cultivated righteousness and despised his enemies, as in this fragment of a speech on foreign affairs:

Use of assembly for legislative reform

> Citizens, none of us men of rank comes before you for nothing, not even I myself. . . . But what I want from you is not money, but a good repute and public distinction.

The true understanding of Gaius is obscured by the un-

certainty of the chronological order of his measures in 123 and 122. But, despite minor confusions, it is clear that Gaius completed the whole of his program that touched the government of the Roman state before he turned to a different problem—the relationship between Rome and her Italian allies—early in his second tribunate and that his bill for the extension of the franchise to the independent peoples of Italy was his last legislative proposal. His preceding measures were criticized by the extreme conservatives as a general attempt to "destroy aristocracy and set up democracy, but they did not satisfy the radicals either." Other writers, while criticizing his methods, conceded the utility of many of his measures. Certainly his laws were "against the Senate," in that they sought to diminish its monopoly of power, and, if he had ended his effort in December 123, he could be dismissed as a statesman who sought personal supremacy by advocating salutary reforms. But the second tribunate, in which he sacrified his career to an unpopular cause, refutes so limited a view.

The measures of 123 were concerned with the abuse of power and with the extension of his brother's economic policy. He began with a demonstration against the enemies of Tiberius: the family vendetta was a regular part of Roman politics. A bill aimed at Octavius denied further office to magistrates deposed by the Assembly. Though Gaius did not press this proposal, it deterred his colleagues from using their vetoes against him. A law forbidding the establishment of political tribunals by the Senate without the sanction of the Assembly was intended to prevent a recurrence of the judicial murders committed by the political court set up to punish the supporters of Tiberius in 132, while providing for the indictment of Popillius Laenas, who had presided over them. This law was reinforced by a measure that penalized any attempt to exert improper influence over the management of tribunals, by the use of an ingenious formula covering "any magistrate or senator who combined or associated to secure the condemnation of a person in a public court."

A third law, concerned with judicial corruption, sought to provide independent juries for the "extortion court." This court had been created only 26 years earlier to curb the malpractices of Roman governors by enabling provincial subjects to sue for the restitution of monies taken improperly from them. Hitherto the jurors of this court had been senators, who had failed to protect the provincials against extortion through their own private interest in the fleecing of provinces. The judiciary law of Gaius excluded senators from the juries altogether and replaced them by Roman knights, who were expected to be more impartial. It is possible that Gaius originally proposed a mixed commission of senators and knights and withdrew that proposal when it became apparent that this scheme would not secure enough support to ensure its passage.

Gaius' judiciary law

Considerable portions survive of the text of what must be either the actual judiciary law of Gaius or a revised version modelled closely upon it. These show the same determination and ingenuity as his laws about special tribunals, to stop corruption and abuse in the working of the court. The exclusion of all magistrates and senators is minutely regulated, and no qualified juror may sit on a case if he and the accused person are members of the same club or confraternity. Lengthy clauses exactly regulated the distribution and collection of voting tablets and the counting of the vote.

This attention to detail is the hallmark of all the work done by Gaius about which there is any substantial information.

Two measures served partisan interests. The first established a beneficial system for the basic food supply of the now overgrown metropolis of Rome, where urban employment and prices were equally irregular. This "corn law" organized a supply of grain to be stored in new public granaries and to be sold to the urban population at a moderate price throughout the year. The second bill transferred the lucrative farming of taxes in the new province of Asia from local businessmen, who farmed the taxes on behalf of the Roman governor, to financial syndicates of Roman knights who dealt directly with the treasury at Rome, thus creating a monopoly for the Roman financiers. Both measures suggest a positive bid for the votes of persons domiciled at Rome. The rural population was wooed by two other measures: one transferred payments for military clothing from the conscript peasantry to the Roman treasury, and the second, modifying the law of Tiberius, proposed the establishment of self-governing communities of colonists instead of isolated individual holdings, not only in Italy but also in provinces overseas. This innovation led in later times to the widespread settlement of Roman colonies that latinized southern Europe.

Food and tax laws

Gaius may have toyed also with two basic constitutional amendments. The first would have enlarged the Senate from a council composed solely of former officials to include a larger number of Roman knights who had not held public office, thus giving the upper bourgeoisie a dominant voice in the formation of policy. The second proposal would have introduced a measure of equality into the complicated structure of the Centuriate Assembly. This body elected the consuls and praetors on the basis of voting groups weighted heavily in favour of the wealthier classes. Apparently Gaius ingeniously proposed to diminish the prepotent effects of the votes of the wealthy by simply changing the order in which they voted.

Second tribunate. Such were the laws proposed or carried by Gaius in 123 in the face of determined opposition. Though in late summer popular enthusiasm swept Gaius into a second tribunate, for which he had not originally intended to stand, his judiciary bill was subsequently passed by the vote of only 18 of the 35 voting groups of the Assembly. In so close a situation his successes are the more remarkable. But he had a yet more difficult project in mind for the next year. The greatest of Roman problems at this time concerned the management of the allies in Italy, who occupied two-thirds of the peninsula. They provided the larger part of the Roman armies that held the world in fee, yet these peoples were treated with increasing disdain and severity by the Roman aristocracy, though they were akin in race, language, and customs.

Gaius proposed a complex solution of the Italian question. The Latin-speaking allies, whose communal life was akin to that of Rome, were to be incorporated into the Roman state as full citizens and organized in locally self-governing municipalities, and the Italic peoples of non-Latin stocks were to have the intermediate status of the Latin allies; in private life the Latin allies shared some privileges of the Romans. This ingenious measure shows the disinterested yet committed character of Gaius as a statesman. Such an enlargement of the Roman state was, however, intensely unpopular with Romans of all classes. Gaius' persistence at once weakened his popular following and in the end wrecked his career. The opposition took heart. The tribune Marcus Livius Drusus undercut the proposal of Gaius with the suggestion that the Latin allies would be satisfied with a minor extension of privilege, while the consul Fannius played on the xenophobia of the masses. The prestige gained from the colonial schemes was undermined by a fictitious proposal of Drusus to found a full dozen such colonies in the name of the Senate. Gaius' position at Rome was not helped by his departure for two months to Africa to manage the foundation of a colony of 6,000 settlers at Carthage. Among the business classes, who had nothing more to gain from Gaius, his support was weakened by the alienation of the numerous corn merchants whose profits had been decreased. On his return Gaius tried by a series of demonstrations to restore his popular following. He moved his residence from an aristocratic quarter down to the plebeian streets around the Forum, insisted on the right of the common people to watch the public games without charge, and tried, though ineffectively, to prevent the execution of a consular decree forbidding Italians to remain in Rome during the vote on the enfranchisement bill. Altogether, opposed by senatorial opinion

Proposals for enfranchising allies

and shorn of his equestrian supporters, Gaius was a more isolated and a more demagogic figure than in 123. The enfranchisement bill was rejected, and Gaius failed to secure a third tribunate at the elections of 122.

In adversity Gaius showed the same stubborn determination as his brother to maintain a good cause at all costs. Like Tiberius he fell defending the agrarian colonization that was the basis of their position. In 121 a tribune proposed the dissolution of the great colony of Carthage. Helped by the remnant of his plebeian supporters, Gaius organized a counterdemonstration in a style that was allowed to no private citizen at Rome, where even a senator could not address a public meeting unless invited by a magistrate. In the fracas one of Gaius' party was killed, and the Gracchans retired uneasily to the Aventine Hill, traditional asylum of the Roman plebeians in an earlier age. Some of them bore arms, but the hostile sources admit that Gaius carried no more than a dagger. The Senate seized the opportunity to pass a novel decree, which urged the consuls to protect the state from any harm. Gaius, appalled, sought a parley. But the consul Lucius Opimius, refusing any negotiations, organized a heavily armed force composed largely of Roman knights and assaulted the Aventine. Massacre followed, as did the suicide of Gaius. But most of his legislation survived, and his unfinished projects were remembered, to become the basis of politics in the next generation. His rejected unification of Italy was finally conceded in 89 BC, after a destructive and unnecessary civil war that came close to destroying the foundations of Roman power. Hardly any substantial reform was proposed in the last century of the republic that did not owe its conception to the political intelligence of Gaius Gracchus. But he had also perfected a method of political organization in the interest of serious policy that proved disastrous in the hands of lesser men. Only too justly did he remark that he had cast into the Forum daggers with which the citizens would destroy themselves.

BIBLIOGRAPHY

Sources: APPIAN, *Civil Wars*, bk. 1, ch. 22–27; and PLUTARCH, *Parallel Lives, Tiberius and Gaius Gracchus*, both translated in the "Loeb Classical Library." For other material, see A.H.J. GREENIDGE and A.M. CLAY, *Sources for Roman History, 133–70 B.C.*, 2nd ed. (1960).

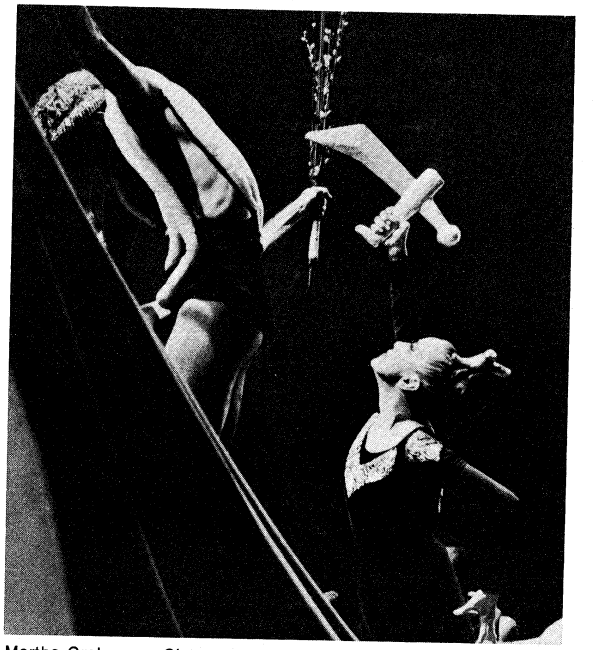
Modern accounts: H.M. LAST, *Cambridge Ancient History*, vol. 9 (1932); J. CARCOPINO, *Autour de Gracques* (1928), a balanced account, stressing the chronological problems, that evaluates the bias of the source but is too speculative; T. FRANK (ed.), *An Economic Survey of Ancient Rome*, vol. 1 (1933), documents the agricultural background; D.C. EARL, *Tiberius Gracchus, a Study in Politics* (1963); E. BADIAN, *Foreign Clientelae, 264–70 B.C.* (1958), for enfranchisement proposals; A.E. ASTIN, *Scipio Aemilianus* (1967), usefully summarizes recent theories.

(A.Sh.-W.)

Graham, Martha

One of the most influential dancers, choreographers, teachers, and innovators of modern dance, Martha Graham created a dance technique rooted in the muscular and neuromuscular responses of the body to both inner and outer stimuli. It is a technique requiring both unrelenting discipline and prodigious virtuosity. It is the most highly developed body-training method in the entire field of modern dance. Explaining her art, she said that it was the function of her theatre of dance "to reveal the inner man" or "to give substance to things felt."

Martha Graham was born on May 11, 1893, in Pittsburgh, Pennsylvania, one of three daughters of a physician who was particularly interested in the bodily expression of human behaviour. Her ancestry, on one side, went directly back to Myles Standish and, on the other, to what she herself has called "Black Irish," a combination that resulted in a strict Protestant upbringing but an early exposure to the ritual beauties of the Roman Catholic Church. After some time in the South, her family settled in Santa Barbara, California, where she discovered the rhythm of the sea and made an acquaintance with Oriental art, influences that were to be evident in her choreography throughout her career.



Martha Graham as Clytemnestra. (Choreography and costumes, Martha Graham; music, Halim El-Dabh; decor, Isamu Noguchi.)
Martha Swope

Denishawn

Her professional career had its beginning at Denishawn, the schools and dance company founded by Ruth St. Denis and Ted Shawn, where as a teenager she was introduced to a repertory and curriculum that, for the first time in America, explored the world's dances—folk, classical, experimental, Oriental, Occidental, and American Indian. She was entranced by the religious mysticism of St. Denis, but Shawn was her major teacher; he discovered sources of dramatic power within her and then channelled them into an Aztec ballet, *Xochitl*. A tremendous success in both vaudeville and in concert performance (in the latter Shawn himself was her partner), *Xochitl* made her a Denishawn star.

She remained with Denishawn from 1916 to 1923, and although she ultimately rebelled violently against its eclecticism, she later mirrored in her own works something of the Orientalism that pervaded the school, and she assumed something of Ruth St. Denis' attitude of a priestess when she herself became a senior dancer.

Martha Graham eventually left Denishawn to become a featured dancer in the Greenwich Village Follies revue, where she remained for two years. Independence was not all she wanted, however; she was also searching for a new way of permitting the body to express itself in dance. In 1924 she went to the Eastman School of Music in Rochester, New York, to teach and to experiment.

Martha Graham made her New York debut as an independent artist in 1926. Though some of the fruits of her experiments were discernible from the first, a good many of her dances, such as *Three Gopi Maidens* and *Danse Languide*, echoed her Denishawn past. The critics found her to be very graceful and lyrical. All of that changed with her 1927 concert, and for the next decade and more, her dances were to be referred to as ugly, stark, and obscure. The exotic costumes and rich staging of Denishawn were in the past. Among the dances of her 1927 program was *Revolt*, probably the first dance of protest and social comment staged in the United States, which was set to the avant-garde music of Arthur Honegger. The audience was not impressed; dancers and theatregoers, famous and unknown, made fun of her. Fanny Brice, the popular comedienne, satirized her in a revue. Graham herself later referred to this decade as "my period of long woolens," a reference to the plain jersey dress that she wore almost as a uniform.

A strong and continuing influence in her life was Louis Horst, musical director at Denishawn, who had left the school two years after Graham. He became her musical

director, and often composer, during her first two decades of independence, and remained close to her until his death in 1964. Among his most noted scores for her were those for the now historic *Frontier*, a solo dance, and *Primitive Mysteries*, written for Graham and a company of female dancers.

Frontier (1935) initiated the use of decor in her repertoire and marked the beginning of a long and distinguished collaboration with the noted Japanese-American sculptor Isamu Noguchi, under whose influence she developed one of her most singular stage innovations, the use of sculpture, or three-dimensional set pieces, instead of flats and drops. This production illustrated her concept of dance.

Concept of dance

For Martha Graham the dance, though it is in part a narrative medium, can also, like the spoken drama, explore the spiritual and emotional essence of man. Thus, when in *Frontier* a leg sweeps high, it is not a chorus girl's high kick but an expression of ecstasy, of a willingness to adventure. *Frontier* is not of the American West alone, but rather a distillation of the word's meaning itself; a geographical frontier, the youth facing the post-school frontier of life, the virgin passing into potential motherhood, the citizen choosing between safe conformity and rebellion, the man between comfort and daring. It is a revolution of "the inner man."

In *Night Journey* (1947), a work about the Greek legendary figure Jocasta, the whole dance-drama takes place in the instant when Jocasta learns that she has mated with Oedipus, her own son, and has borne him children. The enormity of the sin is such that there is no cry of remorse from the lips; only a leg sweeping head high and the body falling forward to the floor as the legs slide into a split. The action, was to Graham, "a cry from the loins." Such a monumental exclamation was necessary to distinguish Jocasta's crime from an ordinary one. In *Letter to the World* (1940), a work about the poet Emily Dickinson, there are two Emilys: one who spoke the lines of Dickinson poems, and was restrained, ladylike, New England in deportment; and another whom Graham danced as the inner Emily, a wild, tempestuous, mercurial creature. It was the inner, not the outer, being who wrote the poems, and Graham knew it—this dance was one more example of her concept of the function of a dance, "to reveal the inner man."

For more than ten years, her company consisted solely of girls, but her themes were beginning to call for men and women dancers. She engaged Erick Hawkins, a ballet dancer, to join her company, and he appeared with her in a major work, *American Document* (1938). Though she and Hawkins ultimately were married, the marriage did not last.

In a career spanning more than half a century, Martha Graham created close to 150 works, ranging from solos to large-scale creations of full-program length such as *Clytemnestra* (1958). For her themes she has almost always turned to man, his conflicts and his emotions. The settings and the eras vary, but her great gallery of danced portraits never fails to explore the inner being. She created some dances from American frontier life, the most famous of which is *Appalachian Spring* (1944), with its score by Aaron Copland. Another source was Greek legend, the dances rooted in classical Greek dramas, stories, and myths. *Cave of the Heart* (1946), based on the Medea figure, with music by Samuel Barber, was not a dance version of Medea herself but rather an exposure of the Medea latent in every woman who, out of consuming jealousy, not only destroys those she loves but herself as well. *Errand Into the Maze* (1947), an investigation of hidden fears presented through the symbols of the Minotaur and the labyrinth; *Alcestis* (1960); *Phaedra* (1962); *Circe* (1963); and others followed. Biblical themes and religious figures have also inspired her: *Seraphic Dialogue* (1955; Joan of Arc), *Embattled Garden* (the Garden of Eden), *Legend of Judith* (1962), and such fanciful abstractions as *Diversion of Angels* (1948) or *Acrobats of God* (1960).

Graham, whose stage presence is one of commanding stature and great dynamic force, is, in fact, a tiny woman. An indefatigable worker, quick-tempered, terrifying in her rages, sunny, kind, and humorous when life is smooth, she is generous with her company and her students and friends.

Assessment

She has maintained a position as the foremost figure in U.S. modern dance, for although movements that were once avant-garde have become familiar, her art and her craftsmanship can still startle and surprise. She has instructed, or guided, generations of modern dance teachers both in America and abroad, and has toured Europe, the Orient, and the Middle East with her company. She has strongly influenced succeeding generations of modern dancers, ballet choreographers, stagers of musicals and operas, creators of dance-dramas, and directors of motion pictures and television. In 1970 she announced her retirement as a dancer but continued to create dances. From the "long woolens" of the 1920s, Martha Graham has moved to some of the most opulent productions to be found in modern dance, with an accent on sculptured pieces, brilliant costumes and properties, and she continues to be the recipient of many awards and honours.

BIBLIOGRAPHY. LEROY LEATHERMAN, *Martha Graham: Portrait of the Lady as an Artist* (1966), a personal and artistic biography; MERLE ARMITAGE (ed.), *Martha Graham* (1937, reprinted 1966), articles, reviews, and pictures of Martha Graham's Dance Theatre; MARGARET LLOYD, *The Borzoi Book of Modern Dance* (1949, reprinted 1969), a study of modern dance with synopses of major works; BARBARA MORGAN, *Martha Graham* (1941), a book of dance photographs; *A Dancer's World* (movie), produced by NATHAN KROLL (1957), a film of Graham technique and excerpts from her repertoire.

(W.T.)

Grammar

Historically, the term grammar has been used in several senses, not all of which need be taken up in detail here.

1. It is the name of a branch of learning that investigates the formal, or abstract, features of a language and the rules that govern their combination, reference, and interpretation. By "features of a language" are understood such matters as sounds, morphemes (the smallest meaningful units), words, and sentences. Morphology, in the narrowest sense, refers to the construction of words from morphemes; syntax has to do with the construction of sentences, often to the exclusion of morphology. A central concern of grammar, then, is syntax, though often also with its phonetic manifestations and representation in writing. A more usual and more manageable meaning of the word grammar, attested in English at least since 1846, is the study of morphology and syntax. Another conception of grammar, particularly current in technical work of the 1930s and 1940s, embraced phonology (the study of the sound patterns that occur within languages) and morphosyntax (the body of rules, taken together, that not only build sentences but also incorporate elements in words and thus predict the formation—morphology—of words along with sentences). This view excluded semantics and phonetics, which such theorists took to involve substantial, nonlinguistic criteria different from the abstract attributes that they supposed to characterize grammar and linguistics proper.

There is a further understanding of the notion of grammar that limits it to those problems that bear on the classification of the parts of speech.

2. "Grammar" also refers to the actual features dealt with abstractly in grammar as it is defined in 1 above and to the rules that unite and relate them; it is in this sense that the term English grammar is used. More formally, a grammar has been called a system of rules that expresses the correspondences between the sounds and meanings of a language. A grammar, then, is a finite theory that specifies the structures of an infinite class of sentence and discourse shapes. A limited version of this sense that has long been current is illustrated by the phrase "English has very little grammar." By this is meant that English has few overt inflections.

3. By "grammar" are understood such features and rules as are sketched in 2 above, or the study of such features and rules as is outlined in 1, when they char-

acterize human language as a whole. When one speaks of "universal grammar," he means grammar in this sense. Such a view could also seek by extension to characterize other creatures' modes of communication.

"Universal grammar"

4. A grammar may also be a book or treatise presenting or formulating those features outlined in 2 above. A special but frequent use of this sense has reference to a handbook for foreign-language learning. Such a grammar for a second language is committed to presenting the grammatical characteristics of one language from the point of view of another. This approach rests upon the accumulation of a body of statements that has been called contrastive grammar, the systematic study of all those features and rules whereby one grammar (of language *A*) differs from another (of language *B*). It is obvious that from such a study two separate formulations may be derived: the divergences of the grammar of *A* with respect to *B* as a base; or the divergences of the grammar of *B* with respect to *A*. Thus, the problem of specifying English for French speakers is not the same as that of specifying French for English speakers. Analogously, the devising of effective teaching materials offers complex and asymmetrical problems in each case.

In contrast to a historical grammar, which sets forth the features and rules of a language and their changes (attested in documents) over a period of time, a comparative grammar presents in manageable form those aspects of grammar (and the supporting evidence) that are recovered by reconstruction for the unattested parent language.

A treatise designed to provide a reasonably exhaustive, technical, formally stated, and parsimonious account of the grammar of a language is called a reference grammar. A handbook that aims to convey in an effective fashion a manageable coverage of the grammar of a language (whether for purposes of information or to impart an active command of the language in writing or in speaking) is called a pedagogical grammar.

5. "Grammar" may also be conceived of as a set of rules or criteria for evaluating sentences or constructions; such a set of criteria might be rather informal. This sense is in a way a subtype of 2 above, although it differs immediately in introducing an element of value judgment. The word grammar is understood in this sense in the locution "He knows his grammar."

6. "Grammar," in colloquial and everyday usage, frequently means the knowledge or employment of preferred or prescribed modes of expression according to some social and regional norm. It also means simply the knowledge or employment of grammatical forms. Both of these allied senses presuppose the establishment of the exact nature of the first three senses indicated above.

Origins of the term grammar

The word "grammar" is descended from the Middle English *gram(m)er(e)*, which in turn was borrowed from the Old French *gramaire*. Until the 16th century this term was used only of the grammar of Latin; in a now obsolete sense "grammar" meant simply "the Latin language." This family of words earlier replaced the Latin term *grammatica*, which in turn was a borrowing by learned Latin scholarship from the Greek term *grammatikē*.

THE STUDY OF GRAMMAR IN THE PAST

Non-Western traditions. Grammatical speculation and investigation, insofar as is known, has gone on in only a small number of societies. To the extent that Mesopotamian, Chinese, and Arabic learning dealt with grammar, their treatments were so enmeshed in the particularities of those languages and so little known to the European world until recently that they have had virtually no impact on Western grammatical tradition. Chinese linguistic and philological scholarship stretches back for more than two millennia, but the interest of those scholars was concentrated largely on phonetics, writing, and lexicography; their consideration of grammatical problems was bound up closely with the study of logic.

Certainly the most interesting non-Western grammatical tradition—and the most original and independent—is that of India, which dates back at least two and one-half millennia and which culminates with the grammar of Pāṇini, of the 5th century BC. There are three major ways in which the Sanskrit tradition has had an impact on modern linguistic scholarship. As soon as Sanskrit became known to the Western learned world the unravelling of comparative Indo-European grammar ensued and the foundations were laid for the whole 19th-century edifice of comparative philology and historical linguistics. But, for this, Sanskrit was simply a part of the data; Indian grammatical learning played almost no direct part. Nineteenth-century workers, however, recognized that the native tradition of phonetics in ancient India was vastly superior to Western knowledge; and this had important consequences for the growth of the science of phonetics in the West. Thirdly, there is in the rules or definitions (sutras) of Pāṇini a remarkably subtle and penetrating account of Sanskrit grammar. The construction of sentences, compound nouns, and the like is explained through ordered rules operating on underlying structures in a manner strikingly similar in part to modes of contemporary theory. As might be imagined, this perceptive Indian grammatical work has held great fascination for 20th-century theoretical linguists. A study of Indian logic in relation to Pāṇinian grammar alongside Aristotelian and Western logic in relation to Greek grammar and its successors could bring illuminating insights.

Sanskrit grammar

Whereas in ancient Chinese learning a separate field of study that might be called grammar scarcely took root, in ancient India a sophisticated version of this discipline developed early alongside the other sciences. Even though the study of Sanskrit grammar may originally have had the practical aim of keeping the sacred Vedic texts and their commentaries pure and intact, the study of grammar in India in the 1st millennium BC had already become an intellectual end in itself.

Classical antiquity. The emergence of grammatical learning in Greece is less clearly known than is sometimes implied, and the subject is more complex than is often supposed; here only the main strands can be sampled. The term *hē grammatikē technē* ("the art of letters") had two senses. It meant the study of the values of the letters and of accentuation and prosody and, in this sense, was an abstract intellectual discipline; and it also meant the skill of literacy and thus embraced applied pedagogy. This side of what was to become "grammatical" learning was distinctly applied, particular, and less exalted by comparison with other pursuits. Most of the developments associated with theoretical grammar grew out of philosophy and criticism; and in these developments a repeated duality of themes crosses and intertwines.

Much of Greek philosophy was occupied with the distinction between that which exists "by nature" and that which exists "by convention." So in language it was natural to account for words and forms as ordained by nature (by onomatopoeia—*i.e.*, by imitation of natural sounds) or as arrived at arbitrarily by a social convention. This dispute regarding the origin of language and meanings paved the way for the development of divergences between the views of the "analogists," who looked on language as possessing an essential regularity as a result of the symmetries that convention can provide, and the views of the "anomalists," who pointed to language's lack of regularity as one facet of the inescapable irregularities of nature. The situation was more complex, however, than this statement would suggest. For example, it seems that the anomalists among the Stoics credited the irrational quality of language precisely to the claim that language did not exactly mirror nature. In any event, the anomalist tradition in the hands of the Stoics brought grammar the benefit of their work in logic and rhetoric. This led to the distinction that, in modern theory, is made with the terms *signifiant* ("what signifies") and *signifié* ("what is signified") or, somewhat differently and more elaborately, with "expression" and "content"; and it laid the groundwork of modern theories of inflection, though by no means with the exhaustiveness and fine-grained analysis reached by the Sanskrit grammarians.

"Analogists" and "anomalists"

The Alexandrians, who were analogists working largely on literary criticism and text philology, completed the development of the classical Greek grammatical tradition. Dionysius Thrax, in the 2nd century BC, produced the

first systematic grammar of Western tradition; it dealt only with word morphology. The study of sentence syntax was to wait for Apollonius Dyscolus, of the 2nd century AD. Dionysius called grammar "the acquaintance with [or observation of] what is uttered by poets and writers," using a word meaning a less general form of knowledge than what might be called "science." His typically Alexandrian literary goal is suggested by the headings in his work: pronunciation, poetic figurative language, difficult words, true and inner meanings of words, exposition of form-classes, literary criticism. Dionysius defined a sentence as a unit of sense or thought, but it is difficult to be sure of his precise meaning.

The Romans, who largely took over, with mild adaptations to their highly similar language, the total work of the Greeks, are important not as originators but as transmitters. Aelius Donatus, of the 4th century AD, and Priscian, an African of the 6th century, and their colleagues were slightly more systematic than their Greek models but were essentially retrospective rather than original. Up to this point a field that was at times called *ars grammatica* was a congeries of investigations, both theoretical and practical, drawn from the work and interests of literacy, scribeship, logic, epistemology, rhetoric, textual philosophy, poetics, and literary criticism. Yet modern specialists in the field still share their concerns and interests. The anomalists, who concentrated on surface irregularity and who looked then for regularities deeper down (as the Stoics sought them in logic) bear a resemblance to contemporary scholars of the transformationalist school. And the philological analogists with their regularizing surface segmentation show striking kinship of spirit with the modern school of structural (or taxonomic or glossematic) grammatical theorists.

The Middle Ages. It is possible that developments in grammar during the Middle Ages constitute one of the most misunderstood areas of the field of linguistics. It is difficult to relate this period coherently to other periods and to modern concerns because surprisingly little is accessible and certain, let alone analyzed with sophistication. In the early 1970s the majority of the known grammatical treatises had not yet been made available in full to modern scholarship, so that not even their true extent could be classified with confidence. These works must be analyzed and studied in the light of medieval learning, especially the learning of the schools of philosophy then current, in order to understand their true value and place.

The field of linguistics has almost completely neglected the achievements of this period. Students of grammar have tended to see as high points in their field the achievements of the Greeks, the Renaissance growth and "rediscovery" of learning (which led directly to modern school traditions), the contemporary flowering of theoretical study (men usually find their own age important and fascinating), and, in recent decades, the astonishing monument of Pāṇini. Many linguists have found uncongenial the combination of medieval Latin learning and premodern philosophy. Yet medieval scholars might reasonably be expected to have bequeathed to modern scholarship the fruits of more than ordinarily refined perceptions of a certain order. These scholars used, wrote in, and studied Latin, a language that, though not their native tongue, was one in which they were very much at home; such scholars in groups must often have represented a highly varied linguistic background.

Use of Latin by medieval scholars

Some of the medieval treatises continue the tradition of grammars of late antiquity; so there are versions based on Donatus and Priscian, often with less incorporation of the classical poets and writers. Another genre of writing involves simultaneous consideration of grammatical distinctions and scholastic logic; modern linguists are probably inadequately trained to deal with these writings.

Certainly the most obviously interesting theorizing to be found in this period is contained in the "speculative grammar" of the *modistae*, who were so called because the titles of their works were often phrased *De modis significandi tractatus* ("Treatise Concerning the Modes of Signifying"). For the development of the Western grammatical tradition, work of this genre was the second great milestone after the crystallization of Greek thought with the Stoics and Alexandrians. The scholastic philosophers were occupied with relating words and things—*i.e.*, the structure of sentences with the nature of the real world—hence their preoccupation with signification. The aim of the grammarians was to explore how a word (an element of language) matched things apprehended by the mind and how it signified reality. Since a word cannot signify the nature of reality directly, it must stand for the thing signified in one of its modes or properties; it is this discrimination of modes that the study of categories and parts of speech is all about. Thus the study of sentences should lead one to the nature of reality by way of the modes of signifying.

The *modistae*

The *modistae* did not innovate in discriminating categories and parts of speech; they accepted those that had come down from the Greeks through Donatus and Priscian. The great contribution of these grammarians, who flourished between the mid-13th and mid-14th century, was their insistence on a grammar to explicate the distinctions found by their forerunners in the languages known to them. Whether they made the best choice in selecting logic, metaphysics, and epistemology (as they knew them) as the fields to be included with grammar as a basis for the grand account of universal knowledge is less important than the breadth of their conception of the place of grammar. Before the *modistae*, grammar had not been viewed as a separate discipline but had been considered in conjunction with other studies or skills (such as criticism, preservation of valued texts, foreign-language learning). The Greek view of grammar was rather narrow and fragmented; the Roman view was largely technical. The speculative medieval grammarians (who dealt with language as a *speculum*, "mirror" of reality) inquired into the fundamentals underlying language and grammar. They wondered whether grammarians or philosophers discovered grammar, whether grammar was the same for all languages, what the fundamental topic of grammar was, and what the basic and irreducible grammatical primes are. Signification was reached by imposition of words on things; *i.e.*, the sign was arbitrary. Those questions sound remarkably like current issues of linguistics, which serves to illustrate how slow and repetitious progress in the field is. While the *modistae* accepted, by modern standards, a restrictive set of categories, the acumen and sweep they brought to their task resulted in numerous subtle and fresh syntactic observations. A thorough study of the medieval period would greatly enrich the discussion of current questions.

The Renaissance. It is customary to think of the Renaissance as a time of great flowering. There is no doubt that linguistic and philological developments of this period are interesting and significant. Two new sets of data that modern linguists tend to take for granted became available to grammarians during this period: (1) the newly recognized vernacular languages of Europe, for the protection and cultivation of which there subsequently arose national academies and learned institutions that live down to the present day; and (2) the exotic languages of Africa, the Orient, the New World, and, later, of Siberia, Inner Asia, Papua, Oceania, the Arctic, and Australia, which the voyages of discovery opened up. Earlier, the only non-Indo-European grammar at all widely accessible was that of the Hebrews (and to some extent Arabic); and Semitic in fact shares many categories with Indo-European in its grammar. Indeed, for many of the exotic languages scholarship barely passed beyond the most rudimentary initial collection of word lists; grammatical analysis was scarcely approached.

New sources of data

In the field of grammar, the Renaissance did not produce notable innovation or advance. Generally speaking, there was a strong rejection of speculative grammar and a relatively uncritical resumption of late Roman views (as stated by Priscian). This was somewhat understandable in the case of Latin or Greek grammars, since here the task was less evidently that of intellectual inquiry and more that of the schools, with the practical aim of gaining access to the newly discovered ancients. But, aside from the fact that, beginning in the 15th century, serious grammars

of European vernaculars were actually written, it is only in particular cases and for specific details (*e.g.*, a mild alteration in the number of parts of speech or cases of nouns) that real departures from Roman grammar can be noted. Likewise, until the end of the 19th century, grammars of the exotic languages, written largely by missionaries and traders, were cast almost entirely in the Roman model, to which the Renaissance had added a limited medieval syntactic ingredient.

From time to time a degree of boldness may be seen in France: Petrus Ramus, a 16th-century logician, worked within a taxonomic framework of the surface shapes of words and inflections, such work entailing some of the attendant trivialities that modern linguistics has experienced (*e.g.*, by dividing up Latin nouns on the basis of equivalence of syllable count among their case forms). In the 17th century, members of the Port-Royal community (a group of hermits who lived in the deserted abbey of Port-Royal in France) produced a grammar that has exerted noteworthy continuing influence, even in contemporary theoretical discussion. Drawing their basic view from scholastic logic as modified by rationalism, these people aimed to produce a philosophical grammar that would capture what was common to the grammars of languages—a general grammar, but not aprioristically universalist. This grammar has attracted recent attention because it employs certain syntactic formulations that resemble in detail contemporary transformational rules, which formulate the relationship between the various elements of a sentence.

The Port-Royal school

Roughly from the 15th century to World War II, however, the version of grammar available to the Western public (together with its colonial expansion) remained basically that of Priscian with only occasional and subsidiary modifications, and the knowledge of new languages brought only minor adjustments to the serious study of grammar. As education has become more broadly disseminated throughout society by the schools, attention has shifted from theoretical or technical grammar as an intellectual preoccupation to prescriptive grammar, which started with Renaissance vernacular nationalism. Grammar increasingly parted company with its older fellow disciplines within philosophy as they moved over to the domain known as natural science, and technical academic grammatical study has increasingly become involved with issues represented by empiricism versus rationalism and their successor manifestations on the academic scene.

Nearly down to the present day, the grammar of the schools has had only tangential connections with the studies pursued by professional linguists; for most people prescriptive grammar has become synonymous with "grammar," and the prevailing view held by educated people regards grammar as an item of folk knowledge open to speculation by all, and in nowise a formal science requiring adequate preparation such as is assumed for chemistry.

19th and early 20th centuries. The past two centuries have witnessed two fundamental changes in the characteristic activities of linguists that must be noted to put grammar in proper perspective. The great triumph of the 19th century was the unfolding of the study of language from a comparative and historical point of view. This had tremendous consequences for linguistics as a whole. It built firmly on the refined text philology developed during the Renaissance and had enormous impact on the field of phonology. It even filtered into some of the popular teaching in the schools, and for this reason linguists are often viewed by the public as essentially etymologists and tracers of the history of old words. The successor field of dialectology has also added to this popular picture. Nineteenth-century comparative philology, however, did not lead to any fundamental questioning of the basic notions of grammar; if anything, scholars increasingly despaired of successfully recovering from the past such abstract structures and their relations in any degree of interesting complexity.

Comparative and historical treatment of grammar

In the late 19th century, however, a series of intellectual earthquakes began to rock the field of linguistics from various directions. Ethnography—the descriptive study of the customs of particular tribes and peoples—contributed new data of a new order of exactitude from exotic aboriginal languages. Certain linguists with typological, rather than comparativist, leanings took a fresh look at what they could then learn of the world's grammar; and formalist literary criticism inspired a fresh analytic attitude from another, and unexpected, angle. These varied points of view converged to provide the direction that grammatical study was to take.

Ethnographers, studying aboriginal cultures in the field, insisted that the student could not truly understand a culture and its values without approaching it through the language; that the language and its grammar must be analyzed in its own terms and not through the lens of Latin or European grammar as had been customary since the voyages of discovery; and that if cultures could differ enormously from what Westerners are accustomed to, as the cultures visibly did, then grammars might, similarly, show infinite variety. Alexandrian and Renaissance scholars drew from literature their textual authority for the data they analyzed, and the product of grammatical analysis was integrated into literary study. It was now the turn of literary study to provide the analytic model for grammatical investigation. The insistence of the formalists on internal analysis of the literary work itself suggested to those of their number working in Moscow at the turn of the century (and later in Prague) who were interested in grammar that a similar internal analysis of oppositions (contrasts) within the language should apply.

Structural grammar

The upshot of all this was the view of grammar prevalent among many theoretical linguists from about 1930 to 1960, a view highly compatible with the emphasis on cultural relativism then current in anthropology. In brief, these linguists held the view that no prediction could be made in advance as to what a given grammar might contain or how it might be structured, that basic categories could vary from language to language indefinitely, and that this relative and arbitrary grammar could in turn serve as a screen through which a society might view the world differently from all other societies. Although some who held these views also had strong positivist or behaviourist leanings that additionally coloured their working methods, such further views were not an essential part of this grammatical outlook. This view of grammar led to a close association between the fields of linguistics and anthropology; at the same time traditional preoccupations of philosophy were not congenial to this view of grammar.

This, in grossly over-simplified terms, is the approach that has been called structural grammar; often in the United States it has been called descriptive grammar. To the extent that such study rigorously excluded evidence from elsewhere on the time axis or from among geographic or social variants, it was called synchronic grammar. Linguists working in a transformational framework often refer to it as taxonomic grammar. The central interest of structural grammar has been to show how the grammar of a given language is unique.

CONTEMPORARY GRAMMAR

During the 1960s, with the striking shift in the field of linguistics to generative-transformational grammar (also called simply transformational grammar or generative grammar), a different emphasis asserted itself. This type of grammar tries to generate, or produce, all of the possible grammatical utterances of a language and only those that are grammatical. In order to accomplish this, a series of rules is posited (some of which are called transformations) that express the relationships among the elements of a sentence. For example, a generative-transformational grammar tries to formulate the relationship between such sentences as "The man is here." "Is the man here?" "Where is the man?" and "Who is here?" It also attempts to produce all of these valid, or grammatical utterances, while excluding such un-English statements as "Man the here is." "Here man the is." and "The man are here."

In the early days of generative-transformational grammar, linguists believed that there was a core of simple

basic sentences in a language from which all of the other sentences in a language could be derived and that there were transformations that applied optionally to different structures. Later these ideas were abandoned, and most linguists advocated the belief that all sentences, even the simplest, are derived by transformations. This derivation of sentences led to the notion of deep structure, the underlying structure of utterances (as opposed to surface structure, what one actually says). It began to appear to workers in this field that the deep structures formulated for various grammars (*i.e.*, the grammars of various languages) came to be more and more alike, thus suggesting that all languages may share to a great extent, or even completely, the same deep structures.

Deep and surface structure in transformational grammar

In the early 1970s there were many linguists who believed that no strict discrimination could be made between "deep" and "surface" structure and that the most fundamental portion of a grammar is a base component consisting simply of abstract semantic structures or configurations. In that case, this source semantic structure would be largely general for all grammars. Regardless of which model is accepted in detail, the important fact was the recognition in the 1960s that all languages may in some sense share a basic facet of grammar. Thus it is evident that a renewed interest in universal grammar has displaced the focus on relativism.

Further conclusions have subsequently been drawn from the claim of a universal base component for all of human language. The existence of such a component would help to explain how a child in any society manages to master to a surprising degree in a few years a large part of the competence that is called "speaking a language." Such theses have important consequences for the study of developmental psychology. The properties of transformational grammars have also been exploited in probing questions of linguistic behaviour that have psychological implications. In brief, a strong and renewed connection has been forged between grammar and the field of psychology; connections of this sort were resolutely rejected by many structural grammarians. Finally, problems of the form of transformational grammars and the nature of the semantic base component have given rise among grammarians to a very active renewed interest in logic; there are indications that philosophers and logicians in turn have a newly developed interest in contemporary grammar, and there are signs, too, that linguistic anthropologists are attempting to bridge their grammatical interests with contemporary psychology and logic. Except in poetics, however, contemporary grammar has as yet had little impact on literary studies; the connections with philology, which were so close during the 19th century, are now rather weak.

Relationships and independence of grammatical study

The Greeks saw an intimate link between questions of philosophy and those of grammar; indeed, they were from one point of view one and the same. Today grammar again enjoys a close relationship with philosophy (and its one-time offshoots, psychology and anthropology); but this time grammar is a well-established field in an autonomous discipline.

THE QUESTION OF GRAMMATICALITY

Older definitions. There has been considerable discussion of what is meant by the notion "grammatical." If grammar is considered to be the art of speaking and writing correctly (as with the ancients), "grammatical" means what the poets use or what the best (*i.e.*, most erudite, tasteful, or socially acceptable) people say. This leads to a normative prescriptive meaning of the term, which is still in use. If grammar formulates what partakes of nature (as with the Stoics), "grammatical" means "physical" or "logical"; if it is concerned with the properties of things (as with the scholastics), it means "real" in their technical sense; if grammar treats of human reason (as with the Port-Royal thinkers), then it means "rational."

If, however, grammar forms part of an autonomous study, then an explicit understanding is needed that may be applied in specific cases; it is difficult to find total agreement on this issue, and much depends in detail on one's particular view of grammar. Glossematic grammar, originated by Louis Hjelmslev, a Danish linguist, views its task as that of partitioning an enormous text (the language); in this sense "grammatical" approaches "attested." Since structural grammar is based on the observation of functional oppositions or contrasts, "grammatical" in one sense means "distinctive"; in another sense it means that the expression in question fits within the outlines sketched by the descriptive grammarian.

If the bias of the grammarian is strongly empirical, positivist, and procedural, he may have limited his grammar to his corpus of data or to one aspect of it; this leaves out of account a very significant aspect of human language—its capacity to produce and predict in any one speaker an indefinite number of sentences. To bridge this gap the notion "potential" has been proposed; *i.e.*, extrapolating from a model, one speaks of certain structures as potentially occurring. Apart from its vagueness, the notion of potentiality is also subject to a kind of indeterminancy or variability. A structural grammar may be specified to varying degrees of completeness. It is clear that as the completeness varies, the amount of room left for potentiality and the degree of restriction in the notion of grammaticality will fluctuate.

Generative views. The introduction of generative grammar has both refined and complicated the question of grammaticality. By "generative" is understood "explicit, complete, precise, and perhaps formal." If a language is taken to be a set of sentences and a grammar is a theory that generates and assigns descriptions to those sentences, then one view is that an adequate and appropriate grammar generates all of and only those sentences; *i.e.*, any other sentences past and future are ungrammatical. This is a demanding requirement.

A refined version of this view clarifies matters in relation to the structuralist position, which would often include in the account lapses or distortions or truncations of various sorts, just as unassimilated loanwords from other languages would be incorporated in the grammar with equivalent status. In such a view the language competence or knowledge of a speaker must be differentiated from his performance amid the accidents of the individual, the temporal, the real, the physical, etc. This distinction bears some relation to the concepts of *langue* (a language system) and *parole* (an individual's use of a language) proposed by a Swiss linguist, Ferdinand de Saussure, in the early 20th century. Thus one may speak of grammaticality in relation to competence.

There is still debate, however, as to how a speaker's knowledge may be tested and determined when individuals differ in minor ways on what they will accept or discriminate or call equivalent (*e.g.*, "He was talked to deliberately." will be accepted by some speakers, but not by others). Despite this variation there is often widespread agreement that an expression is nearly but not quite acceptable or grammatical; a good grammar should not merely rule out such an expression as ungrammatical, but it should characterize it as, and even specify the degree to which it is, deviant. Much poetry and humour are deviant.

Variation and grammaticality. In addition to the problems of acceptability, knowledge, and deviance, there is the complicating fact that no speaking population of consequence is homogeneous; social and regional variation must be not merely recognized but positively provided for and accommodated in the grammar. Otherwise one fails to account for the ability of speakers to communicate readily across some cleavages but not across others; they will both note and evaluate some differences while understanding completely, or they may ignore even greater differences, or they may reject divergences that seem perfectly predictable to an outsider.

Individuals in most societies employ more than one fashion of constructing expressions, depending on the situation and addressee (*e.g.*, "Lug that out of here." "Be so good as to convey that to the efferent system."). Such variant fashions may be called registers. These variations should be clearly distinguished, as they sometimes are not, from style. Style is neither grammar nor a component of a grammar but consists in the introduction, op-

tionally, of values designed to gain a particular effect; it is therefore a complex and cumulative result of semantic choices that the speaker has made.

Diglossia

In some communities (*e.g.*, Baghdad or Greece) two sharply different varieties of the same language are in use under different but well-defined conditions. This special type of bidialectalism, which is quite different from the pervasive phenomenon of register, has been called diglossia. For such situations, it is possible to write a slightly expanded grammar in which the grammatically compatible forms will be correctly selected for the diglossia situation. A similar duplicate marking can accommodate slang, argot, or jargon equivalence (*e.g.*, starboard = right, bulkhead = wall, ladder = stairs, secure = tie, and the special derivations of fore, forward, aft, after, and so on).

In all these cases of register, argot, and code-switching it is clear that the speaker does not make independent choices at every turn, but selects by a single rule the desired set of compatible alternatives. In various ways it is seen that "grammar" and "grammatical" are complex, slightly blurred, flexible, and expandable notions.

MODELS OF GRAMMATICAL DESCRIPTION

Universal grammar. In recent years, as the interests of many students of grammar have shifted from the observation and classification of the facts whereby languages differ on the surface to the more abstract ways in which the broader or "deeper" characteristics of different languages are the same, the empirical investigation of natural human languages has shifted once again to a strong interest in general, or universal, grammar. The object of such study is to distill from the analysis of natural languages what they hold in common; *i.e.*, the grammatical characteristics or rules that hold for all human language.

From a psychological point of view, this holds great interest by potentially stating what is essentially and typically human about language and what types of language are humanly possible or expectable. From a formal point of view, the investigation of this aspect of what has come to be called algebraic linguistics has led to a new rapprochement with certain forms of mathematics, for it is possible to study abstractly selected mathematical structures and relations that have been found to or are believed to resemble those general abstract characteristics called universal grammar. This is more a program for the future than it is a present reality. The universal grammar that may be sifted from empirical language-based study is certain to be a construct of rather complex relations and properties; the mathematical models that have been discussed and proposed to date and that can be manipulated with clarity are not very rich, relatively speaking. Thus there is a considerable gap for scholarship yet to bridge. The false impression is sometimes given that modern grammar has yielded to mathematics; even when the two fields become fully linked, there will always be an empirically based study of natural grammars.

Mathematics and grammar

Taxonomic, or structural, models. A number of models have been proposed for grammars, and several have gained wide currency. In each case they reflect the scholars' intellectual, philosophic, and practical orientation. Among taxonomic, or structural, models, which are strongly (but not necessarily) linked to the search for basic grammatical elements, three versions have been notably cultivated and discussed: Item-and-Arrangement (IA), Item-and-Process (IP), and Word-Paradigm (WP).

Item-and-Arrangement. Item-and-Arrangement stresses the isolation of basic elements (also called primes) and then undertakes to present all arrangements, or juxtapositions, of these elements found in sentences. For example, this type of analysis explains the plural form "dishes" as the juxtaposition of "dish" and the plural element *-es*, the comparative form "older" as "old" and the comparative element *-er*, the participle "drunken" as the juxtaposition of "drunk" (rather than "drink") and *-en*. The items described all occur in actual speech.

Item-and-Process. Item-and-Process isolates primes and then applies processes, or rules, or formulas, that sometimes unite these primes, sometimes adjust their shape or their sequence, and sometimes do both to yield constructions and sentences that are found to occur. For example, it explains the form "dishes" as the result of the process of combining "dish" with the element "plural," and "drunken" as the result of combining "drink" with the function "past participle." One may restrict the processes so that they operate only on some actually occurring surface shape in order to produce the other desired shapes, or one may set up abstract primes that never appear in actual speech but that undergo processes in all instances. Both IA and IP share the characteristic of starting from an inventory of elements and building these up into constructions occurring in observed sentences of the language; in this sense they may be viewed as two devices for assembling language elements. For many grammarians an advantage of the dynamic IP model over the static IA is in the more natural derivation of a form such as "feet" from "foot" + PLURAL.

Word-Paradigm. The grammar is viewed from a different vantage point by the WP model, which observes the relation between the forms that words take and the function of those words in sentences. The various interrelated forms of words are arranged in patterned lists called paradigms; for example, in Latin:

dūcō	"I lead"	*dūcimus*	"we lead"
dūcis	"you lead"	*dūcitis*	"you lead" (plural)
dūcit	"he, she, it, leads"	*dūcunt*	"they lead"
amīcus	"friend"	*amīcī*	"friends"
amīcī	"of the friend"	*amīcōrum*	"of the friends"
amīcō	"to the friend"	*amīcīs*	"to the friends"
amīcum	"friend" (accusative)	*amīcōs*	"friends" (accusative)
amīcō	"with the friend"	*amīcīs*	"with the friends"

Latin paradigms are composed of a fixed set of endings, which can be attached to any stem (or base) belonging to a particular class. In the first example above, the endings are *-ō, -is, -it, -imus, -itis, -unt* and the stem is *dūc-*. The same ending can be used with *cadere* "to fall" (stem: *cad-*), *crēdere* "to believe" (stem: *crēd-*), *īnscrībere* "to inscribe" (stem: *īnscrīb-*), *perdere* "to destroy, lose" (stem: *perd-*), and many other verbs of the same class. Such a grammar prominently depends on and leads to the notion of class. While WP differs markedly in its approach from IA and IP by attaching its primes to the paradigm as a whole, it shares with them an emphasis on class and taxonomy and on the unordered characterization of sentence syntax by relations based on primes of smaller scope.

Transformational models. A transformational grammar, of which the first contemporary exponent was Noam Chomsky, recognizes, essentially, two kinds of rules. One, which may be called an expansion (or phrase-structure) rule, rewrites a unitary element into two or more elements, which themselves may further be rewritten into multiple units by other rules. Such a set of rules yields a branching tree (which may also be displayed as a successive bracketing) with labelled nodes dominating each successive branching. It will be immediately noticed that the starting expressions of such a derivation are the most inclusive (*e.g.*, the sentence, or even some higher unit). The derivation then proceeds through increasingly less inclusive constructions until it reaches all possible terminal strings (*i.e.*, arrangements) of elements for every possible path through the nodes and branches of the tree. A transformational rule, however, may rewrite two (or more) symbols into another expression or, equivalently, may transpose a node in a derivational tree to another position in the same tree; thus a whole phrase or clause may be transformed and rearranged to function differently in the sentence (the arrow in the following examples should be read as "is rewritten as"):

He caused the wine to get to be cool → He chilled the wine.
(The fact) that he chilled the wine improved it → Chilling the wine improved it.

(It should be noted that a great many transformations actually underlie these innocently simple sentences.) The transformational model is able to account for the infinity of sentences that make up a language with a finite (even rather small) set of rules because one can apply the transformational rules over and over again to the original basic sentence and the sentences derived from it.

Debate over form of rules

There is a continuing debate regarding the most appropriate form of the rules of such grammars. This affects not merely the details of their shape, content, and formal composition but also their status and interrelation in the total grammar. A widespread version insists that they should be ordered completely, partially, or by sets in their application; another version would have a rule apply every time a structure appropriate to its input appeared as the output of earlier rule applications. Both these versions require that the output finally be precisely grammatical or acceptable. Still another view would generate many more outputs than those that are grammatical and then pass these through the filters of output restrictions.

FEATURES OF GRAMMATICAL ANALYSIS

There are some properties and topics that have been attributed or implied in one way or another to most types of grammar that have been seriously considered by grammarians.

Distribution. Distribution is the position (or sum of positions) held by an element in a chain, or sequence, of elements; thus it implies "before" and "after." This may be referred to the temporal spoken or written chain ("The bird is seen by the dog."), or to an abstract structure considered to underlie it ("The dog see-s the bird."). These sequential aspects of distribution are referred to as context (or environment). Though it is surely not fruitful to posit context-free grammars for human languages (*e.g.*, the plural *-ren* is found only in the context "child-"), rules may take context into account (*i.e.*, be context sensitive) or may not. Some proponents of taxonomic grammar consider distribution so important that they have claimed that meaning is determined by or is coextensive with the totality of distribution(s) of an element; *i.e.*, the meaning of "optometrist" would mean the sum total of all of its occurrences. This effectively empties "meaning" of any independent standing, and grammar (and indeed language) is then viewed as nothing more than lists and classes of elements and their distributions.

Traditional parts of speech

Classes. All grammatical literature recognizes the notion of class in some form but accords it very different status and goals according to the doctrinal view. Any set of elements sharing a property may be said to form a class; some classes, as the parts of speech of traditional grammar, have been set up on multiple criteria and, hence, present inherent problems of overlap (or intersection) and indeterminancy. Items sharing a given distribution form a class; *e.g.*, stems occurring before the plural may be called nouns or, more inclusively, stems that are not compared, thus excluding adjectives, but that follow "the" may be classed as nouns. In traditional grammar, nouns generally were implied to behave in the manner just mentioned and also to name persons, things, and so on. Such mixed criteria yield mixed classes.

If a form of grammar (*i.e.*, a taxonomic grammar) is much occupied with establishing such classes it rapidly encounters problems of consistency and definition. If all elements that may undergo the same rule in a generative grammar are allotted to a class, the result is a very large number of intersecting classes; there then arises the problem of evolving relevant criteria to choose the important classes that are to be recognized. Proposals have also been made for classes (and matching terminologies) that rest on selected types of criteria; *e.g.*, "nouns" are to be based on criteria of word morphology (plural affixation and the like), and "nominals" on those of phrasal or of sentence syntax (occurrence with "the," as subject of a verb, and so on). Such attempts, however, depend heavily on the theoretician's view of grammar in detail, on idiosyncracies of particular languages (the number of important levels of phrasal or sentence integration), and on a persevering desire to classify regardless of the proliferation in numbers of discriminations.

Though this is an open question in theory, there is a need to recognize some important classes (*e.g.*, stative verbs, verbs of perception, as well as verbs as a whole), while other classes are clearly minor (*e.g.*, "havoc" and "vengeance," which occur with "wreak") or uninteresting (*e.g.*, "bay," "to curry," and other words applying solely to horses). Nevertheless there exist cultures in which it is important to discern features that apply to horses as a diagnostic of that culture—many languages differentiate words for "eat" and "die" according to whether they apply to men or animals. Such a consideration of classes leads naturally to the question of what, in generative grammar, have been called selectional and sub-categorization features and rules.

In much of the discussion concerning distribution, classes, and contextual alternants that depend on distribution and classes, there is a confusion of two independent considerations, the internal constituency or organization of the element in question and the external identity or classification of the adjacent elements that surround it. Many grammarians have regarded these as a procedural means toward the discernment and statement of grammatical elements, classes, and constructions.

Analytical procedures

Levels. The internal organization of a grammar may be viewed as being more or less separable into levels of integration; grammarians differ greatly in this view, since it depends in detail on the doctrine preferred, but all consider it. For almost all schools of thought there is a break between phonology and the other levels. For taxonomists, grammatical (*i.e.*, nonphonological) units are made up of, or constituted by, other grammatical units while they are represented by phonological units; for those who urge more separate levels or strata of grammatical organization, such mapping occurs between each pair of levels. For grammarians who follow J.R. Firth, an English linguist, ranks are discriminated within levels, and rank units are composed of lower rank units. In generative grammar, the phonological component is discriminated from the syntactic (or semantic) component, and these are distinguished by the difference in feature content that is introduced into the rules of each.

Grammatical segmentation. Even though most grammarians today deny that the proper study of grammar is a procedural discipline, certain procedures are well recognized and recur in all discussions. All grammars presuppose a segmentation or partitioning, although only some (proponents of IA analysis) have urged that this is an important aspect of the grammar proper or that all segmentations once made and accorded status must remain undisturbed by later operations governed by the grammar. It is by segmentation that units (elements, primes, and constructions) are recognized. On the basis of provisional segmentations, commutation (*i.e.*, substitutions of replaceable elements within stable frames) is carried out—*e.g.*, "The dog/cat/horse ate his food."

The procedures involved in segmentation of utterances are tolerably well understood and recognized and lead most immediately to the isolation of surface elements. There are procedures being evolved for the recognition of appropriate underlying abstract representations and rule formation, but these are as yet neither well codified nor generally agreed upon. Some of them involve notions familiar from logic and other formalizing disciplines (disjunctive and conjunctive ordering, implication, and so on); others are extensions of intuitive semantic perceptions (including equivalence, paraphrase, ambiguity).

Morphemes as ultimate syntactic units

Morphemes and words. The isolation of units by segmentation (and any other means) leads to the ultimate syntactic units, which are called morphemes or formants. A morpheme is generally considered as a single unit in the surface structure, regardless of its underlying sources. For most grammarians "cat-s" has two morphemes, and "book-bind-er-s" has four. The correct partitioning of "men" (:"man") has caused much debate, but everyone agrees that it results from two formants: "man" + PLURAL.

"Word" is a notion that appears to be useful and applicable in a wide range (probably all) of known languages, but it is very difficult to arrive at a set of criteria applicable to every language to determine exactly which parts of an utterance constitute words. All languages tend to have cohesive units that the speakers normally isolate instinctively and can call words. Such units typically show a large number of specimens with but one element that serves as a nucleus (or root, or base) and that may

in most languages be accompanied by a number of satellite elements. The difficulty is not so much that of finding entities that are obvious "words" (though some languages, such as Eskimo or Oneida, have words that seem very long and complex to speakers of English), as that of finding criteria to segregate all words and not include phrases and the like (*e.g.*, "the king of England's hat"; "of course"). Some morphemes or formants can occur as independent words (*e.g.*, "occur"), while others only occur attached to other morphemes (*e.g.*, *-ence* in "occurrence"). This leads to the recognition of free and bound elements, or morphemes.

This notion, however, does not always lead uniquely to the constructions that one seeks to recognize. In many languages free forms turn out to be roots, or bases (Latin *clam* "secretly"), and in English very many bases may occur as free forms. But in Latin a noun can never stand as a single free form, since a case ending is normally also required; the same is largely true for a Latin verb and certainly for a verb in a declarative sentence. Similarly, some bound elements occasionally turn out to be what might be called separate words in phrases (*e.g.*, "kith" and "kin," "to" and "fro," "wreak havoc"). Therefore the search for a neat universal specification of the "word" is probably illusory; yet the concept is so widespread and easily grasped, apart from being analytically fruitful, that grammarians will probably continue this search.

Prefixes, suffixes, and infixes

Affixes and allomorphs. In those very many instances in which there are words that are not identical with morphemes, one may discriminate the base from its satellites, or affixes. Affixes are conveniently subdivided into prefixes (occurring before the base), suffixes (occurring following the base), and infixes (inserted into the phonetic substance of the base). Infixes are frequent in Indonesian and Southeast Asian languages and are illustrated by the *-n-* in Latin *vi-n-c-ō* "I conquer," as compared to *vic-tum* "conquered," which has the same root *vic-* without an infix.

Among most languages (perhaps excepting a few such as Quechua or some of the Philippines or Oceania) a large portion of the stock of morphemes varies according to context (these are called context-sensitive forms); such systematic (and even irregular or unique) variants have been called allomorphs. Many involve simple widespread adjustments in their sounds (phonetic assimilation and the like), but some languages show large numbers of nonpredictable variants that must be listed or require intricate rules of minor scope. European languages often show such irregularity. An extension of such idiosyncracies turns up in languages of culture such as Japanese, Tamil, or English (usually arising when many forms have been borrowed as stylistic or technical variants from another language). If in English one is justified in stating that a rule relates "go" and the past form "went" ("wen-t" = "slep-t," "learn-t"), one may state a similar relation between "blood" and "hemat-ology" (never "blood-ology" or "blood-lore" as in some languages). Whole words may stand in an allomorphic relation; *e.g.*, among English pronouns, "I" is the variant found immediately before the verb or the auxiliary, even when deleted as in "bigger than I (am)," and "me" in all other contexts.

Distinction between morphology and syntax. The recognition of the widespread notion of "word" leads to the frequent utility of a discrimination between morphology and syntax, the fields of intra-word and inter-word integration, respectively. There are languages (*e.g.*, Chinese and Vietnamese) in which most words are composed of a single morpheme, and in which there is consequently little that can fruitfully be called word morphology. English illustrates this type of structure to a high degree (*e.g.*, in "You see how I can drink cold milk with a straw."), but there are also complex words such as "your," "saw," "could," "drunk(en)," "colder," "milking," "straws," combining more than one morpheme or element of meaning.

Grammarians have sometimes expressed surprise that there exist languages in which many words seem to englobe whole clauses or sentence constituents (*e.g.*, Nootka or Eskimo). This tends to obscure the fact that an English word such as "chilling" is really a surface manifestation for a clause roughly of the form "(the fact) that one causes to get to be cool." One should not be surprised to find words that embrace whole underlying syntaxes; it is rather that in English such cases are comparatively irregular and particular (*i.e.*, apply in idiosyncratic ways to different elements), while in Nootka the process is more regular and pervasive. Moreover, single surface morphemes often represent a complex of underlying grammatical formants; *e.g.*, "bull" = BOVINE + MALE, "calf" = BOVINE + IMMATURE, "veal" = (BOVINE + IMMATURE) + (MEAT + EDIBLE). It is clear that words are surface phenomena, and that their relationships with syntax, sentences, and semantics are complex and intertwined.

Stems and inflections

Stems and bases. For the many languages that show an interesting word morphology, a stem (which may in cases be identical with a base) and its inflections are normally distinguished. Word classes are often discriminated and identified by stems; *e.g.*, "accommodate" is a verb, "friendship," "timeliness," and "politicizationist" are nouns. A stem may comprise a base plus one or more derivative affixes; the last added affix either alters or preserves the class of the stem to which it is added (*e.g.*, the *-ist* ending on "politicizationist" preserves the stem class of noun). A stem may often be composed of a base plus a stem-forming theme; *e.g.*, in Latin, the word *dūcimus* "we lead," with the stem *dūci-*, has *dūc-* as its base and *-i-* as its stem-forming theme, indicating that it is a verb of the third conjugation. Finally, a simple base itself may be specified for class; thus, while most bases in a Semitic language may be viewed as verbs that can be changed into nouns, a restricted number ("father," "dog," "heart") are nouns.

Although the structure of stems, when distinctive, may be a major criterion for the identification of word classes, their use in sentences ultimately determines their class, or part of speech. Thus, while one may argue that "word" is a noun because it can take plural and possessive endings or, conversely, that it is a verb because it can take verbal endings, its function in a sentence like "He worded his argument poorly." indicates "word" to be a member of the verb class; and its function in the sentence "The baby learned three new words." indicates it to be a noun.

Inflection. An important aspect of the morphology of words in many languages is their inflection. This consists in the addition of markers that indicate relations with other words in the sentence or that incorporate into the inflected words abstract formants from elsewhere in the sentence structure. For example, Latin words vary (*i.e.*, are inflected) for case, number, person, tense, mood, and so on. In addition to the variations in the forms of inflectional affixes used in a language, there are also certain widespread types of restrictions concerning the appearance of these affixes. Thus, one element may determine or call forth the form of another; this is called government (or rection). In various European languages a preposition governs a particular case form of its noun or pronoun; *e.g.*, in German, the use of the preposition *mit* "with" determines that the following noun or pronoun be in the dative case (rather than the accusative case)—*Ich ging mit ihm zum Bahnhof* "I went with him to the station."

Concord (or agreement) is the term applied when two or more inflections (on as many words) correspond as to number, gender, case, and other grammatical categories; one is always to be regarded as the head, and the other is said to agree with it. For example, in French, the noun *maison* "house" is classified as a feminine noun; in the phrase *la maison blanche* "the white house" the modifying elements *la* "the" and *blanche* "white" also appear in their feminine forms, in contrast with *le ballon blanc* "the white ball" with the masculine *ballon*. The inflectional property is assigned to the head and thereafter an appropriate mark is transported to the agreeing word or words.

Generative grammarians have proposed that properties analogous to these traditional features of concord and government also apply to the more abstract syntactic (or semantic) structures. Thus there is a type of concord that can be seen by assigning appropriate selectional fea-

tures so that in a single well-formed sentence certain verbs (*e.g.*, compute, parse, pray; eat, give birth) will be prevented from having subjects that are not human or animate, and so forth.

Grammatical categories. The entire subject of categories is a vast one and can only be touched upon here. The term has been applied to very different things, in part confusingly so. Entities such as unit, class, construction, and so on are really frameworks for grammatical relation. Such concepts as noun, verb, and the like have been lumped under this rubric; they are best called form classes or, traditionally, parts of speech. Elements of the sentence such as subject, object, and so on are appropriately referred to as syntactic categories, but their status differs greatly in different theoretical views. For some generative grammarians they are stable basic points of reference in the most abstract representations of a sentence; for others they are categories of syntactic emphasis to which more abstract and basic semantic features are allotted. Then there are typical parameters of word or phrase paradigms: number, gender, definiteness, genericness, case, person, tense, mood, aspect. These are the grammatical categories proper, and they pervade the grammars of most or all known languages, although the list of overt categories will vary somewhat from language to language. It appears that grammatical categories and parts of speech as usually recognized belong to the surface structure of language.

Some form classes (*e.g.*, adjuncts and so-called sentence adverbs) have resisted close definition, sometimes appealing to such a criterion as intonation, and stand on a different footing. Some categories, more abstractly taken, turn out to have very ramified implications; that of person, since it involves and flows from the speaking situation, is bound up with that of deixis, or the kind of pointing that demonstrative pronouns normally carry out (I = this; thou = that (near); 3rd person = yonder). The topic of categories leads naturally to semantic features and to linguistic universals. It is further notable that the notional grammar of the Danish linguist Otto Jespersen also involves extralinguistic categories claimed to be universal but imperfectly expressed in actual languages.

Up to this point the focus has been on those aspects of grammar that are mirrored in overt shapes in the surface structure of sentences. Some grammarians have claimed that true grammar inheres only in such visible or audible "recurrent partials." Yet in their practice all grammarians have been forced to deal with abstract relations, at least in questions of constituency and construction.

Segmentation into immediate and ultimate constituents

Constituents and immediate constituents. When sentences are segmented not all segments are equally revealing or cohesive; thus one seeks to divide spans of speech into immediate constituents, such that the divisions made will show maximal independence (thus "un-gentlemanly" rather than "ungentleman-ly"). The test used to establish relative independence of the constituents is that of their substitutability by other phrases, words, or morphemes. Such divisions, or cuts, usually result in just two constituents, but cases can easily be found in which multiples must be allowed.

This method of reducing a sentence to its ultimate constituents by successive cuts is a useful working procedure; it belongs to the grammarian's kit of field methods, rather than to his body of theory or guides for grammatical formulation. The results, when laid out and studied (with appropriate corrections made) and then formulated, yield the trees (or bracketing or layering) of the phrase-structure (or constituent structure) rules of a transformational grammar. In the strictest taxonomic grammars (exemplified by those of a U.S. linguist, Zellig Harris) the nodes of these trees, constructed from the twigs inward to the root, are unlabelled; in the tagmemic model of the U.S. linguist Kenneth L. Pike and in some work by Firth and his followers, the nodes are labelled, but other types of rules are not provided. In the tagmemic model one inspects all constituents and constitutes (constructions for some grammarians), typologizes these into the most inclusive number of types, and labels the types, each labelling yielding a slot. A tagmeme is a slot together with its list of filler elements, and these lists may have overlapping membership. Thus tagmeme fillers differ from form classes and from constituents, which are mutually exclusive.

The difficulty that arises with immediate constituent (IC) analysis inheres largely in discontinuities that are found (which may additionally be partly ambiguous and idiomatic, thus calling for a unified accounting); *e.g.*, "I *put* a lot of time *in* on it." "I *got* John *up* out of bed." "John *got him*self *up* out of bed." "John *got him*self *up* in a new suit." "John *got* the report *up* for the boss." Such discontinuities, as well as some instances of homophony (*i.e.*, ambiguous sentences like "Flying planes can be dangerous."), are symptoms of the shortcomings of a taxonomic constituent-structure approach. While subordination (traditional hypotaxis) generally is amenable to immediate constituent analysis, coordination (parataxis) offers certain problems of detail; *e.g.*, the assignment of conjunctions.

The typologizing of immediate constituents into endocentric constructions (in which an immediate constituent functionally equals the whole—*e.g.*, "barn" in "the red barn" functions the same as the whole phrase) and exocentric constructions (in which no immediate constituent equals the whole—*e.g.*, in "John slept" neither "John" nor "slept" is the equivalent of the whole phrase) is often clarifying but fails to reach a deep understanding of the grammar involved and says nothing of the grammatical source of the constructions.

DEVELOPMENTS IN TRANSFORMATIONAL GRAMMAR

As transformational grammarians have turned to increasingly subtle problems of sentence formation, important changes in the grammatical model have been introduced and opinion among experts has diverged sharply. Within the constituent structure portion of the grammar, in addition to the branching phrase-structure rules, another set called sub-categorization rules had been introduced by 1965. These rules enable the categories that apply to a constituent to be specified and control the selections and concords that restrict the arrangements or conjoining of elements. A notable feature of this model, too, is the provision for insertion of the lexicon (*i.e.*, the individual base and stem morphemes with their characterizing phonetic and semantic features and shapes) along with the constituent structure rules before transformational rules are applied to that structure. In current grammatical scholarship there is a lack of clarity about the nature and structure of the lexicon, its relation to or inclusion in the grammar proper, and the form of a lexical entry and rule.

Sub-categorization rules

Criticism of earlier forms of transformational grammar has been essentially of two sorts: the detection of inadequacies or insufficiencies and the claim that certain functions are misplaced, unnaturally provided for, or unwittingly duplicated since they are really implied by a proper understanding of some other component. The rule structure of a grammar has been explored and refined: the recognition of general rules has bearing for the formalization of universal grammar; the notion of minor rule helps to recognize, clarify, and formalize what is commonly called irregularity. The properties of trees and node structure have been studied and exploited, leading to several fruitful notions.

A highly important characteristic of recent versions of a transformational grammar is the elimination of double-based and of optional transformations. Earlier versions, for example, allowed a sentence to be negated by the introduction of a negative transformation (schematically, I can go → I can not go); this optional application is somewhat related to the rejected doctrine of kernel sentences that are then transformed into more complex structures, a doctrine that had characterized earlier theories. Around 1965 it became clear that all transformations are best viewed as obligatory and called forth by an abstract element in the underlying (constituent) structure. Such elements include provision for an entire sentence, represented as S, which becomes a node to which a whole embedded clause is attached by appropriate transformations. Thus the derivation of a negative now becomes

Obligatory nature of all transformations

schematically: I Neg can go → I can not go. The implications of this view are far-reaching; this means that all the basic information is already in the most abstract phrase structure of a given sentence, and that transformations, instead of introducing information, simply position, delete, or reshape elements correctly when they are triggered off; thus, except for possible optional stylistic rules (whereby style has a formalized different status), transformations never add information; they are meaning preserving.

A further claim of the 1965 grammatical model was that deep structure exists as an abstract integrated level intermediate between an autonomous semantic representation (connected by rules of interpretation) and the surface structure; this view rests on the assumption that the elements and relations of the deep structure are to be characterized as syntactic and distinct from semantic. In the view of many it has become more and more difficult, however, to maintain this distinction; the syntax of "pregnant," for example, is governed by the presence of an element FEMALE, and the subject and object of "marry" is expected to be opposite in sex even in a gender language. Such a line of reasoning leads to the conclusion that the base component, or the most abstract feature structure of a grammar, simply is semantic. The initial rules generating the structures of such a generative semantics would then be a part of or simply be the rules of logic—an interesting but highly refined return to the preoccupations of the Stoics and Scholastics. In this view a separate syntax in a grammar or a level of structure between the abstract semantic feature bundles and the phonological mapping rules would be just a complication, matching nothing real and occluding the clarity of the rule structure of a transformational semantic grammar.

An additional fundamental departure from the 1965 model of transformational grammar has been urged by the abandonment of the structures of traditional sentence parts that the constituent structure of that model generates. Formerly, a sentence, as the starting unit, was expanded in part into Verb + Noun-phrase (*i.e.*, object) and in that sequential order. The newer view would incorporate in the semantic base features associating unordered roles with various nounlike semantic entities (*e.g.*, agent, patient, instrument, and so on), rather than generated strings equivalent to subject + object and the like. These roles would then be assigned sentence values (subject, object, locative, and so forth) or cases (with inflections for Latin, for example, but prepositions for English) by rule, depending upon inherent feature-specified constraints (restrictions) in the elements that co-occur in a particular sentence. For example, the agent may not be committed a priori to be the subject of a verb ("The medicine healed the wound." "The wound healed from the medicine.") but is assigned to it as the result of a subtle choice on the speaker's part. Subject and predicate have even been claimed as surface structures, while agent and patient would be semantic roles associated with passive function. Many languages show a distinction, which has been called ergative, whereby the subject of an intransitive verb is treated differently from that of a transitive verb and may be treated like the object of a transitive verb; here it is possible that ergativity marks the agent, while the English subject is the natural assignment for assumed information. These questions of basic sentence structure and meaning open up vast and complex debates.

Features of the semantic base

As soon as the prominent part that semantics must play in a grammar is focussed upon, numerous questions of meaning arise that equally occupy the interests of philosophers. For example, the function of reference becomes crucial in dealing with grammatical processes as varied as pronominalization, reflexivization, and relativization; in each of these instances one must distinguish whether a noun has a particular reference or not.

Although it has been customary to consider the sentence (from which parts may of course be deleted or truncated by rule) as the basic unit for grammatical derivations, there is ample evidence that an adequate accounting must go beyond the sentence limit in at least three senses. Some languages, such as Classical Greek and Tonkawa, employ linking particles that show the relation of the sentence to the preceding sentence or sentences. Something similar is involved in discourses with measure expressions: "How much butter would you like?" "Two pounds, please." but never seriously: "Two yards (litres, etc.), please." In another sense one goes beyond the audible sentence when one considers the presuppositions that some structures seem to require; *e.g.*, with the adverb "anymore," as in "Do you see Jean anymore?", one presupposes that at one time the hearer did see Jean. Finally, there are restrictions that an entire conversation or situation places upon the internal interpretation of a sentence: "Won't you have another cup of tea?" is not the negative of "Will you have another cup of tea?" nor of "Gosh, will you really have another cup of tea?"

Because of rapid change in modern grammar, it is not feasible to survey all views. The interested reader should seek further information in the bibliographic references and related articles LINGUISTICS and SEMANTICS.

BIBLIOGRAPHY. Many of the following contain rich and varied bibliographic references. JOHN LYONS, *Introduction to Theoretical Linguistics*, pp. 133–399 (1968), an eclectic survey of grammar; R.H. ROBINS, *Ancient and Mediaeval Grammatical Theory in Europe* (1951, reprinted 1971), a compact, informed survey; ROLAND G. KENT (ed.), *Varro: On the Latin Language*, 2 vol. (1938), includes the text of a leading Roman grammarian, with translation and introduction; IAN MICHAEL, *English Grammatical Categories and the Tradition to 1800* (1970), based on a study of all known English grammars with a long account of the Greek and Latin background; MICHEL ARRIVE and JEAN-CLAUDE CHEVALIER (comps.), *La Grammaire: lectures* (1970), an anthology of works by early modern grammarians, including Ramus, the Port-Royal school, and several others; OTTO JESPERSEN, *Analytic Syntax* (1937, reprinted 1969), an original theoretical work by the greatest grammarian of English; FERDINAND DE SAUSSURE, *Cours de linguistique générale* (1916; Eng. trans. by WADE BASKIN, *Course in General Linguistics*, 1959), the inspiration of most structural schools; DAVID G. MANDELBAUM (ed.), *Selected Writings of Edward Sapir* (1949), this and Sapir's book *Language* (1921) are admired by 20th-century linguists of all persuasions; LEONARD BLOOMFIELD, *Language*, ch. 10–16 (1933), influenced and dominated an entire generation; Z.S. HARRIS, *Methods in Structural Linguistics* (1951), procedure-based summation of structural, or taxonomic, theory without recourse to meaning; ARCHIBALD A. HILL, *Introduction to Linguistic Structures* (1958), the most detailed post-Bloomfieldian presentation of English; E.P. HAMP, F.W. HOUSEHOLDER, and R. AUSTERLITZ (eds.), *Readings in Linguistics II* (1966), contains various European morphological views, 1930–55; LOUIS HJELMSLEV, *Sproget, en introduktion* (1963; Eng. trans., *Language: An Introduction*, 1970), presents a readable view (pp. 91–121) of the glossematic position by its originator; M.A.K. HALLIDAY, "Categories of the Theory of Grammar," *Word*, 17: 241–292 (1961), an original divergent Firthian theory; B.F. ELSON and V.B. PICKETT, *An Introduction to Morphology and Syntax*, 4th ed. (1965), presents the tagmemic position; JOHN P. KIMBALL, *The Formal Theory of Grammar* (1973), a summation of the development of the theory from 1956; NOAM CHOMSKY, *Aspects of the Theory of Syntax* (1965), a basic revision of the theory of generative grammar by its originator; R.A. JACOBS and P.S. ROSENBAUM (eds.), *Readings in English Transformational Grammar* (1970), deals with deep structure and the base component since Chomsky 1965; GEORGE LAKOFF, *Irregularity in Syntax* (1970), a 1965 dissertation that led to much of the dispute since Chomsky 1965; C.J. FILLMORE and D.T. LANGENDOEN (eds.), *Studies in Linguistic Semantics* (1971), treats generative semantics and presupposition; P.H. MATTHEWS, *Inflectional Morphology* (1972), a searching probe of taxonomic and generative theories; ROBERT WALL, *Introduction to Mathematical Linguistics* (1972), chapters 9–11 treat the properties of formal grammars and characterize transformational grammars; Z.S. HARRIS, *Mathematical Structures of Language* (1968), formalized transformations radically different from the mainstream; R.W. ZANDVOORT, *A Handbook of English Grammar*, 4th ed. (1966), a meticulous work in the great tradition; RANDOLPH QUIRK, *Essays on the English Language, Medieval and Modern* (1968), an independent, searching, and resourceful work.

(E.P.H.)

Gran Chaco

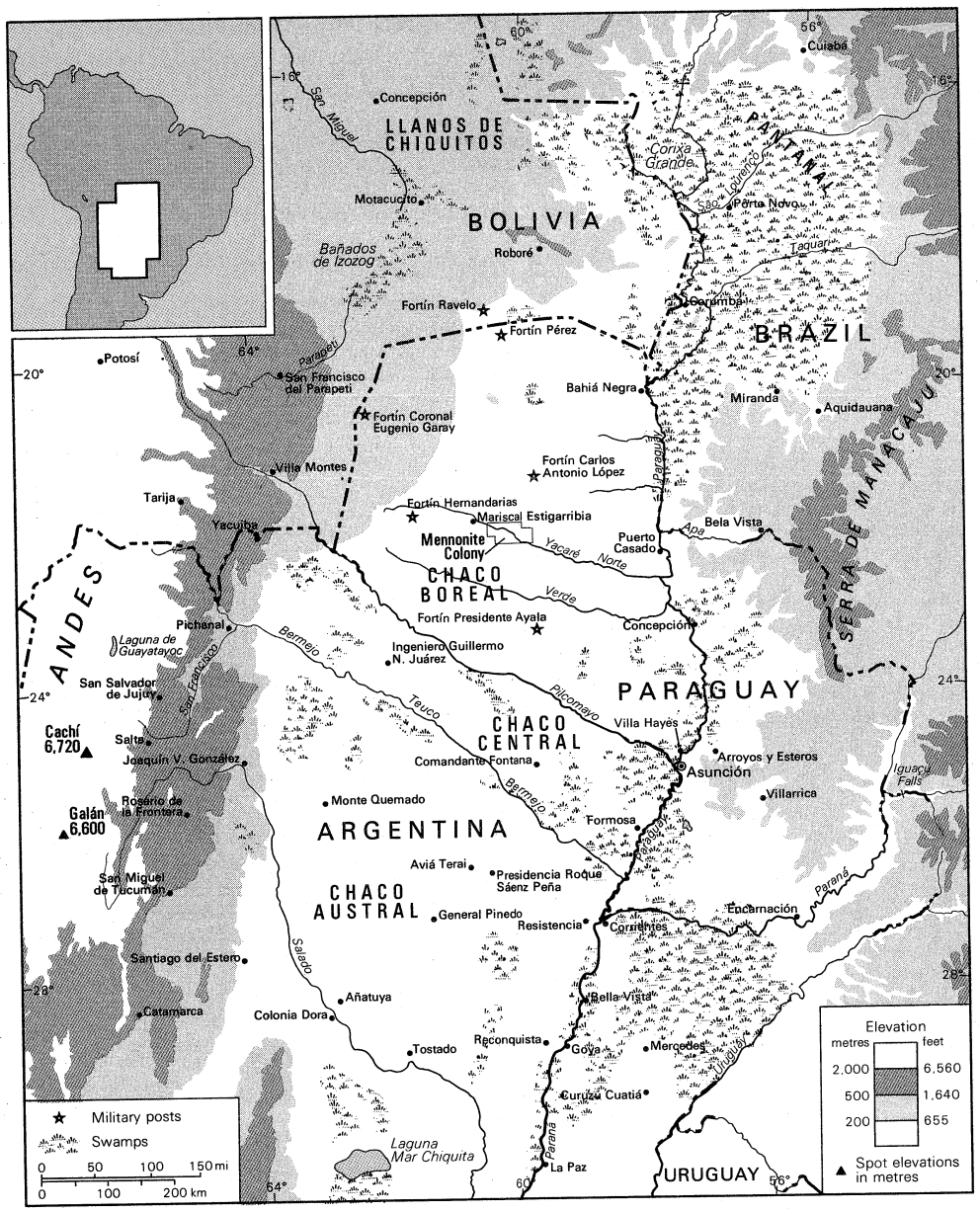
The Gran Chaco is an immense lowland alluvial plain in interior south central South America. The name is of

Quechua origin, meaning "hunting land." Largely uninhabited, the Chaco is an arid subtropical region of low forests and savannas traversed by only two rivers and practically unmarked by roads or rail lines. It is bounded on the west by the Andes mountains and on the east by the Paraguay and Paraná rivers. Its northern and southern boundaries are not as precise: its northern boundary is generally assumed to be the Llanos (high plains) de Chiquitos and Bañados (swamps) de Izozog in Bolivia, and its southern boundary the Río Salado in Argentina. Thus defined, the Gran Chaco covers about 280,000 square miles (730,000 square kilometres), of which slightly more than one-half lies within Argentina, one-third in Paraguay, and the remainder in Bolivia. Two great rivers, the Pilcomayo and the Bermejo, traverse the Chaco from their Andean headwaters to the Paraguay and Paraná rivers in the east. Widely recognized regional divisions are the Chaco Boreal, north of the Pilcomayo; the Chaco Central, between the Pilcomayo and Bermejo; and the Chaco Austral, south of the Bermejo.

Area and boundaries

The natural environment. *Geomorphology.* The Gran Chaco is the alluvial fill of a vast geosynclinal basin formed by downwarping or submergence of the area between the Andean Cordillera on the west and the Brazilian Shield on the east. Because of its alluvial character, the Gran Chaco is nearly stone free and is composed of extremely deep (up to 10,000 feet) unconsolidated sandy and silty sediments. The only rock outcrops of consequence are a few isolated remnants along the Paraguay River (in Paraguay) and some sandstone mesas in northern Paraguay and southern Bolivia.

Rivers and drainage. All but the extreme northwestern sector of the Gran Chaco is drained by west-bank tributaries of the Paraguay and Paraná rivers. The Bermejo (Teuco) and the Pilcomayo, even though they manage to traverse the Chaco, are still typical of most Chaco streams. Their courses are marked by countless sloughs, oxbow lakes, braided channels, sandbars, and vast swamplands; and they sustain such high losses from flooding, seepage, and evaporation that only a meagre portion of their full flow ever reaches the parent stream. Most of the Chaco is so poorly drained that the very shallow, irregular channels on the exceptionally level plain lead to rapid and extensive flooding during the very rainy summers. At the peak of these floods, as much as 42,000 square miles, or 15 percent, of the area of the Chaco may be inundated, although some of this is as much due to improper drainage of the impermeable subsoils as to overflow of the streams. Potable groundwater supplies are poor. Saline water is common in both deep and shallow wells, and the location and maintenance of freshwater supplies is much a matter of chance. The



The Gran Chaco.

problem appears to be greatest in the Chaco Boreal, although some hydrologists feel that large-scale studies might reveal a situation more like that of the remainder of the Chaco or like the Argentine Pampa, where groundwater problems are not now considered to be as severe as early settlers and explorers had postulated.

Climate. With a north–south extent of about 700 miles, the Gran Chaco is subject to climates that vary from tropical in the north to warm temperate in the south. Most of the region is, however, subtropical. Average temperatures vary from 65° to 75° F (18° to 24° C), with an average humidity between 60 and 75 percent. Great temperature contrasts, however, exist. Average maximums are near 80° F (27° C), and absolute maximums may exceed 116° F (47° C). The average minimum is about 57° F (14° C), although freezing winter temperatures are known throughout the region.

The highest average annual rainfall is in the east, with 52 inches, which gradually decreases to about 20 inches in the far west. Although the rainfall would normally be adequate for agriculture, from one-third to one-half the total comes in hot summer. Evaporation losses sharply reduce the effective precipitation and give the Chaco an arid nature lost only in the permanent swamps and forests along the Paraguay River.

Although light breezes are common, outbreaks of cool air from the south, called pamperos in Argentina, bring gusty winds of 32–45 miles per hour that occasionally exceed 60 miles per hour. The windiest season, however, is spring, during the transition from warm to hot weather. In the dry season dust storms are not uncommon.

Soils. Chaco soils range from sandy to heavy clay. Soils in the more humid east have more organic material and lateritic subsoils, whereas in the west are soils with less surface organic material and predominately calcareous subsoils. The local determining factor is drainage, whether a function of soil texture or relative relief. Sometimes elevation differences of less than three feet result in different soil types. Grasslands, or savannas, seem to be associated commonly with sandier soils, bushlands with poorly drained clay soils, and the forestland with better drained clay soils. In many cases the high concentration of dissolved salts in the soil water creates physiologically arid conditions in swampy sites, thus extending the arid appearance even into many areas with an abundance of water.

Vegetation. The vegetation of the Gran Chaco is intimately associated with the pattern of soils and reflects the same general east–west division. The eastern Chaco is called Parque Chaqueño for its parkish landscape of clustered trees and shrubs interspersed with tall herbaceous savannas. To the west, a wide transition zone grades into the *espinal*, a dry forest of spiny, thorny shrubs and low trees. Chaco vegetation is adapted to grow under arid conditions and is highly varied and exceedingly complex. The climax vegetation is considered to be the *quebrachales*, the vast, low hardwood forests of which various species of quebracho tree are dominant and economically important as sources of tannin and lumber. These forests cover extensive areas away from the rivers; nearer the rivers they occupy the higher, better drained sites, giving rise to a landscape in which the forests appear like islands amid a sea of savanna grasses growing as high as a man on horseback. In the more arid western Chaco, thorn forests, the continuity of which is occasionally broken by palm groves, saline steppes, and savannas induced by fire or deforestation, are dominated by another quebracho tree that has a lower tannin content and is used most often for lumber. There is also a marked increase in the number and density of thorny species, among which the notorious vinal (*Prosopis ruscifolia*) was declared a national plague in Argentina because its thorns, up to a foot in length, created a livestock hazard in the agricultural lands it was invading.

Sources of tannin and lumber

Animal life. True to its native name, the Chaco has an abundance of wildlife. Among the larger animals are the jaguar, ocelot, puma, tapir, giant armadillo, spiny anteater, many foxes, numerous small wildcats, the agouti (a large rodent), the capybara (water hog), the red wolf, palustrian deer, the peccary, and the guanaco. It is one of the last major refuges for the South American member of the ostrich family, the rhea, or nandu, and has long been noted for its abundant and varied bird population. The streams are host to more than 400 fish species, among which are the salmon-like dorado and the aggressive piranha. Countless travellers' tales complain of the pestilent insects. Reptiles are also abundant, with numerous lizards and at least 60 known species of snakes, including many pit vipers and constrictors, while at least six species of poisonous tree toads have been identified.

Human ecology. The indigenous tribes of the Chaco were numerous. Because of their subsistence as hunters, gatherers, and fishermen, tribal units were not much larger than extended families. Nevertheless, from among the diverse dialects, anthropologists have described a few major linguistic associations: the Guaicurú, Lengua, Mataco, Vilela, Zamuco, and Tupí. Most of these people lived under extremely primitive conditions, with settlement dictated by the availability of fresh water, making stream courses the most coveted sites. Implements were fashioned largely from wood and bones because of the absence of stones, while the spiny leaves of the pineapple-like groundcover *carraguatas* served as the universal fibre source. The harsh Chaco forest, quite surprisingly, contains more plant sources of human sustenance in the form of edible pods, fruits, berries, and tubers than surrounding areas, and this factor was well exploited by the native tribes. Game was gathered by trapping, netting, clubbing, and spearing, often in conjunction with large group drives. For those Indian groups still living outside the limits of European settlement, conditions are only slightly modified today, although these people now have domesticated animals and metal tools. Most tribes, however, exist as sort of a peasant pioneer fringe and practice some form of shifting subsistence agriculture.

Aside from the scattered, although successful, *reducciones* (agricultural communes) of the Jesuits, the Chaco defied effective European occupancy until well into the 19th century. Hostile Indian tribes, in concert with the forbidding nature of the Chaco itself, limited European influence in the colonial period to a situation much like a state of siege.

Natural resources and their exploitation. The limited early colonization in the Argentine and Bolivian Chaco was based on exploitation of the longhorn criollo cattle that roamed half wild throughout the region. The western Chaco Austral, near Salta, was also exploited as a source of heavy timbers for the mines in the highlands of Bolivia and Peru. In the late 19th century, the Chaco in Argentina and southern Paraguay became a land of great ranches trading in the criollo cattle; and numerous, small, independent camps of woodcutters (*obrajes*) exploited the abundant hardwoods of the Chaco forests for lumber and firewood. Cattle grazing is still the most extensive land use, with few substantial changes from pioneer days. One of the key problems in improving industry is the apparent endemic nature of many serious cattle diseases and pests against which the criollo cattle have developed some immunity, whereas purebred cattle are fully susceptible.

In the eastern Chaco, vast, highly capitalized industrial ventures established large plants to process the great quantities of tannin found in the various quebracho species. Unlike the *obrajes* of the woodcutters, these operations were large, centralized mills adjacent to rivers or rail lines from which the selective cutting of quebracho has proceeded at a systematic pace. The slow growth habits of the quebracho trees, however, may result in the demise of the tannin industry, as the pace of the harvest has greatly exceeded reforestation. It is uncertain if the relatively untouched Bolivian Chaco holds sufficient quebracho timber to offset diminished production in the exploited sectors or if synthetic products will supplant the present demand for quebracho tannin. Other forest products include lumber and heavy timbers from a variety of other species, firewood, and palo santo oil (holywood oil), from the wood of *Bulnesia sarmientii*, a tree found in the more arid portions of the Chaco.

The tannin processing industry

Although wild cotton had been known in much of the Chaco since pre-Columbian times, it had never been raised as more than an agricultural curiosity until the 20th century. During World War I, with cotton prices at a peak, large areas in Chaco province of Argentina were turned over to cotton culture. Despite bad markets, insect plagues, and often poor weather in many years since that time, the crop area has increased to over 1,000,000 acres. Production methods are antiquated, however, with nearly total reliance on thousands of migrant labourers. The crop is used for both fibre and seed-oil production. Lesser quantities are found in the Paraguayan and Bolivian Chaco.

Discovery of oil in the Bolivian piedmont in the 1920s led within a decade to the disastrous Chaco War between Bolivia and Paraguay, whose leaders held hopes of finding more oil in the neighbouring Chaco Boreal. Paraguayan claims were eventually honoured but did not include any part of the oil-rich piedmont. Subsequent explorations have been disappointing, but hopes still exist for future discovery.

Future prospects. Since World War II, efforts have been made by the respective governments to spur colonization and settlement of the Chaco. Argentine efforts have concentrated along the railways out of Resistencia and Formosa, with pioneer settlements composed mainly of eastern European immigrants and based on cotton production. In the central Paraguayan Chaco, where the Trans-Chaco Road from Villa Hayes to Fortín Coronel Eugenio Garay was completed only in 1965, Mennonite immigrants from Canada had settled in the 1920s and were joined by co-religionists from the Soviet Union in the 1930s. These settlers established self-sufficient colonies and were joined by another large contingent of refugees from the Soviet Union after World War II. These colonies support a population of about 30,000. The primary land use in the Bolivian Chaco is still open cattle range. With nearby supplies of oil and gas in addition to hydroelectric and water storage on the fast-flowing piedmont streams such as the Pilcomayo, future development of the Bolivian Chaco may be more feasible than in some other parts of the Chaco.

The Mennonite colonies

In each of the various countries, the Gran Chaco is viewed as a region with immense yet untapped potential. Notwithstanding this optimism, there exist severe inherent limitations on its development, among which are its isolation from markets due to poor transportation capabilities; extreme climatic variability, with both severe flooding and extended drought; serious endemic pests posing threats to humans, cattle, and crops; saline soils; and poor groundwater supplies. Any large-scale settlement effort must have great financial and technical capacities if it is to overcome these obstacles.

BIBLIOGRAPHY. PRESTON E. JAMES, *Latin America*, 4th ed. (1969); and OSCAR SCHMEIDER, *Geografía de América Latina* (1965), are two complementary works; James focusses on the political units and land use, while Schmeider discusses physiographic regions. "The Indians of the Gran Chaco," in JULIAN STEWARD (ed.), *Handbook of South American Indians*, vol. 1, pt. 2 (1946), consists of two chapters by ALFRED METRAUX and JUAN BELAIEFF. SIR CHRISTOPHER HERBERT GIBSON, *Enchanted Trails* (1948), is the tale of the life of a cattle rancher in the Gran Chaco in the mid-20th century. SIR JOHN GRAHAM KERR, *A Naturalist in the Gran Chaco* (1950, reprinted 1968), is an account of expeditions along the Bermejo, Pilcomayo, and Paraguay rivers in 1889–97.

(G.E.Ma./M.D.H.M.)

Grand Canyon

Cut by the Colorado River, the Grand Canyon, the world's most intricate and complex system of canyons, gorges, and ravines, is part of the Colorado Plateau, an upthrust area generally circular, located mainly in northwestern Arizona in the southwestern United States. The plateau, about 140,000 square miles (360,000 square kilometres) in area, comprises essentially horizontal, layered rocks and lava flows. The trunk gorge, the unifying link for all canyons and gorges, is the continuous canyon cut by the Colorado River, the Grand Canyon. This deepest and most spectacular of all the Colorado Plateau canyons, frequently called the awesome abyss, is about 5,300 feet (1,620 metres) deep.

Location, dimensions, and colouring

The Grand Canyon is a broad, intricately sculptured chasm that contains between its outer walls many imposing peaks, buttes, and canyons. It ranges in width from 4 to 18 miles (6 to 29 kilometres) and extends in a winding course from the head of Marble Gorge, near the northern boundary of Arizona, to Grand Wash Cliffs, near the Nevada line, a distance of about 217 miles. The deepest and most impressively beautiful section, 56 miles long, is within Grand Canyon National Park, through which the river winds for 105 miles. In its general colour, the canyon is a red, but each stratum or group of strata has a distinctive hue—buff and gray, delicate green and pink, and, in its depths, brown, slate-gray, and violet. At 8,200 feet (2,500 metres) above sea level, the North Rim is 1,200 feet higher than the South Rim.

The first sighting of the Grand Canyon is credited to the Coronado expedition of 1540 and subsequent discovery to two Spanish priests, Francisco Garcés and Silvestre Vélez de Escalante, in 1776. In the early 1800s trappers examined it, and sundry government expeditions exploring and mapping the West began to record information about the canyon. In May 1869 the explorer-scientist John Wesley Powell descended the Green River and on August 10, 1869, the party arrived at the mouth of the Little Colorado River at the northeast edge, the upstream entry point to the Grand Canyon. In 1870 reports on the geography, geology, botany, and ethnology of the area were published, and mineral exploration in the most accessible parts of the canyon was begun.

Grand Canyon National Park, containing 673,575 acres (272,586 hectares), was created in 1919. The North and South rims are connected by a paved road and by a transcanyon trail. Scenic drives and trails lead to all important features. There are about 2,000,000 visitors annually. Mule-pack trips and adventuresome scenic rides down the river in rafts and power-driven craft are intensively sought-after ways of viewing and experiencing the vast beauty of the canyon. Many pueblo and cliff-dweller ruins, with accompanying artifacts, indicate prehistoric occupation. There are five Indian tribes living on nearby reservations.

Geological history. Although the awesomeness and grandeur of its beauty are major attractions of the Grand Canyon, probably the most vital and valuable aspect of the canyon lies in the time scale of Earth history that is revealed in the exposed rocks of the canyon walls. No other place on Earth displays such an extensive and profound record of Earth events for analysis, dating, and study. Three aspects of Earth processes are superlatively, if not always clearly, presented: first, the building materials of the Earth's crust, the composition, kind, origin, and age of the exposed rocks; second, the destructive processes, the cutting of the canyon, erosion; and, third, the Earth-building processes, submergence, deposition, uplift, folding, faulting, and volcanism. Extending north from the Colorado River to Bryce Canyon, a 23,000-foot section of rocks is exposed in normal sequence from old to recent. The record, however, is far from continuous and complete. There are immense time gaps; probably several million years are unaccounted for by an unconformity in which vast quantities of Earth materials were either removed by erosion or there was little or no deposition of Earth materials. Thus rock formations of vastly different ages are separated only by a thin, distinct surface that reveals the vast unconformity in time.

The rock strata of the canyon walls are mostly limestones, freshwater shales, and cemented sandstones of windblown origin, the result of limey ooze, mud, and sand laid down in water and later hardened into rock by the great weight of the overlying layers. The crystallized, twisted, and contorted unstratified rocks of the inner gorge are granite and schist about 4,000,000,000 years old. They constitute the roots of lofty mountains, their tops eroded away. Overlying the canyon rocks are the precipitous butte remnants from the Mesozoic Era (225,000,000 to 65,000,000 years ago) and the vermilion,

white, and pink cliff terraces of southern Utah, which have been entirely eroded away in the south. Of relatively recent origin are sheets of black lava and volcanic cones covering portions of the plateau tops, some estimated to have been active within the past 1,000 years.

The cutting of the mile-deep Grand Canyon by the Colorado River is one of the greatest events of Earth history. River velocity, volume, stream gradient, and cutting tools (mud, sand, and gravel) account for the incredible cutting capacity of the river. Sediments carried by the Colorado have been measured at an average of 500,000 tons per day. The canyon actually was cut by a reverse process, for the river remained in place and cut as the land moved slowly upward against it. Only thus can be explained the canyon's east-to-west course across a south-facing slope and the presence of plateaus that stand across the river's course without having deflected it.

Inevitably, questions arise not only as to how the Grand Canyon was formed but also as to how its magnificent features, once formed, have persisted. The most significant aspect of the environment that is responsible for the canyon is frequently overlooked or not recognized. Were it not for aridity, there would be no Grand Canyon. Slope wash would have removed the canyon walls, the stair-step topography would long ago have been excavated, the distinctive sculpturing and the multicoloured rock structures could not exist, the Painted Desert would be gone, and the picturesque Monument Valley would have only a few rounded hillocks.

Biological past and present. Plant and animal fossils are abundant in the sedimentary rocks, ranging from primitive algae in the lower strata to trees in the upper strata and from seashells and trilobites to the remains of dinosaurs (both bones and footprints), camels, horses, ground sloths, and elephants. These fossils suggest the evolution of life through the ages.

Fossil and living plant and animal life

Because of the location and vast scale of the Colorado Plateau and the Grand Canyon, noticeable differences exist in soil types, temperatures, moisture distribution, slope exposure, and elevation. All of these are reflected in the plant and animal life of the canyon.

In the bottom of the canyons are the hydrophytes, plants that require abundant water during the growing season. This plant association grows along the riverbanks of the main stem Colorado and along the tributary streams on low sand and gravel bars. Willows and cottonwoods are prominent examples of the hydrophytes.

At the other end of the moisture scale are the xerophytes, the desert, drought-resistant plants with a wide diversity of adaptations. Most are perennials, widely spaced with well-developed lateral root systems. Common among the xerophytes are the yucca, agave, and numerous species of cactus.

The phreatophytes, a pestiferous group of bush and tree vegetation, grow near the streams in canyon bottoms. They have very deep root penetration, up to 50 feet, and consume immense quantities of precious water, diminishing the discharge of all streams. Mesquite, acacia, tamarix, and cottonwood are common phreatophytes.

On the canyon rims, north and south, there is a wide assortment of plant life. Typical of the South Rim is a well-developed ponderosa pine forest, with scattered stands of piñon pine and juniper. Bush vegetation consists mainly of scrub oak, mountain mahogany, and large sagebrush.

On the North Rim, in the areas of moist and deep soil are magnificent forest communities of ponderosa pine, white and Douglas fir, and aspen. Under less optimum conditions the plant life reverts to the desert varieties.

Animal life in the Grand Canyon area is both varied and abundant. In some instances it is confined to a specific area and habitat because of the extremely difficult terrain that limits movements. The common animals are the many varieties of squirrels, coyotes, foxes, deer, badgers, bobcats, rabbits, chipmunks, and kangaroo rats.

BIBLIOGRAPHY. W. KENNETH HAMBLIN and JOSEPH R. MURPHY, *Grand Canyon Perspectives* (1969), is a systematic division of the Grand Canyon into subregions and panoramas. For each panorama there is a lucid description of the rock outcrops and their colour, the canyon profile, the nature of the Colorado River, and the related plant and animal ecology. The FOUR CORNERS GEOLOGICAL SOCIETY, *Geology and Natural History of the Grand Canyon Region* (1969), the record of the Fifth Field Conference of the Powell Centennial River Expedition, accounts for, in a very comprehensive manner, the main structural features of the Grand Canyon and the major rock formations with age description and location. A brief resumé of the Powell party, its members, and the exploration of the Grand Canyon serves as the introduction to the main work.

(M.J.Lo.)

Grant, Ulysses S.

Ulysses S. Grant was a U.S. soldier and the 18th president of the United States. During the Civil War (1861–65) he commanded the Federal forces and led the Union to victory over the Confederacy. He was elected president in 1868 and served two full terms (1869–77).

By courtesy of The Library of Congress, Washington, D.C.



Grant.

Grant was born Hiram Ulysses Grant at Point Pleasant, Clermont County, Ohio, on April 27, 1822, the son of Jesse Root Grant, a tanner, and Hannah Simpson Grant. Detesting the work around his father's tannery, Ulysses performed his share of chores on farmland owned by his father, developing considerable skill in handling horses. Ulysses' boyhood appears normal for the time and place.

Early life

Jesse secured for Ulysses an appointment to the United States Military Academy at West Point, New York, in 1839. Ulysses had no interest in military life but accepted the appointment, realizing that the alternative was no further education. Ulysses decided to reverse his original name and enroll as Ulysses Hiram; his appointment to West Point was erroneously made in the name Ulysses S. Grant, the name he eventually accepted, maintaining that the middle initial did not stand for anything.

At West Point, Grant ranked 21st in a class of 39 but distinguished himself in horsemanship and showed considerable ability in mathematics. Upon graduation in 1843, he was assigned as a brevet second lieutenant to the 4th U.S. Infantry, stationed near St. Louis. There he fell in love with Julia Boggs Dent, whom he married in 1848. They had four children.

During the Mexican War (1846–48) Grant showed gallantry in campaigns under Gen. Zachary Taylor, then was transferred to Gen. Winfield Scott's army, where he first served as regimental quartermaster and commissary, which gave him invaluable knowledge of army supply but galled the young officer who wanted action, and later distinguished himself in battle, earning brevet commissions as first lieutenant and captain, though his permanent rank was first lieutenant.

On July 5, 1852, when the 4th Infantry sailed from

New York for the Pacific coast, Grant left his family behind, unwilling to endanger them on a dangerous crossing of the Isthmus of Panama. Assigned to Ft. Vancouver, Oregon (later Washington) Territory, he entered unsuccessful business ventures to supplement his army pay, but he could not reunite his family. A promotion to captain in August 1853 brought an assignment to Ft. Humboldt, California, a dreary post with an unpleasant commanding officer. On April 11, 1854, Grant resigned from the army. Allegations that he found consolation in drink during the lonely years on the Pacific coast and in later life were never proved, yet they affected his reputation nonetheless. There were many reasons for his resignation, and it was submitted by his own choice.

Grant settled on the Dent estate of White Haven, in Missouri, and began to farm 80 acres (30 hectares) given to Julia by her father. Grant's farming venture and a real estate partnership in St. Louis in 1859 were both unsuccessful. In 1860 Grant took a post in a leather goods business, owned by his father and operated by his brothers, in Galena, Illinois.

The Civil War

At the outbreak of the Civil War, Grant helped recruit, equip, and drill troops in Galena, then accompanied them to Springfield, where Gov. Richard Yates appointed him an aide and assigned him to the state adjutant general's office. Yates appointed him colonel of an unruly regiment (later named the 21st Illinois Volunteers) in June 1861. Before he had even engaged the enemy, Grant was appointed brigadier general through the influence of Elihu B. Washburne, U.S. congressman from Galena. He soon gained command of the District of Southeast Missouri, headquartered at Cairo, Illinois.

In January 1862, dissatisfied with the use of his force for defensive and diversionary purposes, Grant received permission from Gen. Henry Wager Halleck to begin an offensive campaign. On February 16 he won the first major Union victory of the war, when Fort Donelson, on the Cumberland River in Tennessee, surrendered with about 15,000 troops.

Now a major general, Grant drove off an unexpected Confederate attack on April 6–7 at Shiloh Church, near Pittsburg Landing, Tennessee, but outcry over heavy Union losses (1,754 killed) hurt Grant's reputation, and Halleck took personal command of the army. When Halleck was called to Washington as general in chief in July, Grant regained command. Before the end of the year, Grant began his advance toward Vicksburg, the last major Confederate stronghold on the Mississippi River. Grant displayed the qualities of aggressiveness, resilience, independence, and determination that led to final victory on July 4, 1863. When Port Hudson, Louisiana, the last Confederate post on the Mississippi, fell a few days later, the Confederacy was cut in half.

Command over Union armies

Grant was appointed lieutenant general in March 1864 and was given command over all the armies of the United States. His basic plan for the 1864 campaign—to immobilize Gen. Robert E. Lee near the Confederate capital at Richmond, Virginia, while Gen. William Tecumseh Sherman led the western Union army through Georgia—was successful. By mid-June Lee was pinned down at Petersburg, near Richmond, while Sherman's army cut through Georgia, and cavalry forces under Gen. Philip Sheridan destroyed railroads and supplies in Virginia. On April 2, 1865, Lee was forced to abandon his Petersburg line, and the inevitable surrender followed on April 9 at Appomattox Court House.

That Grant's army vastly outnumbered Lee's at the close of the conflict should not obscure Grant's achievements; the Union had had numerical superiority in Virginia throughout the war, yet Grant was the first to make these numbers count. Grant had rebounded from initial defeat at Shiloh; his success was due in large measure to administrative ability, receptiveness to innovation, versatility, and capacity for growth.

In late 1865 Grant toured the South at Pres. Andrew Johnson's request, was greeted with surprising friendliness, and submitted a report recommending a lenient Reconstruction policy. In 1866 Grant was appointed to the newly established rank of general of the armies of the United States. In 1867 Johnson removed Secretary of War Edwin M. Stanton in order to test the constitutionality of the Tenure of Office Act, which required the assent of Congress to removals from office, and in August he appointed Grant secretary of war ad interim. When Congress insisted upon Stanton's reinstatement, Grant resigned his secretaryship (January 1868), thus infuriating Johnson, who believed that Grant had promised to remain in office to provoke a court decision. Johnson's angry charges brought an open break and strengthened Grant's Republican ties, leading to his nomination for president in 1868. The last line of his letter of acceptance, "Let us have peace," became the Republican campaign slogan. Grant was elected with a small popular margin over his Democratic opponent, Horatio Seymour, former governor of New York.

Grant's presidency

Grant entered the White House on March 4, 1869, politically inexperienced and, at age 46, the youngest man yet elected president. His appointments to office were uneven in quality but sometimes refreshing, as when he appointed a Seneca Indian, Ely S. Parker, his former staff officer, as commissioner of Indian affairs.

On March 18 Grant signed his first law, pledging to redeem in gold the greenback currency issued during the Civil War, thus placing himself with the financial conservatives of the day. During his first term he backed the recommendations of the first Civil Service Commission but abandoned the effort in view of congressional intransigence. He was more persistent but equally unsuccessful when the Senate rejected a treaty of annexation with Santo Domingo. His negotiation of the Treaty of Washington provided for the settlement by international tribunal of American claims against England arising from the wartime activities of the British-built Confederate raider "Alabama."

Grant won re-election easily in 1872, with a large margin over Horace Greeley. During the campaign, newspapers discovered that prominent Republican politicians were involved in the Crédit Mobilier of America, a shady corporation designed to siphon profits of the Union Pacific Railroad. More scandal followed in 1875, when Secretary of the Treasury Benjamin Helm Bristow exposed the operation of a whiskey ring that had the aid of high government officials in defrauding the government of tax revenues. When the evidence touched the President's private secretary, Orville E. Babcock, Grant regretted his earlier statement, "Let no guilty man escape." Grant blundered in accepting the resignation of Secretary of War William W. Belknap, who was impeached on charges of accepting bribes; Belknap escaped conviction since he was no longer a government official. Discouraged and sickened, Grant closed his second term by assuring Congress that "Failures have been errors of judgment, not of intent."

Scandals have become the best remembered feature of the Grant administration, obscuring more positive aspects. Grant supported amnesty for Confederate leaders and protection for Negro civil rights. His veto of a bill to increase the amount of legal tender currency (1874) diminished the currency crisis during the next quarter century. He dealt gracefully with the controversy caused when both Republican Rutherford B. Hayes and Democrat Samuel Jones Tilden claimed election to the presidency in 1876.

Later life

In 1879 Grant found a faction of the Republican Party anxious to nominate him for a third term. Although he did nothing to encourage support, he received more than 300 votes in each of the 36 ballots of the 1880 convention, which finally nominated James A. Garfield. In 1881 Grant bought a house in New York City and began to take an interest in the investment firm of Grant and Ward, in which his son Ulysses Jr. was a partner. Grant put his capital at the disposal of the firm and encouraged others to follow. In 1884 the firm collapsed, swindled by Ferdinand Ward. This impoverished the entire Grant family and clouded the General's reputation.

In 1884 Grant began to write reminiscences of his campaigns for the *Century Magazine* and found this so congenial that he began to prepare his memoirs. Despite ex-

cruciating throat pain later diagnosed as cancer, he signed a contract with his friend Mark Twain to publish the memoirs and resolved grimly to complete them before he died. In June 1885 the Grant family moved to a cottage at Mount McGregor in the Adirondacks, and there Grant died on July 23. He was buried in Riverside Park, New York City, where an elaborate tomb was dedicated in 1897.

Grant completed his memoirs shortly before his death. Written with modesty and restraint, exhibiting equanimity, candour, and a surprisingly good sense of humour, they retain high rank among military autobiographies.

BIBLIOGRAPHY. An excellent brief biography is BRUCE CATTON, *U.S. Grant and the American Military Tradition* (1954). More detailed biography is available in LLOYD LEWIS, *Captain Sam Grant* (1950); BRUCE CATTON, *Grant Moves South* (1960), *Grant Takes Command* (1969); and WILLIAM B. HESSELTINE, *Ulysses S. Grant, Politician* (1935). Analyses of Grant as a soldier include J.F.C. FULLER, *The Generalship of Ulysses S. Grant* (1929); T. HARRY WILLIAMS, *Lincoln and His Generals* (1952); and KENNETH P. WILLIAMS, *Lincoln Finds a General*, vol. 3–5 (1952–59). *The Personal Memoirs of U.S. Grant*, 2 vol. (1885–86), are also available in a one-volume edition (1952). Grant's correspondence is in JOHN Y. SIMON (ed.), *The Papers of Ulysses S. Grant*, 4 vol. (1967–72).

(J.Y.S.)

Grasslands

The natural vegetation of the earth can be conveniently zoned on a combined physical and environmental basis into grassland, woodland, tundra, and desert. Each region contains distinctive plants, animals, climate, and soil. The proportion of land surface originally occupied by each region is difficult to assess because of gradual transitions from one to another and because of the modifying influence of man in altering the biotic and environmental characteristics of the landscape. Evidence suggests that, before man began the present rapid modification of environments by extensive agricultural and industrial operations, between 40 and 45 percent of the land surface of the earth was occupied by grasslands, in which grasses and grasslike plants dominated in the absence of trees or with widely spaced trees. Although tracts of grasslands and savannas still survive in a modified state as rangelands for domesticated livestock and game animals, much has also been exposed to such intensive agriculture that the original plants and animals have been virtually eliminated.

Original extent of grassland

This article is concerned with natural grasslands other than those occurring at high latitudes north of the circumpolar coniferous forest in the Northern Hemisphere and above forests on mountains. Likewise, seminatural grasslands of forest areas denuded by man are not considered, since they relate more properly to special situations within the woodland biome—*i.e.*, the worldwide complex of woodlands characterized by prevailing climate and soil conditions. Seeded grasslands are agricultural systems on arable land and will be treated here only in relation to ecological problems that arise when natural grassland flora is replaced by other species.

THE ENVIRONMENTAL SETTING

Regions of natural grassland occur where the environment is too arid for the development of closed forest but not so adverse as to prevent smaller, nonwoody but long-lived plants from forming a dense layer. Climate controls the biotic components of a region directly, through temperature and moisture extremes, as well as indirectly, through its influence on soil development by such long-term physical and chemical processes as freezing and leaching. Within the grassland zone, however, local variations in topography and soils result in microclimatic conditions so different from the typical regional norms that nongrassland ecosystems dominated by plants and animals not characteristic of the remainder of the region may develop locally. In certain warm regions, for example, the woody habit is exhibited by many plant species even under arid conditions, so that specially adapted trees are scattered throughout the grassland to form a characteristic parklike vegetation type known as a savanna.

Climate. The subhumid to semi-arid climates of natural grassland areas are characterized by marked periodicity of precipitation, both from season to season within the year and between years. Consequently, annual drought periods of several weeks to several months are typical. The severity of drought increases with the distance from the forest margin. It is also accentuated by a cyclical climatic pattern that often results in several consecutive dry years. Droughts are particularly difficult for plants and animals in regions where the dry season is lengthy—as in the Mediterranean-type climate of some temperate grasslands and in most tropical and subtropical grasslands and savannas—or where a series of moister than average years is followed by several years of below-average moisture (see SCRUBLANDS).

Annual precipitation in grasslands

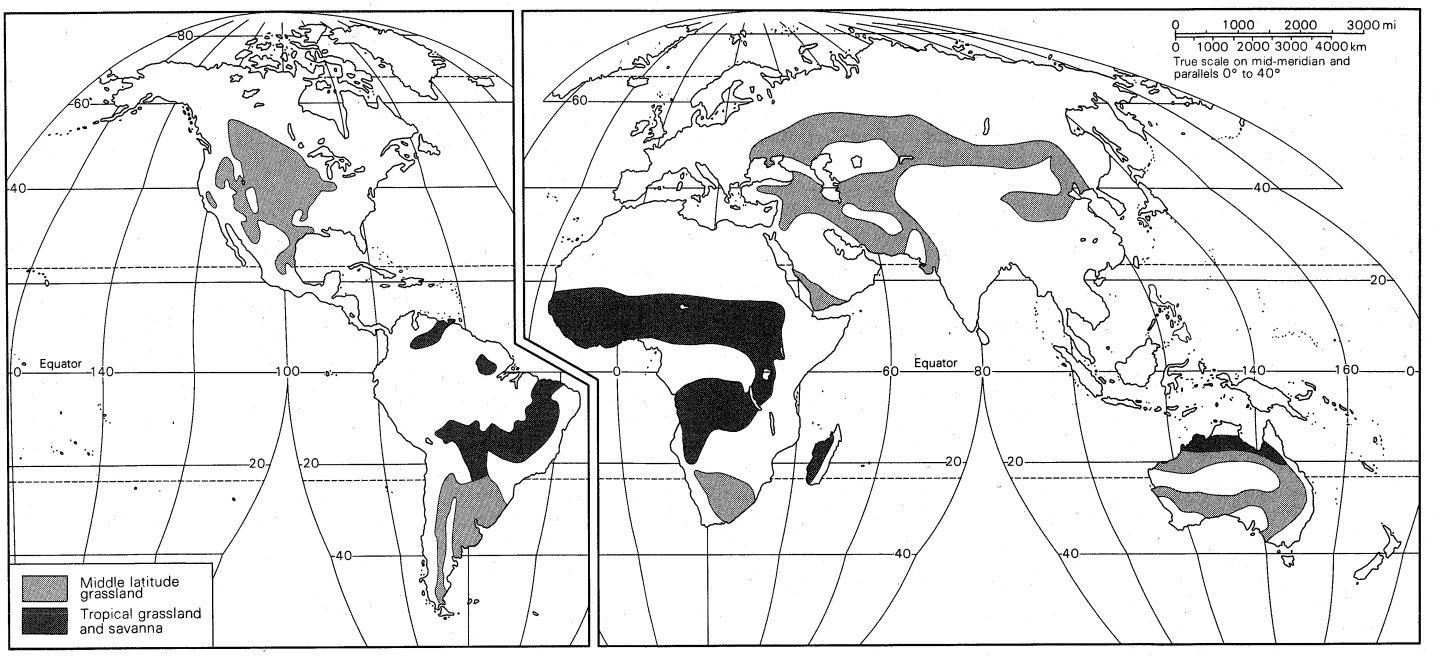
Mean yearly precipitation in most temperate grasslands ranges from ten to 30 inches (250 to 750 millimetres) and in tropical and subtropical grasslands and savannas from 25 to 60 inches (600 to 1,500 millimetres). Precipitation determines the nature and extent of natural grasslands through its effect on soil-moisture supply, since trees have difficulty competing with grasses in areas where upper soil layers are moist during part of the year but where deeper layers are continuously dry.

The adverse effect of drought on organisms is made more severe when accompanied by high temperatures. Increased duration and intensity of sunlight associated with decreasing cloudiness during periods of low rainfall tend to raise air and soil temperatures. In regions where the dry season is warm, the combined effects of drought and high temperatures make living conditions particularly adverse. Many organisms adapt to the hot dry season in grasslands of tropical, subtropical, and Mediterranean climates and to the cold season (also a time of low precipitation) of many temperate regions by becoming dormant.

In the grasslands of tropics and subtropics, the length of the growing season is determined by the length of the rainy season, ranging from 120 to 190 days. In cold, temperate regions, however, temperature is a controlling factor. Growth begins in such regions when the mean daily temperatures reach 40° to 50° F (5° to 10° C). Although the frost-free season in cold, temperate grasslands may be as short as 100 continuous days, the vegetative season usually exceeds 165 days. Less temperature difference occurs among various grasslands during the growing season than during the dormant season. The temperatures during the warmest month in temperate grasslands are usually not lower than 60° to 70° F (15° to 20° C); in tropical and subtropical regions, the mean during the rainy season is commonly no more than 20° F (10° C) higher.

Grassland regions tend to have higher than average wind speeds, in part because of low wind resistance due to the relatively smooth vegetative layer.

Fire. Fire is an important feature of the grassland environment. Lightning-induced fires are common and, before their control by man, periodically burned over extensive areas. The causes of fire are so closely related to the more or less arid nature of the grassland climate that it is not reasonable to separate fire from the more traditional climatic factors. Some geographers, however, have not been prepared to accept fires as a climatic factor and consequently have considered the natural vegetation to be that which would exist in the absence of fire. It seems more practical, in regions where aridity, continuity of combustible plant cover, and frequent thunderstorms lead to the frequent occurrence of natural fires, to consider the vegetation as it has been modified by fire to be the stable one. Fire maintains the boundaries of grasslands in climatic regions (usually the wetter portions of the grassland zone) capable of supporting forest growth. The effect of fire here, therefore, has been principally to prevent the extension of the forest into the grassland. The nature of the soil in such transition zones between grassland and forest is such, moreover, that the long-term existence of grassland is indicated.



World distribution of grassland biomes.
Adapted from *Biological Sciences Curriculum Study Green Version High School Biology*, 2nd ed.; Chicago: Rand McNally & Co., 1968

In the temperate grasslands, a climate adverse to tree growth exists except near the forest edge, but, in the tropics and subtropics, trees and shrubs are often scattered throughout the grassland to form a savanna. The luxuriant growth of herbaceous (nonwoody) plants during the rainy season provides fuel, which, in the dry season, supports fires almost annually in many regions. Whether or not a forest would develop in the absence of fires in these tropical areas is speculative. It is most satisfactory to consider them as savannas in which the herbaceous layer predominates.

Fire effects

Fire is more destructive to trees and shrubs than to grasses and other herbs, because the buds, from which new growth will begin the next season, are located well up on the stems in a position quite vulnerable to fire. In herbaceous species, the growth zones are at or below the soil surface.

Topography. Many grassland regions are characterized by level to gently undulating topography. Often, however, the land surface is sufficiently irregular to provide variations in microclimate that result in a mosaic of vegetation types, as, for example, in the glacial moraines and loessial (loess is a soil material carried to its present location by winds) hills in parts of the Central Lowland and Great Plains of North America and in sandhills everywhere. Near the forest–grassland margin, grassland is favoured, even within the forest zone, on dry slopes that are subjected to greater radiation and wind than is level terrain. Conversely, patches of trees occur within the grassland zone on protected slopes.

Contrasts in vegetation on slopes with different characteristics are particularly noticeable in the higher latitudes, where the slopes facing the equator regularly receive more radiation than do poleward-facing ones. Various habitats also are found in sheltered basins and on exposed hilltops. Forest occurs within the grassland zone in protected basins, and patches of grassland occur within the forest zone on exposed hilltops.

Within the grassland zone, the herbaceous vegetation of depressions and poleward-facing slopes often consists of species that dominate on level ground in poleward areas. In the same regions, the species in exposed habitats are those that dominate communities that range toward the Equator. Similarly, where grassland gives way to desert, desert species appear in exposed locations.

Soil. In regions where natural grasslands have developed, physical and biotic components have combined to produce soil quite different in its character from that of forest regions. In the temperate grassland, soil leaching (downward movement of water-soluble soil constituents) is restricted both by the scarcity of percolating water and by the low solubility of soil materials. The soil of natural grasslands, consequently, contains an abundance of organic matter at all depths; and chemical substances that are leached precipitate at the bottom of the soil profile (*i.e.*, the layering of soil as seen in a trench). Consequently, temperate grassland soils have the capacity to release nutrient elements to plants slowly and over a long period of time, thus permitting the production of annual crops, even after destruction of the natural plant cover. The abundance of organic matter (humus) in grassland soils is reflected in the colour of the soil profile; in temperate grassland, the colour varies from black near the forest margin, where the organic content approaches 10 percent in some areas, to chestnut and brown in areas where the organic content is below 3 percent.

Colour of grassland soils

In the tropical and subtropical grasslands, soil development is influenced by high temperatures and high precipitation in a process called laterization. By this process, soils become highly leached, and there is rapid decay of organic matter with low levels of humus accumulation. These soils are often reddish or yellowish in colour, reflecting the high content of iron left in the profile by the leaching of other minerals under conditions of high temperature. They are also very much more subject to nutrient depletion under agricultural practice than are temperate grassland soils.

Differences in soil texture and soil structure throughout the grassland zones modify the vegetative cover because of influence on water regimes. Although sandy soils hold less moisture per unit volume, they permit more rapid percolation of surface moisture than do soils of finer texture, with a lower resultant loss by runoff and evaporation. Because of their greater moisture supply and the availability of moisture to greater depths, sands support stands of tall grasses or even trees in regions where grasses of smaller stature and less depth of rooting occur on soils of finer texture.

THE BIOTIC COMPONENTS

The living members of an ecosystem are most conveniently considered as the floral (plants), faunal (animals), and micro-organism elements, which, respectively, function in a general way as producers of organic matter, consumers, and decomposers.

Flora. The vegetation of natural grasslands is primarily composed of seed-bearing herbaceous plant species, which are classified, on the basis of gross morphology, in-

to two groups: (1) grasses and grasslike plants (particularly sedges), which are collectively referred to as graminoids; and (2) nongrasslike herbs, mostly broad-leaved species, known as forbs. Woody components also occur, which, unlike the herbs, maintain perennial stems above the soil surface—*i.e.*, they live through several growing seasons. Frequently, there are dwarf shrubs that do not exceed the stature of the grasses; taller shrubs occur in some areas as isolated individuals or in patches. Scattered trees characterize the savannas, and, in some grassland areas, groves of trees are found in locally modified habitats such as along stream courses. Occasionally, colonies of non-seed-bearing plants, particularly lichens, club mosses, mosses, and algae, are found on the soil surface.

A considerable number of species comprise the vegetation of grasslands in areas of favourable moisture and temperature conditions, but the number decreases with increasing environmental adversity. In a square mile (about 2½ square kilometres) of rolling grassland in the black-soil region of eastern Nebraska, for example, over 200 different prairie species are found; in a similar area of level brown soil 1,000 miles (1,600 kilometres) to the northwest in southern Saskatchewan, the flora of seed-bearing plants is about 50 species. Similarly, the number of species in tropical savannas is limited to those that can resist the combined effects of fire and drought.

Success of grasses in occupying habitats

The graminoids (grasses and grasslike plants) are particularly well adapted to dominate in herbaceous communities, so that, although they typically make up less than 20 percent of the number of species present in grassland, they often furnish 90 percent or more of the plant biomass (dry weight) present. Similarly, though only a small percentage of the seed-bearing plant species belong to the grass family, their contribution to plant communities is great everywhere.

Grasses evolved from tropical woody ancestors, of which the bamboos are a present-day example. This origin is reflected in the relatively primitive nature of grasses composing the grasslands adjacent to forests in both tropical and temperate regions. Their adoption of the herbaceous habit, however, has permitted survival through such adversities as periodic burning and seasonal drought. Other adaptations that increase the chances of seed production in arid, windy, and cold environments include wind pollination, increased protection of flowers, and a reduced stature.

The graminoid life-form seems particularly well adapted to dominate in competition with forbs. Success of grasses seems to be associated with their ability to provide a dense plant cover in which seedlings of less aggressive species have difficulty in becoming established. Graminoids are well adapted to exist in conditions of frequent fire and in areas of animal grazing by their manner of growth. New growth is from a zone of cell division located at the bases of leaves and stems rather than at the tips, as in forbs and woody plants. This permits continued growth when the upper ends of shoots or the leaves are cut, eaten, or burned. The growth zones in the stem also assist shoots that have been bent by trampling to regain a vertical position.

Classification of grasses by height

Grasses and the grasslands they dominate are sometimes classified as high, tall, mid, and short. High grasses, frequently reaching heights of ten feet (three metres) or more, are limited to tropical grasslands with high precipitation. Tall grasses are found in the tropical and subtropical savannas and also in the most favoured parts of the temperate grasslands, moist habitats being particularly suited to them. Midgrasses, the flowering stems of which usually reach heights ranging from one to three feet (30 to 90 centimetres), dominate portions of the grassland closer to the woodlands of temperate regions. The short grasses, with leafy shoots often only three to six inches (eight to 15 centimetres) high, are dominant in the most arid grasslands. Between these extremes, the short and mixed grasses intermingle to form mixed-grass prairie, as in most of the Great Plains of North America.

The adaptation of each species of grass is also dependent on whether it is a turf former or a bunch grass. Turf-forming grasses spread by means of buds on lateral stems below the soil surface (rhizomes) or, much more rarely, above it (stolons). Bunch grasses are able to spread horizontally only far enough to extend the size of the dense bunch, which eventually declines with age. Occasional development of seeds is necessary for the perpetuation of a bunch-grass cover but not of turf grasses. Bunch grasses tend to dominate in arid habitats and turf formers in moist ones.

Grass species also differ physiologically in ways that adapt them to withstand adversity. Some species can continue growth during drought by extracting a greater amount of moisture from the soil. Others endure drought by achieving a dormant state before desiccation becomes excessive. Under extreme conditions, drought or overwinter survival is achieved only by means of seeds, in which case, the plants are known as annuals. Annuals are common in disturbed habitats in grassland, but they are more characteristic of the temporary herbaceous cover of deserts.

Root systems of grasses

The root systems of grasses are typically finely branched and rebranched to form a dense network extending through the soil. Depth of rooting of tall and midgrasses is characteristically not greater than six feet (about two metres) and those of short grasses only one to three feet. These distances are also the maximum depths of moisture penetration in the soils of the habitats in which they dominate.

In some grasslands of both temperate and tropical regions, sedges (*Carex*) are associated with the grasses and have similar ecological relationships. The composites (*e.g.*, asters, sunflowers) commonly rank next to grasses and sedges in abundance. Under the favourable growing conditions near the forest boundary, forbs are abundant; the adverse conditions in the drier grassland regions, however, reduce their abundance, relative to grasses, in number of both species and individuals. Many forbs are adapted to arid conditions by rooting to depths much greater than the grasses, by surviving as dormant underground organs for periods as long as many years and by storing water in their tissues. The root systems of some forbs are similar in structure to those of grasses, but a variety of other types exist, including taproots, bulbs, corms (rootlike or bulblike stem structures), and roots capable of forming shoot-producing buds. Species with the last type are particularly vigorous competitors with grasses and are especially abundant in the grasslands of the U.S.S.R., where in some areas forbs are considered to dominate over grasses. Eurasian forbs with such root systems have become the most persistent introduced weeds of arable land in many parts of the world. Among them are perennial species of *Euphorbia* (spurge), *Cardaria* (white top), *Centaurea* (thistle), and *Convolvulus* (morning glory).

The complexity of the vegetative cover of natural grassland is much greater than it appears to the casual observer. The many plants occupying grasslands may exhibit as much variety within a square yard as is present in an acre of forest.

The dominant and associated species are arranged horizontally into colonies and societies and vertically into layers, both above and below the ground. The number of layers discernible depends on the nature of the grassland. The uppermost one is composed of the tallest grasses and forbs. Some forbs have vegetative growth confined to lower layers but thrust their flowering parts, often on leafless stems, upward above the grass layer to insure pollination and seed dispersal. Others grow vegetatively at a rate similar to that of the taller grasses and bear their flowers on leafy stems. A lower layer is made up of species of shorter stature, some of which are secondary species that grow early in the season, before the dominants cover them.

The lowest layer above the ground is the crust of lichens, mosses, club mosses, and algae that often occurs on the soil surface among the fallen stems and leaves (litter) of the upper layers. Various groups of plants tend to root to different depths, so that short grasses and small sedges form a shallow layer underground and taller grasses a deeper layer. The roots of some forb species often branch

only below the deepest layer of grass roots; they sometimes penetrate to depths twice that of the dominants.

Seasonal aspects of grasslands

The appearance of temperate grasslands changes from season to season as a result of the successive flowering of different species. A few species thrust their flowering stems upward shortly after the disappearance of snow and before they are shaded by the spring growth of the cool-season grasses. These are followed by the spring flowers, which, in the temperate grasslands, are composed of a great variety of forbs including early legumes (members of the pea family). Early grasses and some sedges begin to flower near the end of this period. Most of the dominant grasses, of both warm and cool seasons, blossom during the summer, as do most of the legumes, all of which give a predominantly bluish and whitish tinge to the landscape at this time. The predominant colour of autumn flowers is yellow, and the floral display is made up principally of an abundance of composites. During these shifts in the seasons, the grasses also change in the early spring from the grayish colour of the preceding year's old growth through rich greens as the new leaves emerge to a straw-coloured hue of pale yellow in the late summer and early fall. Curing of shoots to a yellowish colour is characteristic of the more arid grasslands; those in more humid areas turn gray more rapidly. The growing season varies in character from year to year as differences in weather differentially stimulate or retard flowering of the various species. The grassland tends to be very drab and relatively lifeless during the winter. The changes in the season have not been as well documented in the tropical grasslands, in which the main changes apparently take place with the shift of the season from dry to rainy, with the resultant prompt resumption of luxuriant green growth that ends only with the renewed onset of drought. Some tree and forb species in tropical savannas anticipate the change in the season by leafing out before the arrival of the rainy season.

Animal life. The natural animal populations in grasslands are much more diverse than is generally realized, as many surface species are small and inconspicuous, and many other species live underground for at least part of their lives. The best known are the large grazing mammals, or game animals. Other well-known groups in grasslands include grazing marsupials (pouched animals, such as the kangaroo), predators belonging to the cat and dog families, rodents of various sizes, birds, lizards and snakes, and the larger insects, particularly grasshoppers and locusts. A large proportion of surface species are either running or burrowing types. The characteristic aggregation into colonies or herds provides some measure of protection in the open type of habitat. Invertebrates of various kinds are abundant in the soil.

Vertebrates. The large grazing mammals, subject to the influence of man have been severely reduced, and some species survive only in protected areas. The former ranges of these animals have become densely populated and are being grazed by cattle, sheep, horses, and goats. In some of the managed reserves in Africa, the natural and domesticated populations intermingle, but for the most part, the native grazing species are considered intruders, and their predators also are excluded by poisons, hunting, or fences.

Large animals of grasslands

The African grasslands were once abundantly supplied with a large number of species, of which perhaps the wildebeest, gazelles, and zebra were the most numerous over large areas. Other species included buffalo, giraffe, eland, antelope, hartebeest, hippopotamus, waterbuck, and impala. Some species, which are browsers as well as grazers, have had a great influence in restricting the growth of trees. These herbivores (plant-eating animals) are accompanied by a variety of omnivores (animals, such as the warthog, that feed on both animal and vegetable substances) and carnivores (meat eaters), particularly the lion, leopard, hunting dog, and fox. The hyena is perhaps more truly a scavenger, but it also fills the role of a predator while travelling in packs. The characteristic grazing species of the Asian grasslands include wild cattle, saiga antelope, wild horse, marmot, stag, and boar; predators are represented by the wolf and the fox. Wild horses and cattle, formerly abundant, have been exterminated by man, and antelope and marmot have largely been forced out of the grassland region.

The number of species of large grazing mammals in the Western Hemisphere, while it was larger during the early geological stages of development of grasslands, has been low in recent geological time. In North America, the principal species once were the bison and the pronghorn antelope. Their most important predators were the coyote and bobcat. In South America, the large grazing mammals have been less influential; rather, the llama and its relatives and the ostrich are the most important. The isolated fauna of Australia includes several kangaroo species as the major large grazing animals, with the dingo as their prime predator.

In some grasslands, the large grazing mammals are as important as climate and soil conditions in determining the nature of the vegetative community, since some plant species are more sensitive to grazing pressure than others.

Smaller herbivorous mammals, common in natural grasslands, include many species of rodents (*e.g.*, mice, voles, shrews, ground squirrels, gophers, prairie dogs, hares, and rabbits). Some species have increased their populations in regions where man has reduced the numbers of their predators, others (*e.g.*, the prairie dog in the Great Plains of North America) have suffered from a direct effort by man to exterminate them. Rodents, now reported to be the most frequently found mammals in some grasslands, are credited with causing widespread range deterioration, because they modify the speciation of the plant cover and expose the soil surface to erosion.

A wide variety of birds, including herbivorous, carnivorous, and omnivorous species, inhabit grasslands. The herbivores are the least numerous in terms of different species, and the omnivores are most abundant. The passerines (perching birds) are particularly common and include larks, longspurs, meadowlarks, starlings, grouse, cranes, partridges, and doves. Hawks, owls, and eagles are important predators.

The amphibians and reptiles are important organisms in many grasslands. Lizards, toads, and box turtles are mainly functional as predators of insects; snakes are predators of rodents and some other small vertebrates. These organisms, probably the least conspicuous components of grasslands, are often, however, among the most abundant vertebrates.

Invertebrates. Great numbers of species of insects and other invertebrates inhabit grasslands, forming large populations of individuals. Most conspicuous are the grasshoppers and their relatives. The numbers of surface insects present in grasslands have been estimated to be as many as 1,000 per square metre (about 100 per square foot), and the number of species present in one type of grassland may exceed 200. Insect bulk, or biomass, has been reported to be greater in some areas than that of large grazing mammals. Next to the grasshoppers (Orthoptera), the insects with the most pronounced effect on herbage are probably those of the orders Heteroptera (bugs) and Homoptera (aphids and leafhoppers). Ants (Hymenoptera) and termites (Isoptera) are found particularly in tropical and subtropical grasslands and savannas. The larvae of beetles (Coleoptera) feed on roots of plants and on feces, especially those of large grazing animals. Larvae of flies (Diptera) often affect seed production and are active in decomposing carrion. Larvae of moths and butterflies (Lepidoptera) feed on the crowns of grasses; and thrips (Thysanoptera) also affect seed production. The most abundant predatory invertebrates above the soil surface are spiders.

Insect numbers and bulk

Small soil animals are important to the grassland ecosystem. Chief among them are nematodes (roundworms), collembolans (springtails), mites, and enchytraeids (small segmented worms), as well as insects. Earthworms, although common in semipermanent and seeded grasslands, are rare in natural grassland conditions.

Micro-organisms. Microflora (bacteria, actinomycetes, fungi, and algae) probably are found in greater numbers in grasslands than are microfauna (protozoans). The great majority of micro-organisms present in the soil are

apparently dormant or inactive at any one time, although the bacterial biomass to a depth of 15 centimetres (six inches) in agricultural soil has been estimated to range from 330 to 720 kilograms per hectare (290 to 640 pounds per acre). Grassland soils, because they have considerable root systems, which provide habitats for markedly higher numbers of bacteria, thus contain an even greater bacterial biomass. Fungi, although present in appreciably lower numbers than bacteria, have much larger cell sizes and are estimated to possess a biomass twice that of the bacteria. Algae are commonly estimated to contribute less than 10 percent of the total microfloral biomass in mature soils.

Bacteria and fungi

COMMUNITY DEVELOPMENT

Stages. Grassland ecosystems have developed through an orderly series of changes in plant cover and soil, as one has affected the other through climatic influences. These progressive changes in plant cover are known as primary plant succession. As plant succession progresses toward a stabilized plant community, a soil gradually develops a profile that reflects the nature of the plants occupying it. The end of this process, a climax situation in which the organisms and soil are in equilibrium with the climate, is a stable system with many buffers to protect it from climatic changes and animal populations.

Primary succession of grasslands begins on bare, dry land surfaces or in water. In each situation, development is toward a more moderate water regime. This is accomplished in the dry habitat by the development of an insulating plant cover that protects the surface from the desiccating effects of sun and wind and by the incorporation of organic matter into the soil, which increases its water-holding capacity. In wet habitats, the process involves gradual elevation of the substrate (bottom of the water body in question) by deposition of both organic and mineral particles.

The series of changes in plant cover resulting in grassland begins with a layer of crustose lichens, followed by foliose (leafy growth form) lichens. Eventually, annual herbs and a series of perennial species—first forbs and short-lived grasses, finally a stable community of long-lived grasses and forbs—becomes established.

Gradual succession of water bodies to grassland

The succession beginning in a body of water starts with submerged aquatic species and progresses through floating-leaf, reed-swamp, and sedge-meadow stages as the water level decreases. This moist habitat may survive for thousands of years and permit the development of various vegetative stages (even of shrubs and trees) before the depression containing the water gradually fills from the edges and the vegetation becomes dependent, for the first time, on climatic factors (chiefly, precipitation) that permit the establishment of a climax grassland cover.

The uniformity of the climax vegetative cover depends on the degree of irregularity of the landscape, for the final stages of the dry-habitat succession will not be reached while exposed locations survive on sunny and windy slopes and on hilltops, and protected situations are preserved on sheltered slopes and in depressions. The regional characteristics of the vegetative cover, however, are expressed in topographic locations between these extremes. It is on this portion of the landscape, which in most grassland areas constitutes the major area, that the following brief survey of the grasslands of the earth is based. While this survey considers the temperate regions separately from the tropical and subtropical grasslands, the distinction is very arbitrary. Subtropical grasslands give way imperceptibly to temperate communities through the arid areas where contact usually takes place.

Temperate grasslands. The most extensive grassland areas of the temperate zone are the steppes of Eurasia, the prairies and plains of central and western North America, and the pampa and adjacent areas of Argentina. Less extensive areas are found in the velds of South Africa, in the mountain grasslands of South America, and in Australia and New Zealand.

North America. The North American grassland is composed of seven regional types. The largest, the mixed prairie, extending from southern Alberta and Saskatchewan southward to western Texas, lies between the 100th meridian (west longitude) and the foothills of the Rocky Mountains to the west. It is dominated in the north by bunch-forming spear grasses (*Stipa*) and turf-forming wheat grasses (*Agropyron*), which are midgrasses that gradually give way to short grasses southward, particularly blue grama (*Bouteloua*) and buffalo grass (*Buchloë*). The next largest area of grassland east of the Rocky Mountains occurs in a higher precipitation belt extending from the mixed prairie to the eastern deciduous forest. Although it is now the "corn belt," this area was formerly dominated by spear grass and dropseed (*Sporobolus*) in the drier habitats and by bluestems (*Andropogon*) in more favourable locations. Both types occur in areas east of the Rocky Mountains, where there is an early-summer peak in precipitation, followed by late-summer drought. West of the Rockies the most northerly grassland area is the palouse prairie, which occurs in eastern Washington and adjacent parts of Idaho and Oregon, and in the valleys of British Columbia. The dominant grass species is a large, bunch-forming wheat grass. The upper slopes in the northern part of this region are occupied successively by remnants of mixed prairie to the east and then by fescue prairie. The latter also occurs as a belt around the mixed prairie's northern edge in the eastern foothills of the Rocky Mountains and in parklands of Canada's "prairie provinces." Southward, in the Mediterranean climate of the valleys of California, the native, perennial grass cover has been replaced—through overgrazing—by annual weedy grasses, but spear grasses formerly dominated. Southwestern United States and north central Mexico contain the most arid prairies, the desert plains grasslands, which surround the warm desert at elevations between 1,000 and 5,000 feet (300 and 1,500 metres). Short grama (*Bouteloua*) and wire grasses (*Aristida*) abound here in an environment similar to that of the dry subtropics. The region of Texas adjacent to the Gulf of Mexico also has subtropical vegetation characteristics.

South America. The best known temperate grassland in South America is the pampa of east central Argentina, which occupies a region of level, black soil. The annual precipitation of 40 to 50 inches (100 to 125 centimetres) is the highest of the temperate grasslands, and frequent drought periods do not occur. Similar grassland extends on more rolling topography to the north through Uruguay into the semi-arid campos region of southern Brazil and northwestward in Argentina. Westward, the pampa becomes drier and the soils gradually change to chestnut and brown and then to gray in colour near the Andes mountains. In the driest areas, shrubs and small trees (*Prosopis, Cassia*) form a dwarf savanna. Southward, the pampa gives way to the semi-arid regions of Patagonia. Only part of these areas have grassland character. The extremely dry regions contain shrubby, cold desert vegetation. The pampa and the adjacent less humid grasslands have been exposed to grazing by domesticated livestock for about 400 years. Heavy grazing, particularly in the last century, has modified the vegetation to the extent that the identity of the natural dominant species is not known, although spear-grass species are thought to have been included among them. The vegetation is now composed principally of introduced species. Cool mountain grasslands are extensive at various altitudes in the Andes, occurring at high elevations even in the tropics. These areas are dominated by species of spear grass, fescue, bluegrass (*Poa*), and reed grass (*Calamagrostis*).

Eurasia. The body of natural grasslands in Eurasia lies within the U.S.S.R., extending from north of the Black Sea in the southern Ukraine eastward through northern Kazakhstan and southern Siberia. These steppes are bordered on the north by forest steppe where the elements of the boreal forest intermingle with the grassland, as is the case in the similar climate at the northern edge of the Great Plains grassland of North America. Southward, the climate becomes warmer and drier, particularly in the Asiatic portion, and the soils gradually lighten in colour from black to dark brown, brown, and gray at the edge of the cold-shrub desert region. The most important dominants are species of spear grass, feather grass (also

Stipa), and fescue. Here, the species of fescue that dominates over large areas is more drought-resistant than the dominant spear grasses and feather grasses, a situation the reverse of that in the northern American grasslands.

Africa and Australia. The South African veld, the tussock grasslands of New Zealand, and the grasslands of Australia are much less extensive. They are similar mainly in that oat grasses (*Danthonia*) are dominants in each. In Australia, however, the *Danthonia*-dominated grasslands are restricted to the southeast, with more extensive grasslands dominated by Mitchell grasses (*Astrebla*), which occur in arid situations blending into subtropics. Grassland of temperate character also occupies the highlands of East Africa.

Tropical and subtropical grasslands and savannas. Tropical and subtropical grasslands and savannas are located in central Africa, central South America, and northern Australia.

Africa. In Africa, they occur northward and southward from the tropical rain forest. The tropical high-grass savanna forms a belt up to 400 miles (650 kilometres) wide adjacent to the forest. This savanna is dominated by many grasses of the tribes Andropogoneae and Paniceae, of which elephant grass (*Pennisetum*) and cogon grass (*Imperata*) are probably the most abundant. The dominants reach five to 15 feet (1.5 to 4.5 metres) in height. Even though there is a pronounced dry season, the stability of this vegetation is considered to be more dependent on fire than on climate. The high-grass savannas give way, northward and southward, to tropical and subtropical tall-bunchgrass savannas and then to short-bunchgrass savannas, with decreasing precipitation and increasing length of dry season. All of the widespread dominants of this region are of the tribe Andropogoneae, the most abundant species being themeda grass (*Themeda*). The tall-grass areas stand three to five feet (one to 1.5 metres) high, while, in the short-grass savannas, dominants grow to about one foot (30 centimetres). The trees in these savannas are scattered or in clumps and reach heights of ten to 50 feet (three to 15 metres). In the drier areas they are shorter and more thorny than elsewhere. As the degree of aridity increases further, in areas where the annual rainfall averages less than 20 inches (50 centimetres), the savanna is replaced by semidesert and desert grassland in which important species include spear grasses and wire grasses and in which the woody element is limited to low, thorny shrubs.

Fire-maintained high-grass savanna

South America. The most extensive tropical grassland area in South America is the llanos of Venezuela, which lies between the Andes Mountains and the Orinoco River. This region is composed of a complex of plant communities including grassland and savanna, with high grasses dominating in some parts and tall grasses in others. Over large areas there are no trees, apparently because of waterlogging of soil. Less extensive savannas are found, interspersed with forest, in many other parts of tropical America, especially the Orinoco plains and lowland along the Gulf of Mexico in Colombia, parts of the Amazon River Basin, parts of Bolivia, the Pacific coast of Costa Rica, and parts of the West Indies. The best known dominants in many of these areas are Para grass and guinea grass (both of the genus *Panicum*). The semi-arid region of northeastern Brazil is similar to those on the equator side of the deserts in Africa, with wire grasses among the dominants and also with scattered shrubs.

Australia. The "savanna woodlands" of tropical Australia include many areas in which the tree cover (mostly *Eucalyptus*) is sufficiently thin to permit grasses to dominate. Here, many high and tall grasses develop during the rainy season, but (as elsewhere in the subhumid tropics) the nutrient quality of herbage in the dry season is very low. Similar types of grassland cover occur in Java, Sumatra, Ceylon, and the Philippines in places where the forest has been destroyed or thinned by various means.

Lands surrounding the Mediterranean Sea, those in Asia Minor, and parts of southwestern Asia (Iran, Pakistan, Afghanistan, Tadzhikistan, and India) have been exposed so long to the effects of man's occupation that it is no longer possible to determine their original character. It is probable that many areas that are presently desert-like were previously grassland. Elsewhere, the change from grassland to desert, because of man's occupancy, has been observed during the last century.

FUNCTIONING AND PRODUCTIVITY OF GRASSLAND ECOSYSTEMS

The organic component of a grassland ecosystem is a complex of producers, consumers, and decomposers that are highly organized into a food web. The system is driven by energy fixed from sunlight during the process of photosynthesis in the green tissues of plants (the producers). This energy is passed on in the form of organic matter—used as food—from one to the other of various groups of organisms, each of which liberates some of the energy in its own body functions. Finally, the photosynthetically produced organic matter is transformed back into mineral elements, which may again be used by plants in photosynthesis.

Biological productivity

The biological productivity of such a system is expressed in terms of the rate of fixation of solar energy by the producers and is known as primary productivity. The rate at which the consumer biomass increases, by transformation of plant materials into animal tissue, is referred to as secondary productivity. Primary productivity for natural, temperate grasslands ranges mostly between 200 and 1,000 grams per square metre of dry matter per year; that in the most favourable tropical areas is much greater. Secondary productivity is considerably less than primary productivity in all regions. In the central Great Plains of North America, for example, yearling cattle consume only a portion of the vegetation present in their grazing area, and of this it is estimated that only 9 percent is used in meat production, 48 percent being lost as heat to the atmosphere and 43 percent to other food chains in the form of feces and urine.

Plant species in natural grasslands are sensitive to the removal of more than a moderate proportion of their shoots. Unconsumed shoots die and remain standing, in some communities for periods as long as two or three years, during which time they protect the soil against erosion and contribute toward increasing the rate of its water intake. In some grasslands, there is more of this "old growth" present than the current season's crop of green shoots. The old shoots gradually decay or fall to the ground to produce a litter layer. The depth of litter depends on amount of growth, amount of consumption, and rate of transformation (by microorganisms and soil animals) to soil humus.

The plant biomass (including underground parts) of natural temperate grasslands is usually in the range of 1,000 to 3,000 grams per square metre of dry matter. As much as 85 percent of this biomass occurs underground. Dead material comprises only a fraction of the aboveground parts (both standing and as litter) but up to 75 percent of the underground parts. About half of the energy fixed in a given year is deposited in underground tissues. The rate of decomposition of dead, unconsumed plant parts is such that, on the average, the amount of biomass present disappears in about two years aboveground and about four years underground. The amount of plant litter present usually reaches an equilibrium state (*i.e.*, the condition that exists when the rate of disappearance equals the rate of new litter added), which varies from less than 100 grams per square metre in dry, temperate grasslands to about 1,000 grams per square metre in black-soil areas. Lower values than these prevail in areas where a higher than usual proportion of the shoots enter the consumer food web or where recurrent fires occur.

In grassland stands in the subhumid tropics and subtropics, biomass of herbage sometimes exceeds 5,000 grams per square metre; but the rate of decomposition of organic matter is very high in such areas, leading to low quantities of litter and soil organic matter.

It has been estimated that about two-thirds of the energy of the grassland ecosystem is released through the activity of reducers and decomposer organisms that feed on detritus and animal wastes. This activity is almost all in the soil or in the litter at the soil surface. Only general

Importance of decomposer organisms

conclusions can be given concerning the proportion of this activity caused by invertebrates and other decomposing reducers, since it has not yet been quantitatively evaluated. Under some conditions such organisms are very numerous. The liveweight of termites in the tropical savanna of the Ivory Coast, for example, has been found to be 40 grams per square metre in an area where aboveground herbivores measure only one gram per square metre. It is generally agreed, however, that the soil microflora (microscopic plants, including fungi and bacteria) is the single most important group of organisms affecting the turnover of energy. The biomass of soil microflora has been estimated to be 400 to 600 grams per square metre of soil surface on a dry-weight basis; however, probably less than 1 percent of this is active at any one time.

It is not yet possible to identify to what extent various groups of soil microflora contribute to the total of decomposer activity in any given system. The rate of decomposition in standing dead vegetation is usually slower, however, than in litter and in vegetation in contact with the ground. The bulk of the microbial decomposition does not occur until the litter has either made contact with the soil or has become densely compressed just above the soil surface. The smaller macrofauna (animals bigger than microbes), such as litter-feeding insects and worms, are active in bringing litter into more intimate contact with the soil.

UTILIZATION OF GRASSLANDS

Many early human civilizations developed in grassland regions, so man should be familiar with the ecology of grasslands. Greater interest, however, has been shown in converting grasslands to the growth of annual crops than has been devoted to considering whether they would have been a more valuable resource in an untilled state.

Domesticated grazing animals occupy the most important role in man's conversion of natural-grassland plant growth to a form of food that satisfied him. While conservationists in the past may have looked upon the domesticated grazing animal as an intruder in natural grassland, this does not necessarily mean that the grassland environment will deteriorate through the replacement of natural by introduced animals. The philosophy of range management that has developed in North America is based on the concept of obtaining the highest sustained level of animal production on natural grassland that is compatible with maintenance of the resource. Range ecologists have been much more conscious of the need to conserve land resources than have agriculturalists. In the management of arable lands, for example, the guiding principle has been almost exclusively determined by the need to produce the maximum harvestable yield, a practice hardly compatible with conservation. The advantage that the ecologist sees in the use of domesticated livestock in the rational management of rangelands is that the distribution and density of these animals is under his control to a far greater extent than would be possible with native animals. The domestication of the range is thus seen as a stabilizing situation.

The success achieved in increasing the harvestable yield of intensively managed arable land and improving semipermanent grasslands of woodland climate has led to the suggestion that the plant cover of nonarable grasslands should be changed as much as possible by the introduction of domesticated forage crops that have been selected and bred for high yield and for optimum response to fertilization and management. A high degree of environmental control is needed, however, to utilize the higher potential of these species, and the indication is that the native grass cover cannot be excelled by introduced species for range production on nonarable land.

Effects of herbicides and fertilizers

Attempts have been made to increase the productivity of natural grassland by the use of herbicides and fertilizers. Weed control of rangeland is economically practical only in cases where the weedy situation has been induced by mismanagement and when this situation is corrected. The effect of fertilizers is variable, yielding better returns where moisture conditions are most favourable. Some native species do not respond to increased levels of nutrient supply and may be replaced after fertilization by species that are more productive but dependent for survival on a continued supply of fertilizer nutrients.

After the initiation of tillage, highly fertile, temperate grasslands gradually (over a period of 50 to 100 years) attain a new level of equilibrium, which is associated with a lower content of soil organic matter. The full impact on organic-matter content will not occur until the original organic material is replaced by that formed from the annual agricultural plant cover. The concept that corrective measures can be taken by future human generations by addition of chemical fertilizer does not account for changes in soil structure that may be associated with declining organic content.

The maintenance of both arable and nonarable ecosystems in grassland zones is vital to the continued provision of food for the world. The temperate grassland zones include a very important portion of the cropland of the earth (for example, 90 percent of the grain for commerce originates here); the tropical and subtropical grasslands and savannas provide a possible means for expanding agriculture when technology is developed to manage these lands on a long-term basis.

BIBLIOGRAPHY. J.W. BEWS, *The World's Grasses: Their Differentiation, Distribution, Economics and Ecology* (1929), a comprehensive review of the nature of grasses and grasslands of Africa and their relation to those of other parts of the earth; J.L. CLOUDSLEY-THOMPSON, *The Zoology of Tropical Africa* (1969), a scientific account of the wildlife of the African savannas; R.M. MOORE (ed.), *Australian Grasslands* (1970), a compilation by many specialists (prepared for the International Grassland Congress) of grazing resources, both natural and man-modified; H.L. SHANTZ and C.F. MARBUT, *The Vegetation and Soils of Africa* (1923), a classical account, with maps, resulting from a trip through Africa by two distinguished Americans who were leaders in the vegetation and soils disciplines; G.M. ROSEVEARE, *The Grasslands of Latin America* (1948), an account of grazing resources based on a survey of the scientific literature; J.E. WEAVER, *North American Prairie* (1954), an authoritative, popular work on the history, nature, and response to man's occupation of the natural vegetation of the corn belt; and with F.W. ALBERTSON, *Grasslands of the Great Plains* (1956), an account similar to the above on the natural vegetation of the Great Plains.

(R.T.C.)

Gravitation

Gravitation is a universal force of attraction acting between all matter. The trajectories of bodies in the solar system are determined, except for relatively tiny effects, by the laws of gravity, while on Earth all bodies have a weight or downward force of gravity proportional to their mass, which the Earth's mass exerts on them. Gravity is by far the weakest known force in nature and therefore plays no role in determining the internal properties of everyday matter. But because of the long reach and universality of the gravitational attraction, gravity plays a central role in shaping the structure and evolution of stars, galaxies, and the entire universe.

The works of Isaac Newton and Albert Einstein dominate the development of gravitational theory. Newton's classical theory of gravitational force held sway from its presentation in his *Principia*, published in 1687, until Einstein's work in the early 20th century. Even today, Newton's theory is of sufficient accuracy for all but the most precise applications. Einstein's modern field theory of general relativity predicts only minute quantitative differences from the Newtonian theory except in a few special cases. The major significance of Einstein's theory is its radical conceptual departure from classical theory and its implications for further growth in physical thought (see RELATIVITY).

DEVELOPMENT OF GRAVITATIONAL THEORY

Early concepts. The classical Greek philosophers considered the motions of the celestial bodies and of objects on Earth as basically unrelated. The former was not considered as gravitationally determined, as the celestial bodies were seen to follow perpetually repeating, nonde-

scending trajectories in the sky. Aristotle envisioned such bodies as possessing "natural" motions that did not require external causes or agents. In this view, celestial bodies underwent their own particular "natural" motion, while massive earthly objects possessed a natural tendency to move toward the Earth's centre. Two other Aristotelian viewpoints were: that a body moving at constant speed required a continuous force acting on it and that force must be applied by contact rather than interaction or force at a distance. These views impeded understanding of the principles of motion and hence retarded the development of a theory of universal gravitation. During the 16th and early 17th century, however, several scientific contributions to the problem of earthly and celestial motion set the stage for Newton's gravitational theory.

Kepler's laws of planetary motion

Johannes Kepler, accepting the Copernican perspective —in which the planets orbited the Sun rather than the Earth—and using Tyco Brahe's improved measurements of planetary movements, succeeded in describing the planetary orbits by simple geometrical and arithmetical relations. Kepler's three quantitative laws of planetary motion were: (1) the planets describe elliptic orbits, of which the Sun occupies one focus; (2) the line joining a planet to the Sun sweeps out equal areas in equal time; and (3) the square of the period of revolution of a planet is proportional to the cube of its average distance from the Sun. During this same period, Galileo made major progress in understanding the properties of "natural" motion and simple accelerated motion for earthly objects. He realized that bodies uninfluenced by forces would continue indefinitely to move and that force was necessary to change motion, not to maintain constant motion. Galileo performed experiments to show that the Earth's gravity produced constant downward acceleration and that the downward gravitational acceleration was independent of the bulk or composition of bodies.

Newton's law of gravity. The modern quantitative science of gravitation began with the work of Newton. He assumed the presence of an attractive force between all massive bodies; this force does not require bodily contact but acts at a distance. By invoking his law of inertia (bodies not acted upon by a force move at constant speed in a straight line), Newton concluded that a gravitational force exerted by the Earth on the Moon was needed to keep it in a circular motion about the Earth rather than in a straight line and that this force could be, at long-range, of the same kind as the force with which the Earth pulled objects on its surface downward. Galileo had previously measured the downward acceleration of bodies on Earth to be approximately 980 centimetres (32 feet) per second per second, while Newton calculated that circular orbital motion of radius R and period T required a constant inward acceleration A equal to the product of $4\pi^2$ and the ratio of the radius to the square of the time:

$$A = \frac{4\pi^2 R}{T^2}. \tag{1}$$

Applying this formula to the Moon's orbit, which has a radius of about 384,000 kilometres (about 60 Earth radii) and a period of 27.3 days, the inward acceleration is approximately 2.7×10^{-3} metres per second per second; this is the same as $1/3{,}600$ $(1/60^2)$ times the Earth's surface acceleration. Newton deduced that the gravitational force between bodies diminishes as the inverse square of the distance between the bodies, as he could thus relate the two accelerations to a common interaction. A further assumption, that the mass of the Earth acts gravitationally on the outside world as if the mass were concentrated at the Earth's centre, was needed to obtain his relationship. Newton proved mathematically that this assumption was true for all spherically symmetric bodies.

Newton saw that the gravitational force between bodies must depend on masses of the bodies. Since a body of mass M experiencing a force F accelerates at a rate F/M, a force of gravity proportional to M would be consistent with Galileo's observation that all bodies accelerate under gravity toward Earth at the same rate. Newton's law of gravity can therefore be expressed mathematically by a relation expressing the law. If F_{12} is the magnitude of the gravitational force acting between masses M_1 and M_2 separated by distance D_{12}, then the force equals the product of these masses and of G, a universal constant divided by the square of the distance, D_{12}^2:

The universal constant G

$$F_{12} = \frac{GM_1M_2}{D_{12}^2}. \tag{2}$$

The constant G is a quantity having the physical dimensions (length)^3/(mass)(time)^2, its numerical value depending on the physical units of mass, length, and time used. (G is discussed more fully in a later section.) To obtain the total gravitational force on a body produced by many masses represented as M_i, where the subscript i stands for the positions 1, 2, . . . n, the individual forces must be vectorially added together, as force has direction as well as magnitude. Letting $\boldsymbol{D}_i$ be the spatial vector from the mass M to the mass M_i in the same way, the force on M due to several masses becomes the sum of the forces due to each mass separately. In the following expression, in addition to the quantities already explained, the symbol Σ represents the sum, and the factor $\boldsymbol{D}_i/D_i^3$ is needed to give the direction and numerically is equivalent to division by D_i^2:

$$F = GM \sum_{i=1}^{n} \frac{M_i \boldsymbol{D}_i}{D_i^3}. \tag{3}$$

This is Newton's gravitational law in its vector form. The simpler expression gives the surface acceleration on Earth; setting a mass equal to the Earth's mass M_E and the distance equal to the Earth's radius r_E, the downward acceleration of a body at the surface g is equal to the product of the universal gravitational constant and the mass of the Earth divided by the square of the radius:

$$g = \frac{GM_E}{r_E^2}. \tag{4}$$

Weight and mass

The weight W of the body can be measured by the equal and opposite force necessary to prevent the downward acceleration; this is M_g. The same body placed on the Moon's surface has the same mass, but as the Moon has a mass of about $1/81$ times that of the Earth and a radius of only 0.27 of that of the Earth, the body on the Moon's surface acquires a weight of only $1/6$ its Earth weight, as demonstrated by the U.S. Apollo astronauts. In orbiting satellites, where no force prevents the free fall of the satellites in the gravitational field, the cargo of humans and instruments experiences weightless conditions although the masses remain the same as on Earth.

The two equations above can be used to derive Kepler's third law, for the case of circular planetary orbits. By putting the expression for the acceleration A in equation (1) equal to the force of gravity for the planet, GM_pM_S/R^2, divided by the planet's mass M_p, M_S being the mass of the Sun, and R, T being the radius and period of the orbit, respectively, the following equation is obtained:

$$\frac{GM_S}{R^2} = \frac{4\pi^2 R}{T^2}$$

or

$$R^3 = \left(\frac{GM_S}{4\pi^2}\right) T^2. \tag{5}$$

Newton was able to show that all three of Kepler's observationally derived laws followed mathematically from the assumption of his own laws of motion and the law of gravity stated above. In all observations of the motion of a celestial body, only the product of G and the mass can be found. Newton first estimated the magnitude of G by assuming the Earth's average mass density to be about 5.5 that of water, somewhat greater than the Earth's surface rock density, and calculating the Earth's mass from this. Then, taking M_E and r_E as the Earth's mass and radius, respectively, the value of G was

$$G = \frac{gr_E^2}{M_E}, \tag{6}$$

which numerically comes close to the accepted value of

6.7×10^{-8} (cm)³/(gm)(sec)², first directly measured by a Cavendish balance experiment (see below).

Comparing equation (4) above for the Earth's surface acceleration g with the R^3/T^2 ratio for the planets, a formula for the ratio of the Sun's mass, M_S, to Earth's mass, M_E, was obtained in terms of known quantities, R_E being the radius of the Earth's orbit:

$$\frac{M_S}{M_E} = \frac{4\pi^2 R_E^3}{g T_E^2 r_E^2} \cong 325{,}000\ . \qquad (7)$$

By using observations of the motion of the moons of Jupiter discovered by Galileo, Newton determined that Jupiter was 318 times more massive than Earth but only ¼ as dense, having a radius 11 times larger than Earth.

Interaction among celestial bodies

When two celestial bodies of comparable mass interact gravitationally, the bodies each orbit about a fixed point (the centre of mass of the two bodies), which lies between the bodies on the line joining them at a position such that the distances to each body multiplied by each body's mass are equal. Observing that the Sun's apparent position in the ecliptic (the plane in which the Sun seems to be moving around the Earth) oscillates every month by about 12 arc-seconds (superimposed upon its annual motion) and accounting for this by means of a motion of the Earth around the Earth–Moon centre of mass, it was concluded that this centre of mass is placed about 4,800 kilometres (3,000 miles) toward the Moon from the Earth centre. From this, the Moon was found to be about 1/81 (4,800/384,000) as massive as the Earth. With slight modifications Kepler's laws remain valid for systems of two comparable masses; the foci of the elliptical orbits are the two-body centre-of-mass positions, and putting $M_1 + M_2$ instead of M_S in the expression of Kepler's third law, equation (5) above, the third law reads:

$$R^3 = \frac{G(M_1 + M_2)}{4\pi^2} T^2. \qquad (8)$$

This agrees with equation (5) when one body is so small that its mass can be neglected. The rescaled formula can be used to determine the separate masses of binary stars (pairs of stars orbiting around each other; see STAR) that are of a known distance from the solar system. Equation (8) determines the sum of the stars' masses; and, if R_1, R_2 are the distances of the individual stars from the centre of mass, the ratio of the distances must balance the inverse ratio of the masses, and the sum of the distances is the total distance R. In symbols,

$$\frac{R_1}{R_2} = \frac{M_2}{M_1}\ ;\ R_1 + R_2 = R. \qquad (9)$$

These relations are sufficient to determine the individual masses. Observations of the orbital motion of double stars, of the dynamical motion of stars collectively moving within their galaxies, and of the motion of the galaxies themselves verify that Newton's law of gravity is valid to a high degree of accuracy throughout the visible universe.

Ocean tides, phenomena that mystified thinkers for centuries, were also shown by Newton to be a consequence of the universal law of gravitation, although the details of the complicated phenomena were not understood until comparatively recently. They are caused specifically by the gravitational pull of the Moon and, to a lesser extent, of the Sun (see TIDES).

In the period following Kepler and Newton, improved accuracy in the measurements of planetary motion led to small discrepancies from the simple predictions of Kepler's laws. All but a few were later shown in accord with the universal aspect of Newton's law of gravity. Small corrections due to the fact that all the planets must perturb each other explained almost all variations in the planets' motions. The exceptions proved to be of large importance. Uranus, the seventh planet from the Sun, was observed to undergo variations in its motion that could not be explained by perturbations due to Saturn, Jupiter, and the other planets. The English mathematician John Couch Adams and the French mathematician Urbain-Jean-Joseph Le Verrier independently assumed the presence of an unseen eighth planet that could produce the observed discrepancies in the motion of Uranus.

The discovery of Neptune

They calculated its position within a degree of which the planet Neptune was discovered in 1846. Measurements of the motion of the innermost planet, Mercury, over an extended period led astronomers to conclude that the major axis of this planet's elliptical orbit precessed (precession is the gyration or wobble of the axis of a rotating body affected by a gravitational field) in space at a rate 43 arc-seconds per century faster than could be accounted for from perturbations of the other planets. In this case, however, no other bodies could be found that could produce this discrepancy, and very slight modification of Newton's law of gravitation seemed to be needed. Einstein's theory of relativity precisely predicts this observed behaviour of Mercury's orbit (see RELATIVITY).

INTERPRETATION OF GRAVITY MEASUREMENTS

Potential theory. For irregular, nonspherical mass distributions in three dimensions, the vector equation (3) above, which expresses Newton's law of gravity essentially in its original form, is inefficient, though theoretically it could be used for finding the resulting gravitational field. The main progress, after Newton, in classical gravitation theory was the introduction of potential theory, which allows practical as well as theoretical investigation of the gravitational variations in space and anomalies due to the irregularities and shape deformations of the Earth.

Potential theory led to the following elegant formulation: The gravitational acceleration, g, a function of position R, $g(R)$, at any point in space is given from a function, Φ, called the gravitational potential, by means of a generalization of the operation of differentiation:

$$g(R) = \frac{\partial \Phi}{\partial x}\, i + \frac{\partial \Phi}{\partial y}\, j + \frac{\partial \Phi}{\partial z}\, k,$$

in which i, j, k stand for unit basis vectors in a three-dimensional Cartesian coordinate system. Whatever restriction on g is introduced by the mass density ρ, that restriction is then transferred to the potential function Φ and is expressed in an equation that was discovered by the French mathematician Siméon-Denis Poisson:

Poisson's equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) \Phi(R) = -4\pi G \rho(R),$$

in which the equation is to hold for well specified values of R. The significance of this approach is seen by observing that Poisson's equation can be solved under rather general conditions, which is not the case with Newton's equation. When ρ is non-zero, the solution is expressed as a definite integral:

$$\Phi(R) = G \int \frac{\rho(R')dR'}{|R - R'|}\ .$$

When $\rho = 0$ (that is, outside the Earth), Poisson's equation reduces to the simpler equation of Laplace and has a general solution expressed as a series of powers of the trigonometric cosine function of θ, the latitude angle measure from the north pole:

$$\Phi(R) = \frac{GM_E}{R}\left[1 - J_2\left(\frac{R_E}{R}\right)^2 \frac{3\cos^2\theta - 1}{2} - J_3\left(\frac{R_E}{R}\right)^3 \frac{5\cos^3\theta - 3\cos\theta}{2} + \ldots\right],$$

in which, if R is the distance from the Earth's centre, R_E is the average Earth radius, θ is the latitude angle measured from the north pole, and J_2, J_3, etc., are constants. The constants J_2, J_3, etc., are determined by the detailed mass distribution of the Earth; and, since Newton showed that for a spherical body all the J_n are absent, the J_n must be measurements of the deformation of the Earth from a spherical shape; J_2 measures the magnitude of the Earth's rotational equatorial bulge, J_3 measures a slight pear-shaped deformation of the Earth, and so on. By comparison of gravimeter measurements (see below) from many parts of the Earth's surface and by observations of perturbations on satellite orbits and the Moon's orbit, the parameters J_2 and J_3 have been found to be 1,082.7 $\times 10^{-6}$ and -2.4×10^{-6}, respectively. Higher terms in

the potential series are being detected and measured by continuous observation of near-Earth satellite orbits.

Effects of local mass differences. The method used to describe global features of the Earth's mass distribution is inadequate to represent gravitational variations due to local mass differences such as mountain ranges, mineral deposits of unusual density, ocean basins, etc.; so other methods have been developed for these purposes. Several geological features of the Earth were first discovered from gravity measurements. Using gravimeters and horizontal pendulums, observers expected the additional bulk of mountain ranges compared with surrounding plains to produce slight attractive gravitational forces; instead, in many cases, slight repulsion was found. Such repulsion supports the view that mountain ranges float, possessing deep roots, much as an iceberg does, of underlying lightweight material that displaces the denser material of the Earth's interior. Portable gravimeters, which can detect variations of a part in 10^9 in the gravitational force, are today in wide use for mineral and oil prospecting, unusual underground deposits revealing their presence by creating local gravitational variations (see below).

Man-made satellites tracked during orbital motion around the Moon and around the planet Mars have followed trajectories that are best understood by assuming that these bodies possess large regions of anomalous mass densities (called mascons) that produce gravitational field variations. The lunar mascons are of sufficient size to have perturbed the Apollo manned lunar landings, and the paths of the approaching capsules were found to need adjustment to account for the perturbations and to reach their desired landing sites.

Mascons

GRAVITATIONAL THEORY AND OTHER ASPECTS OF PHYSICAL THEORY

The Newtonian theory of gravity is based on an assumed force acting between all pairs of bodies; that is, an action at a distance. When a mass moves, the force acting on other masses has been considered to adjust instantaneously to the new location of the displaced mass. Special relativity theory states that all physical signals travel no faster than the speed of light. This theory, with the field theory of electrical and magnetic phenomena, have met such empirical success, however, that most modern gravitational theories are constructed as field theories consistent with the principles of special relativity. In a field theory the gravitational force between bodies is formed by a two-step process: (1) One body produces a gravitational field that permeates all surrounding space but has weaker strength farther from its source. A second body in that space is then acted upon by this field, experiencing a force. (2) The Newtonian force of reaction is then viewed as the response of the first body to the gravitational field produced by the second body, there being at all points in space a superposition of gravitational fields due to all the bodies in it.

Special relativity and electromagnetic field theory

Field theories of gravitation. If the gravitational field has a theoretical and conceptual existence of its own, various new predictions of gravitational phenomena can be made. The equations governing the time evolution of the field predict a finite speed of propagation of disturbances through the field, replacing the direct instantaneous action at a distance by a delayed interaction transmitted by the field. The gravitational field in most modern theories is, in fact, free to change dynamically in certain modes independent of sources and to transmit energy and momentum in these modes, in a manner similar to electromagnetic wave radiation. These gravitational waves, which have been extensively studied mathematically since Einstein first showed their existence in theory, may have recently been demonstrated experimentally in the form of pulsed radiation received from the centre of the Galaxy. The possible detection of these extra-terrestrial gravitational waves was performed by monitoring cotemporaneous mechanical vibrations present in a pair of isolated, several-ton aluminum cylinders located about 1,500 kilometres (900 miles) apart. Field theories of gravity predict that gravity waves will excite such mechanical oscillations in these rigid bodies.

Field theories of gravity, Einstein's general relativity being an important example, also predict specific corrections to the Newtonian force law, the corrections being of two basic forms: (1) When matter is in motion, additional gravitational fields (analogous to the magnetic fields produced by moving electric charges) are produced; also, moving bodies interact with gravitational fields in a motion-dependent way. (2) Unlike electromagnetic field theory, in which two or more electric or magnetic fields superimpose by simple addition to give the total fields, in gravitational field theory nonlinear fields proportional to the second and higher power of the source masses are generated, and gravitational fields proportional to the product of different masses are created. Gravitational fields themselves become sources for additional gravitational fields. Examples of some of these effects are shown below. The acceleration, $\mathbf{A}$, of a moving particle of negligible mass that interacts with a mass, m, which is at rest, is given, in the following formula, derived from Einstein's gravitational theory. The expression for $\mathbf{A}$ now has, as well as the Newtonian expression as given in equation (1), further terms in higher powers of Gm/R^2—that is, in G^2m^2/R^4— and V is the particle's velocity vector, $\mathbf{A}$ its acceleration vector, $\mathbf{R}$ the vector from the mass m, and c is the speed of light. When written out, the sum is

Corrections of Newtonian force law

$$\mathbf{A} = -\frac{GmR}{R^3} + 2\frac{G^2m^2\mathbf{R}}{c^2R^2} - \frac{3}{2}\frac{GmR}{R^3}\left(\frac{V^2}{c^2}\right) - \frac{\mathbf{V}\cdot\mathbf{AV}}{c^2} - \tfrac{1}{2}\frac{V^2}{c^2}\mathbf{A} + \ldots.$$

This expression gives only the first post-Newtonian corrections; terms of higher power in $1/C$ are neglected. For planetary motion in the solar system the $1/C^2$ terms are smaller than Newton's acceleration term by at least the factor 10^{-8}, but some of the consequences of these correction terms are measureable and important tests of Einstein's theory. It should be pointed out that prediction of new observable gravitational effects requires particular care; Einstein's pioneer work in gravity has shown that gravitational fields affect the basic measuring instruments of experimental physics—clocks, rulers, light rays—with which any experimental result in physics is established. Some of these effects are listed below:

1. The rate that clocks run is reduced by proximity of massive bodies; *i.e.*, clocks near the Sun will run slowly compared with identical clocks farther away from it.
2. In the presence of gravitational fields the spatial structure of physical objects is no longer describable precisely by Euclidean geometry; for example, in the arrangement of three rigid rulers to form a triangle, the sum of the subtended angle's will not equal 180°. A more general type of geometry, Riemannian geometry, seems required to describe the spatial structure of matter in the presence of gravitational fields (see PHYSICAL THEORIES, MATHEMATICAL ASPECTS OF).
3. Light rays do not travel in straight lines, the rays being deflected by gravitational fields. To distant observers the light-propagation speed is observed to be reduced near massive bodies.

Gravitational fields and general theory of relativity. In Einstein's general theory of relativity the physical consequences of gravitational fields are stated in the following way. Space-time is a four-dimensional non-Euclidean continuum, the curvature of space-time's Riemannian geometry being produced by or related to the world's matter distribution. Particles and light rays travel along the geodesics (shortest paths) of this four-dimensional geometrical world.

The experimental foundations for modern theories of gravity can be classified into two categories—null experiments and the detection of extremely small (post-Newtonian) effects. Null experiments establish the absence of conceptually possible gravitational effects, usually thereby greatly restricting the class of acceptable laws of gravity. The small differences from Newtonian gravitation, and their interpretation, are discussed below.

Null experiments and post-Newtonian effects

At the turn of the century the Hungarian physicist Baron Lóránt (Roland) Eötvös found that different materials accelerated in the Earth's field at identical rates to an ac-

curacy of one part in 10^9. Recent experiments have increased the observed equality of accelerations to one part in 10^{11}. Newtonian theory is in accord with these results because of his postulate that gravitational force is proportional to a body's mass.

Inertial mass is a mass parameter giving the inertial resistance to acceleration of the body when responding to all types of force. Gravitational mass is determined by the strength of the gravitational force experienced by the body when in the gravitational field g. The Eötvös experiments, therefore, show the equality of gravitational and inertial mass for different substances.

Einstein's special theory of relativity views inertial mass as a manifestation of all the forms of energy in a body according to his fundamental relationship $E = mc^2$, E being the total energy content of a body, m the inertial mass of the body, and c the speed of light. Viewing gravitation, then, as a field phenomenon, the null result of the Eötvös experiments indicates that all forms of nongravitational energy must identically couple to or interact with the gravitational field, because the various materials in nature possess different fractional amounts of nuclear, electrical, magnetic, and kinetic energies, yet they accelerate at identical rates.

In the general theory of relativity the gravitational field also interacts with gravitational energy in the same manner as with other forms of energy, an example of that theory's universality not possessed by most other theories of gravitation. Eötvös' experiments using celestial bodies that contain a detectable fraction of internal gravitational energy were testing this feature of the general theory of relativity in the 1970s; these experiments are intended to determine whether the various solar-system bodies accelerate at identical rates in the Sun's field. Measurements of great precision, using radar and laser ranging, of the time-dependent interbody distances between the Earth and Moon, the Earth and Jupiter, and other such relationships are now possible (see below).

Gravitational consequences of the equivalence principle. A decade before he completed his full mathematical theory of gravity, Einstein predicted new gravitational effects using the equivalence principle. Observing that the equality of gravitational and inertial mass made impossible a distinction between uniform gravitational fields and accelerated coordinate systems, he proposed the null principle: no experiment can distinguish between local gravitational fields and accelerated coordinate systems. He then was able to show that clocks would run slower when near massive bodies and that light would be deflected toward massive bodies by their gravitational field.

Einstein's null principle

Newton's third dynamical law states that every force implies an equal and opposite reaction force. Modern field theories of force contain this principle by requiring every entity that is acted upon by a field to be also a source of the field. A recent null experiment established to a one-part-in-20,000 accuracy that different materials produce gravitational fields with a strength the same as they are acted upon by gravitational fields. In this experiment a sphere of solid material was moved through a liquid of identical weight density. The absence of a gravitational effect on a nearby Cavendish balance instrument during the sphere's motion is interpreted as showing that the two materials had equal potency in producing a local gravitational-field anomaly.

Other experiments have brought confirmation of Einstein's predictions to an accuracy of a few percent. Using the Mössbauer effect to monitor the nuclear reabsorption of resonant gamma radiation (hard X-rays), a shift of wavelength of the radiation which travelled vertically tens of metres in the Earth's gravitational field was measured, the slowing of clocks (in this case the nuclear vibrations are clocks) as predicted by Einstein has been confirmed to 1 percent precision. If ν and $\Delta\nu$ are clock frequency and change of frequency, respectively, h is the height difference between clocks in the gravitational field g. This is

$$\frac{\Delta\nu}{\nu} = -\frac{gh}{c^2}.$$

For a height of ten metres (about 30 feet) this effect produces only a one-part-in-10^{15} change in clock rates. The predicted deflection of light in gravitational fields was first detected during a 1919 solar-eclipse experiment. Recently progress has been made in measuring a related effect, the slowing of light's speed of propagation when near massive bodies. Timing the round-trip travel time for radar pulses between Earth and other inner planets or artificial satellites passing behind the Sun, experiments have confirmed to about 4 percent the prediction of an additional time delay, Δt, given by a formula in which M_S is the Sun's mass, R_1 and R_2 are the distances from the Sun to Earth and to the other reflecting body, D is the distance of closest approach to the Sun of the radar pulses. The additional time delay Δt is expressed as $4G$ times the Sun's mass over the cube of the velocity of light, c, times the logarithm of the quantity: four times the product of R_1 and R_2 divided by the square of D. In symbols,

$$\Delta t = \frac{4GM_S}{c^3} \ln\frac{4R_1R_2}{D^2}.$$

These radar ranging experiments to bodies in the solar system naturally complement the classical astronomical optical measurements of bodies' angular positions as seen from Earth. Radar tracking of the inner planets confirms to a few percent the precessional motion of the planets' elliptical orbits predicted by the general theory of relativity.

SOME ASTRONOMICAL ASPECTS OF GRAVITATION

Recent astrophysical discoveries such as quasi-stellar objects, pulsars, and the apparent gravitational radiation pulses being received from our galactic centre indicate that unusual astrophysical objects containing intense gravitational fields may exist in the universe. Modern physical theory predicts that a sufficiently massive object must, upon exhausting its nuclear fuel, inevitably collapse under its gravitational self-attraction, in most cases expelling part of its material in a supernovae explosion. Making only very general assumptions about the interaction properties of matter, it is concluded that the core of these collapsed stars will end either in a new superdense state of matter—the neutron star of typical density 10^{17} kilograms per cubic metre—or it is believed the core will collapse indefinitely toward a singularity, altering the properties of the surrounding space because of the enormously strong gravitational fields produced, so that physical communication with the outside world is quickly cut off, even light rays being trapped within this gravitational or black hole.

Pulsars

It is presently thought that the pulsars discovered in 1968 are neutron stars, having masses comparable to the Sun but diameters of only ten to 100 kilometres (six to 60 miles). The highly periodic electromagnetic radiation pulses coming from known pulsars with periods of from 1/30 of a second up to about a second require something like very small, but massive, rotating neutron stars as their sources.

The search for black holes

A black hole is the name given to the volume surrounding a collapsed star (an enormous mass in a tiny space), in which the gravitational field is so large that no radiation can get out; as a result it cannot be observed from outside. The astronomical search for a black hole is made difficult because of the very short time required for their formation (fractions of a second) and the relative infrequency of their creation. After being formed they emit no radiation or signals for the astronomer to detect. Present efforts are concentrated on the possibility of detecting a gravitational hole with a visible companion binary star. The gravitational wave pulses believed to be detected by the hundreds per year may be the gravitational radiation that would result from massive bodies being strongly accelerated into a large black hole (of thousands of solar masses) located near the centre of the Galaxy.

This field is experiencing such rapid theoretical and experimental development that these present viewpoints must certainly be considered as somewhat tentative and temporary.

(K.L.N.)

THE ACCELERATION OF GRAVITY ON THE EARTH'S SURFACE

More than 300 years ago Galileo, in studying how things fall toward the Earth, discovered that the motion is one of constant acceleration. He was able to show that the distance a falling body travels from rest in this way varies as the square of the time. The acceleration due to gravity at the surface of the Earth is about 980 centimetres (about 32 feet) per second per second; that is, following its release, an object will gain in speed 980 centimetres per second for each second it falls.

Perhaps Galileo's most noted conjecture was that in a perfect vacuum all bodies would fall at the same rate; if they are released together, they will strike the ground together. This conjecture, in combination with his observation that the motion of free fall is one of constant acceleration, leads to the result that all bodies fall with the same acceleration. This assumption of a common acceleration has proved to be one of the cornerstones of gravitational physics; it has been tested many times—most notably in Eötvös experiments and in the recent extraordinarily precise experiments of the U.S. physicist Robert H. Dicke and his colleagues. Common acceleration indicates that a body's weight (the Earth's gravitational pull on the body) is proportional to its inertial mass —that is, to the resistance it offers to an accelerating force. The constant of proportionality is simply the acceleration of gravity. Thus, g plays a dual role: the value of the acceleration observed in free fall (the acceleration of gravity) is also the constant of proportionality between mass and weight (the force of gravity per unit mass).

Importance of g in many fields

Knowledge of the acceleration due to gravity is of importance to several different disciplines of the physical sciences. Its absolute value provides a base that, together with the standard of mass, establishes the derived standard of force. The standard of force, in turn, is a necessary quantity in the assignment of values to the electrical units of current and voltage. The acceleration due to gravity is also an important factor in the accurate pressure determinations needed for the thermodynamic temperature scale and the establishment of the International Practical Temperature Scale. Absolute measurements of the acceleration due to gravity are also of interest to the science of geodesy. Rapid advances have recently been made in setting up a world gravity network to establish reliable gravity values at a large number of base points located strategically over the Earth.

Variations in g. *Changes due to location.* Though often thought of as a constant over the surface of the whole Earth, the acceleration g varies by about ½ of 1 percent with position on the Earth's surface, from about 978 cm/sec^2 at the Equator to approximately 983 cm/sec^2 at the Earth's poles. This variation stems chiefly from the rotation of the Earth, as part of the Earth's pull is balanced by keeping objects rotating with the Earth instead of flying tangentially off into space (as mud does off a spinning wheel). This effect is also responsible for the bulge of the Earth at the Equator and the slight flattening at the poles. The distance to the centre of the Earth, therefore, increases with the bulge from the poles to the Equator, and consequently g, which for a spherical Earth of radius r_E and mass M_E is given simply by GM_E/r_E^2, is less toward the Equator.

In addition to this broad-scale variation, local variations of a few parts in 10^6 or smaller are caused by variations in the density of the Earth's crust as well as height above sea level. Further, at any one particular place, g, considered to be the resultant force, varies with time as a result of the changing gravitational attraction of the Sun and the Moon.

Changes with time. For most purposes, only knowledge of the variation of gravity with time at a fixed place (tidal variation; see TIDES) or of the gravity differences from place to place is required. Accordingly, the great bulk of gravity measurements that have been made are relative; that is, they measure only the differences between gravity values at various places. In tribute to Galileo the unit gal (equivalent to 1 cm/sec^2) has been adopted as the unit of acceleration of gravity, and this has been subdivided to milligal, essentially one part in 10^6 of g, for convenience; thus, one milligal (mgal) = 0.001 gal = 0.001 cm/sec^2.

Great progress was made during the 1960s in the development of new standards of accuracy for the measurement of g on land, on ships, and in the air, as well as in space. Recent progress in gravimetry has been influenced by improvement in the accuracy and reliability of absolute gravity determinations. As a result, a worldwide network of gravity base stations and calibration lines has been established. The methods used for the measurement of the acceleration due to gravity in terms of the fundamental units of length and time are described below.

Only two basic methods have been used to measure the absolute acceleration of gravity: by timing (1) freely falling bodies and (2) those that move under gravity but whose motion is constrained in some way, as in the case of a pendulum.

Pendulum measurements of g. Pendulum measurements of g are familiar to almost everyone who has taken an introductory course in physics. A pendulum is called simple if it consists of a heavy bob at the end of a nearly weightless arm or compound if the weight is distributed also through the arm. In the case of a simple pendulum, the time of swing is proportional to the square root of the length divided by the acceleration of gravity. This relative ease of timing of a pendulum swing offers a distinct timing advantage over free-fall measurements. The constraint on the body's free fall permits a large number of "drops" to be made very conveniently during a measured interval of time. The price one pays for this gain is that, having introduced a constraint, the effects of the constraint on the performance of the pendulum must be taken precisely into account. The English physicist Henry Kater showed (1817) that if the period of swing was the same about each of the points of support for a pendulum that could be hung from either of two fixed points, the distance separating these points of suspension was equal to the length of a simple pendulum having the same period. Once the equality of periods has been established (by adjusting the position of attached weights), the problem is then reduced to that of measurement of the common period and of the distance separating the two supports. Reversible pendulums, which can reach an ultimate accuracy in the range of a few parts in 10^6 (1 mgal), provided the only practical basis for absolute measurements of g from Kater's time until the middle of the 20th century.

The work of Henry Kater

The most accurate and most straightforward way of measuring the acceleration due to gravity is now to measure directly the acceleration of a freely moving body. Only relatively recently has it been possible by electronic control to realize the necessary accuracy in the measurement of short time intervals to permit effective measurement of g by direct free fall.

Accuracy attained in the known value of g. The early free-fall experiments (dating from 1952) used geometrical optics to define the position of an object as it fell. From 1963, direct interferometric methods using corner cube mirrors, one of which was dropped, have led to more accurate distance measurements during the free fall. More recently still, lasers have been used as light sources in free-fall interferometric devices. By the 1970s the best absolute-gravity experiments had demonstrated accuracies in the 0.01–0.05 mgal range. Furthermore, measurements with a semi-portable laser interferometer apparatus have now been made at a number of stations covering a gravity range of 4,800 mgal with an absolute accuracy of five parts in 10^8 (0.05 mgal).

For most purposes only a knowledge of the variation in gravity from place to place is required. Accordingly the great bulk of gravity measurements are relative and give gravity differences between places on the Earth. These can then be referred to an absolute system to produce gravitational values for the sites.

Accuracy of pendulum measurements. Since the time of Newton, measurements of gravity differences (strictly of the ratio of one value to another) have been made by timing the same pendulum at two sites where gravity is to be compared. Already in 1818, such relative measure-

ments reached an accuracy of a few parts in 10^6. The most accurate work is now done with two pendulums swinging in opposite phase; in this way the sway of the support due to the reaction of the pendulums is eliminated, and also any movements of the support will produce equal and opposite changes in the periods of the two pendulums, which can be made to cancel out if the mean period of the two pendulums is used.

The accuracy of relative measurements with a pendulum depends on timing accuracy and the constancy of the conditions. Further, the difference in gravity is obtained in absolute units and therefore does not require any instrumental calibration. Accuracy of modern pendulum observations is limited by the scatter of the results when a pendulum is swung repeatedly in one place and, mainly, by changes in the pendulums that occur during transportation from place to place. The best claimed accuracy is a few tenths of a milligal.

Use of gravity meters. Up to about 1930 the pendulum was the only instrument available for relative gravity measurements, even for small scale geophysical prospecting. The development of static gravimeters restricted pendulum measurements to providing the calibration for these gravimeters. The growing number of truly absolute determinations can be expected to make even this use obsolete.

Spring gravity meters

Spring gravity meters balance the force of gravity mg on a mass m in the gravity field to be measured, against the elastic force of the spring, using either electronic or mechanical means to achieve high sensitivity. Vibrating string gravity meters in which the string's vibration frequency is determined by g have also been developed. A device of this type was employed by the Apollo 17 astronauts on the Moon to conduct a gravity survey of their lunar landing site. One of the most recent developments has been the superconducting gravimeter, an instrument in which the position of a magnetically levitated superconducting sphere is sensed to provide a measure of g.

Modern gravity meters may have sensitivities greater than 0.005 mgal, the standard deviation of observations in exploration surveys being, in the best performance, of the order of 0.01–0.02 mgal.

Differences in gravity measured with gravimeters are obtained in quite arbitrary units—divisions on a graduated dial, for example. The relation between these units and milligals can only be determined by reading the instrument at a number of points where g is known as a result of absolute or relative pendulum measurements. Further, because an instrument will not have a completely linear response, known points must cover the entire range of gravity over which the gravimeter is to be used.

Gravimetric surveys. Recently, by combining all available absolute and relative measurements, it has been possible to obtain the most probable gravity values at a large number of sites to a high degree of accuracy. The culmination of gravimetric work begun in the 1960s has been a worldwide gravity reference system having an accuracy of one part in 10^7 (0.1 mgal) or better.

Since g is an acceleration, the problem of its measurement from a vehicle that is moving and therefore unavoidably accelerating relative to the Earth raises a number of fundamental problems. Pendulum, vibrating string, and spring-gravimeter observations have been made from submarines; using gyrostabilized platforms, relative gravity measurements with accuracies approaching a few mgal have been and are being made from surface ships. Experimental measurements with various gravity sensors on fixed-wing aircraft as well as on helicopters have been carried out. Additional information about the Earth's gravitational field has been made possible through the use of artificial satellites. Tidal gravity variations are observed through the use of sensitive recording gravimeters. One of the most remarkable recent measurements was the mapping of variations in the gravitational field on the visible side of the Moon by observed perturbations in the orbits of Lunar Orbiter satellites.

The value of gravity measured at the surface of the Earth is the resultant of such component factors as (1) the gravitational attraction of the Earth as a whole, (2) centrifugal force caused by the Earth's rotation, (3) elevation, (4) unbalanced attractions caused by surface topography, (5) tidal variations, and (6) unbalanced attractions caused by irregularities in underground density distributions. Most geophysical surveys are aimed at separating out the last of these in order to interpret the geological structure. It is therefore necessary to make proper allowance for the other factors.

The free-air and Bouguer corrections factors. The first two factors imply a variation of gravity with latitude that can be calculated for an assumed shape for the Earth. The third factor, the decrease in gravity with elevation, due to increased distance from the centre of the Earth, amounts to -0.3086 mgal/m (-0.09406 mgal/ft). This value, however, assumes that material of zero density occupies the whole space between the point of observation and sea level, and it is therefore termed the free-air correction factor. In practice, the mass of rock material that occupies part or all of this space must be considered. Where the topography is reasonably flat this is usually calculated by assuming the presence of an infinite slab of thickness equal to the height of the station, h, and having an appropriate density σ; its value is $+0.04185\ \sigma h$ mgal/m or $+0.01276\ \sigma h$ mgal/ft. This is commonly called the Bouguer correction factor.

Terrain or topographic corrections can also be applied to allow for the attractions due to surface relief if the densities of surface rocks are known. Tidal effects the amplitudes of which are lower than 0.3 mgal can be calculated and allowed for.

In defining anomalies, the observed gravity g_0 is compared with the theoretical value $g\gamma$ for the latitude of the station. The difference is then corrected for the elevation, h, of the station, using the free-air correction factor, F, with or without the Bouguer correction factor, B. The topographic correction, T, is also applied, giving in symbols: free-air anomaly $= g_0 - g_\gamma + Fh + T$ and Bouguer anomaly $= g_0 - g_\gamma + (F - B)h + T$. In exploration surveys Bouguer anomalies are most commonly used. Free-air anomalies or isostatic anomalies, in which a further correction for crustal material based on the compensation of mass above a certain depth has been applied, are those generally adopted in geodetic work.

(Ja.F.)

THE GRAVITATIONAL CONSTANT, G

Difficulties in measuring G

The gravitational constant, G, has been introduced in the first part of this article. Although G is one of the most fundamental constants in nature, it probably is the least accurately known because of the extreme weakness and universality of the gravitational interaction. The weakness is such that the force of attraction between two spheres each weighing one kilogram spaced 0.1 metre apart is only about 1.3×10^{-13} of the pull of gravity on one of the spheres. The universality of the gravitational force, already assumed by Newton, is supported by the failure of experiment to show any variation of G that depends on the kind or size of the attracting masses, their temperature, or the amount of other matter placed between them. Consequently, in order to determine G, it is not only necessary to measure very tiny forces or torques but also to do so in the presence of the much larger perturbing forces due to all of the other matter in the universe, as it is impossible to shield the masses under investigation from the rest of the universe. The classical theory of celestial mechanics is based upon Newton's law and is used to predict with great accuracy the paths of the Moon, the Sun, the planets, and other bodies through space; but solution of the mathematical equations obtained from astronomical observations does not give G uniquely, but rather—if M is the mass of one of the interacting bodies—the highly precise product MG.

Principal methods of measuring G. There are three principal methods of measuring G: (1) in which the pull of the Earth is compared with that of a large natural mass, such as a mountain or other topographical mass, on that of a known mass called a test mass; (2) in which a comparison is made between the Earth's attraction and that of a known mass on a test mass, as in the common

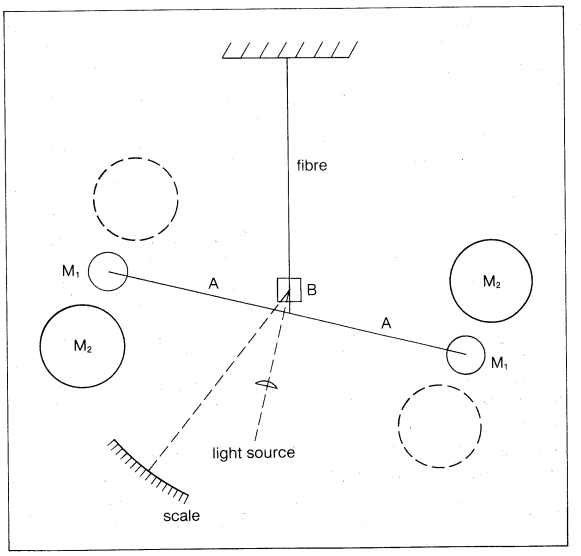


Figure 1: Modified Cavendish deflection experiment. The masses M_1 are deflected horizontally by the masses M_2, first in position shown by full circles, then at dotted circles. The twist is measured on the scale from reflection of light from a small mirror at B (see text).

chemical balance experiments; and (3) in which direct determination of the force between known masses is made in the laboratory. At the present time, experiments in the first category are of historical interest only and have not yielded a reliable value of G. As more accurate values of G become available by other methods, however, such experiments in connection with modern geophysical investigations can give valuable information concerning the densities and density gradients, especially in the Earth and Moon.

The various experimental methods of determining G in the laboratory have in the past employed some form of torsion balance or the common balance. According to the 18th-century English physicist Lord Henry Cavendish, the torsion balance was invented by the British geologist and astronomer, the reverend John Michell, for the purpose of measuring G, though Michell died before the apparatus was completed. The first reliable measurement of G was carried out by Cavendish in 1798 with a torsion balance of the Michell type. Figure 1 shows schematically the method used by Cavendish, though many details of the actual apparatus differed from those of the Figure. Two small spherical masses, M_1, are mounted on a stiff rod, A, and suspended by a torsion fibre, f. A light beam is reflected from a mirror, B, rigidly attached to A and brought to focus on a scale, S. When the large spherical masses, M_2, are placed in the fixed position shown in Figure 1, the gravitational interaction produces a torque on the small mass system, M_1, which causes A to rotate around its vertical axis. This rotation continues until the torque produced by the gravitational attraction of the small mass system, M_1, and the large mass, M_2, is balanced by the restoring torque due to the twist in the fibre. The deflection of the light beam on the scale, S, is then a measure of the twist in the fibre from which the torque, due to gravitational interaction, can be found. If the large masses, M_2, are shifted from the positions shown in the Figure into the positions indicated by the dotted circles, the small mass system is twisted in the opposite direction, and the deflection of the light beam on the scale S is reversed. The extreme smallness of the deflection of the light beam usually limits the precision of the method. This torsion-balance-deflection method was perfected by the English physicist Sir Charles Vernon Boys, who first developed the quartz-fibre suspension.

The work of Lord Henry Cavendish

The common balance method consists in supporting equal spherical masses, M_1, from the weighing pans of an equal arm chemical balance. A large spherical mass, M_2, mounted on a turntable is rotated directly under one of the masses M_1. If d is the distance between the centres of the two spheres, the downward gravitational attraction of M_2 on M_1 given by GM_1M_2/d^2 adds to the pull of gravity M_1g due to the Earth, and it tips the equal arm balance through a small angle that is measured by a pointer or optical magnification device. M_2 is then rotated until it is under the weight attached to the opposite arm of the balance, and the deflection that now takes place in the opposite direction is measured. From calibration of the balance, G can then be calculated in terms of g and d. Careful experiments were carried out with this method, principally by the English physicist John Henry Poynting, using a specially constructed balance; but M_1g was so much larger than the gravitational interaction of the two masses that the results were not as reliable as those obtained with the torsion balance, in which the pull of the Earth is not directly superposed upon the gravitational interaction.

Figure 2 shows a schematic diagram of the torsion-balance-oscillation method. The torsion balance that supports the small masses is similar to that described in Figure 1, but the large attracting masses, M_2, are placed in the same straight line that passes through the small masses, as shown in Figure 2. The torsion balance containing the small masses is then given a small displacement and the period of oscillation measured. This period can also be calculated in terms of G and the other quantities. The large masses, M_2, are then placed in the positions indicated by the dotted lines, which are in a line perpendicular to the line joining the small masses, M_1, and passing through its centre. A second determination of the period of the torsion pendulum is then measured. From these values it is possible to determine G. The advantage of this method over the torsion-balance-deflection method is that periods of oscillation can be measured with greater precision than small deflections, but it suffers from uncertainties due to the effect of gravitational gradients and neighbouring masses. This method produced what until recently was usually regarded as the most reliable value of G. In 1971 it was again used to obtain a new value of G. The new study is being carried out in the Grotta Gigante in Italy. The small masses, M_1, each weigh 10 kilograms and the large masses 500 kilograms while the torsion-wire suspension is 90 metres long. With the oscillation method, however, the torsion pendulum must be in a vacuum to prevent large damping through air resistance. The method has been modified in another study so that both the small and large mass systems are on separate torsion suspensions and are brought into resonance, but the result with this arrangement is believed to be less precise than that achieved by the U.S. physicist Paul Heyl (published in 1930; see Table).

The Table is a partial list of the measured values of G. The accuracy of the measurements made before 1930 is difficult to evaluate, but it is probable that the values re-

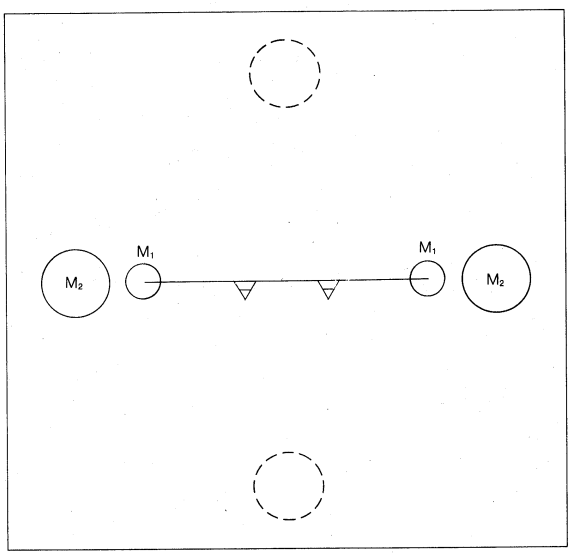


Figure 2: Elements in the torsion-balance resonance experiment (see text).

Best Values of the Gravitational Constant G
(up to 1942)

	year	method	G unit: $(10^{-11}Nm^2Kg^{-2})$
Cavendish	1798	torsion-balance (deflection)	6.754
Reich	1838	torsion-balance (deflection)	6.61
Baily	1842	torsion-balance (deflection)	6.475
Von Jolly	1881	common balance	6.465
Wilsing	1889	metronome balance	6.596
Poynting	1891	common balance	6.698
Boys	1895	torsion-balance (deflection)	6.6576
Braun	1896	torsion-balance (oscillation)	6.6579
Eötvös	1896	torsion-balance (oscillation)	6.65
Richarz	1898	common balance	6.685
Burgess	1901	torsion-balance (deflection)	6.64
Heyl	1930	torsion-balance (oscillation)	6.670
Zahradnicek	1932	torsion-balance (resonance)	6.659
Heyl and Chrzanowski	1942	torsion-balance (oscillation)	6.673

ported by Heyl and his associates—$G = (6.670 \pm 0.015) \times 10^{-11}$ Newton-square metre per square kilogram—is the most reliable among those listed.

A new method was devised in the 1960s and is illustrated in Figure 3. Two comparatively large spherical masses (ten-kilogram tungsten spheres) are mounted on a rotary table that can be turned about its vertical axis by a servo-controlled electric motor. An airtight cylindrical chamber also is rigidly mounted upon the rotary table with its vertical axis coincident with the axis of rotation. The small mass system, which consists of a small cylinder with its axis horizontal, is supported by a torsion fibre fastened to the top of the cylinder so that the fibre hangs in the axis of rotation of the table, as shown in Figure 3. The gravitational interaction between the large and small mass systems tends to rotate the small mass system in a direction that brings the axis of the small cylinder into coincidence with an imaginary line passing through the centres of mass of the large spheres. This changes the angle θ. A change in θ, however, also produces a change in the angle β between a light beam and its reflection from a mirror mounted on the small mass system. The light source and the photo-diode sensing system are rigidly mounted on the table in such a way that a change in β generates a photo-diode signal that actuates the servo-motor, which in turn rotates the table so that β and θ remain constant to less than one-half second of arc. Since the angle θ remains constant, the gravitational interaction between the small and large mass systems produces a constant torque on the small cylinder. This in turn produces a constant angular acceleration that results in an equal constant angular acceleration of the rotary table. This method possesses three novel features that should contribute to improved reliability and accuracy. First, the acceleration of the table can be measured with much greater precision than a deflection, because the change in frequency of rotation can be measured over a long period of time, and this automatically increases the precision. Second, the two mass systems rotate about their common axis many times during a measurement, and consequently the effects of gravitational fields and field gradients due to unavoidable stationary masses are essentially cancelled. Also, the effect of the masses moving with the rotary table, as well as the table itself, is eliminated by determining the acceleration of the rotary table with the large spheres removed from the table. Finally, the coordinates or positions of the interacting masses with respect to a rotating coordinate system do not change during a measurement; hence, the necessary distance can be precisely determined. A preliminary value of $G = (6.674 \pm 0.012) \times 10^{-11}$ nm^2/kg^2 has been reported, but the method is claimed to promise much higher accuracy.

From an illustration by Marina Givotovsky in *Physics Today* (May, 1971)

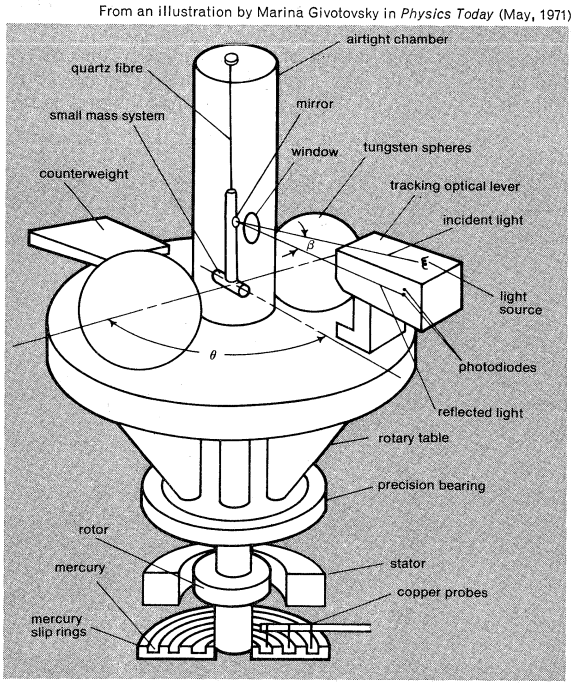


Figure 3: Angular acceleration method of measuring gravitational interaction (see text).

Weighing the Earth. If Newton's equation mentioned above is applied to the attraction of the Earth on a small mass, M_1, with g as the acceleration of gravity and M_E and r_E as the mass and effective radius of the Earth, respectively, then $g = GM_E/r_E$. Since g and r_E are known, the mass of the Earth, M_e, can be obtained if G is known. For this reason Cavendish and some of the early workers referred to their work as "weighing the Earth." Actually, no way is yet known for reliably obtaining the mass of the Earth, Moon, planets, or other heavenly bodies—say in kilograms, tons, etc.—without a knowledge of G. The mass of the Earth is approximately 5.98×10^{24} kilograms (13.18×10^{24} pounds). Since the radius of the Earth is known, its volume can be calculated so that the mean density of the Earth is obtained. This mean density of the Earth is approximately 5.52 times that of water, while the mean density of the Sun is 1.43 times that of water. These mean densities are becoming of considerable importance in geophysical research.

Important gravities dependent on G

Fundamental character of G. In addition to the above practical needs for more accurate values of G, because of its fundamental nature it necessarily must enter into all major cosmological questions. Some cosmological theories predict a minute decrease in G of about one part in 10^{10} per year. Very precise measurements made in the early 1970s of radar-echo time delays between the Earth and Mercury indicate that G does not vary by more than four parts in 10^{10} per year, unless there are unknown compensating factors in the experiments. Speculation that G may be influenced by the position in space led to proposals for the measurement of G in space vehicles. Other suggestions postulate effective changes in G at very small and at very large distances and with time. The absolute measurements of G are not yet accurate enough nor is the time base long enough to resolve these or many other important questions. On the other hand, G seems, to the limit of present accuracy, to be a truly fundamental constant, both in magnitude and sign, and independent of perturbing effects not only in the case of the gravitational interaction of matter but also that of antimatter as well.

(J.W.B.)

BIBLIOGRAPHY

Theories and laws of gravitation: PETER G. BERGMANN, *The Riddle of Gravitation* (1968), a general introduction to the history and contemporary status of gravitation theory; Y.B. ZELDOVICH and I.D. NOVIKOV, *Relativistic Astrophysics*, vol. 1, *Stars and Relativity* (1971; orig. pub. in Russian, 1967), develops modern gravitational theory, tests, cosmology, and astrophysics; C. MISNER, K.S. THORNE, and JOHN A. WHEELER, *Gravitation* (1972), presently the definitive book on gravity and research applications; R.W. DAVIES (ed.), *Proceedings of the 1970 Conference on Experimental Tests of Gravitation Theories* (1971), a progress report on experimental efforts in gravitation; REMO RUFFINI and JOHN A. WHEELER, "Introducing the Black Hole," *Physics Today*, 24:30–36 (1971), a discussion of the physics of astrophysical bodies in which gravity is of particular significance.

Gravitational constant: HENRY CAVENDISH, "Experiments to Determine the Density of the Earth," *Phil. Trans. R. Soc.*, 88:469–526 (1798), describes the first reliable measurement of the gravitational constant; F.C. CHAMPION and NORMAN DAVY, *Properties of Matter*, 3rd ed. (1959); C.V. BOYS, "Measurement of Mean Density of Earth," in R.T. GLAZEBROOK (ed.), *A Dictionary of Applied Physics*, vol. 3, pp. 279–285 (1922, reprinted 1950), excellent accounts of work before 1943; P.R. HEYL, "A Redetermination of the Constant of Gravitation," *J. Res. Natn. Bur. Stand.*, 5:1243–1290 (1930), and with PETER CHRZANOWSKI, "A New Determination of the Constant of Gravitation," *ibid.*, 29:1–31 (1942), usually considered to be the most reliable until recently; A.H. COOK, "A New Determination of the Constant of Gravitation," *Contemp. Phys.*, 9:227–238 (1968), an authoritative review article; *U.S. National Bureau of Standards Misc. Publ. 253* (1963), a critical analysis of measurements up to 1963; J.W. BEAMS *et al.*, "New Method of Measuring the Gravitational Constant," *Bull. Am. Phys. Soc.*, Series 2, 10:249 (1965); R.D. ROSE *et al.*, "Determination of the Gravitational Constant *G*," *Phys. Rev. Lett.*, 23:665–658 (1969), a report on recent measurements; I.I. SHAPIRO *et al.*, "Gravitational Constant: Experimental Bound on Its Time Variation," *ibid.*, 26:27–30 (1971), reports the results of a search for a change in the gravitational constant with time.

Earth's gravitational field: W.A. HEISKANEN and F.A. VENING MEINESZ, *The Earth and Its Gravity Field* (1958), a classic text but portions, particularly those dealing with instrumentation, are somewhat out of date; B.F. HOWELL, *Introduction to Geophysics* (1959), sections dealing with instrumentation are dated; J.A. JACOBS, R.D. RUSSELL, and J. TUZO WILSON, *Physics and Geology* (1959); A.H. COOK, *Gravity and the Earth* (1969), a fairly elementary treatise; IUGG Report, pt. 1, *Trans. Am. Geophys. Un.*, vol. 52, no. 3 (1971), an up-to-date summary.

(K.L.N./Ja.F./J.W.B.)

Graywackes

Graywacke is the name applied to generally dark-coloured, very strongly bonded sandstones that consist of a heterogeneous mixture of rock fragments, feldspar, and quartz of sand size (1/16–2 millimetres [0.002–0.078 inches]), together with appreciable amounts of mud matrix (less than 1/16 millimetres [0.002 inches]). Almost all graywackes originated in the sea, and many were deposited in deep water by density (turbidity) currents.

The name graywacke (from the German *Grauwacke*) was first used by the German mineralogist Abraham Werner in 1787. It describes the colour and texture of the rock, "wacke" being a term used for the heterogeneous weathering products derived from igneous and metamorphic rocks. The classic German locality for graywacke is the Devonian and Carboniferous sequence of the Harz Mountains.

In 1818 John Mawe wrote, "Geologists differ much as to what is, and what is not, Grey Wacce." This is still true. Recent definitions by various geologists characterize graywackes as sandstones that (1) contain more than 10, 15, or 20 percent mud matrix; (2) contain less than 75 percent quartz and more than 25 percent nongranitic feldspars and rock fragments; (3) have similar mineralogy and chemistry to those from the Harz Mountains; or (4) display associations of sedimentary structures that are characteristic of "turbidites"—*i.e.*, rocks deposited by turbidity currents. Many geologists now use the name only in a general sense, in reference to the dark, usually quartz-poor, often well-indurated sandstones that occur in turbidite (or flysch) sequences.

There are great differences between the Harz and Bunter sandstones in Germany, the Aberystwyth and Millstone grits in Great Britain, and the Franciscan and Navajo formations of the United States. The differences in structure and mineralogy reflect fundamental differences in both mode of deposition and relationship to phases of earth history. The persistence of the term graywacke, despite the confusion surrounding it, is testimony to the need to express these differences succinctly, for this term has been applied to the first of each pair cited above. The plethora of definitions indicates the difficulties experienced in finding generally applicable criteria that adequately express the differences.

This article treats the composition, properties, occurrence, and origin of graywacke. See SANDSTONES; SEDIMENTARY ROCKS; MARINE SEDIMENTS; and SHALES for information on related rock types; see DENSITY CURRENTS for additional discussions of origin; and see MOUNTAIN-BUILDING PROCESSES and EARTH, GEOLOGICAL HISTORY OF for the geological significance of graywackes.

PHYSICAL AND CHEMICAL PROPERTIES

Texture and structure. Graywackes typically are poorly sorted, and the grain sizes present range over three orders of magnitude—*e.g.*, from 2 to 2,000 microns (8×10^{-5} to 8×10^{-2} inch). Commonly, the coarsest part of a graywacke bed is its base, where pebbles may be abundant. Shale fragments, which represent lumps of mud eroded from bottom sediments by the depositing current, may be concentrated elsewhere in the bed.

Size, shape, and orientation of particles

Many graywackes contain much mud, typically 15–40 percent, and this increases as the mean grain size of the rock decreases. The particles forming the rock are typically angular. This, and the presence of the interstitial mud matrix, has led to these rocks being called "microbreccias." The fabric and texture indicate that the sediments were carried only a short distance and were subject to very little reworking by currents after deposition.

Although many geologists believe that the mud matrix accumulated simultaneously with the coarser material, some believe it was derived from alteration of rock fragments subsequent to burial. This question remains unresolved. Some graywackes are notably deficient in matrix (less than 10 percent, or even less than 5 percent). This deficiency can occur at the base of beds or be characteristic of beds as a whole. No detailed explanation has yet been suggested.

Sections of graywackes cut parallel to the bedding display alignment of the long axes of the grains. This usually is best developed in the lower-middle part of a bed. The alignment direction varies from level to level in the bed and commonly deviates from the direction of current markings on the underside of the bed, although some workers report parallelism between grain orientation and current markings. Imbrication (overlapping) of grains is observed in sections cut normal to the bedding. The origin of this variability in grain orientation and its deviation from current markings are not understood.

Graded bedding and interbedding

Graywacke sequences are noted for having a characteristic, and usually well-developed, association of sedimentary structures. Typically the beds are sheetlike and are interbedded in a regular fashion with shales, each bed paralleling its predecessor with almost mathematical precision (Figure 1).

The thicknesses of beds in a graywacke sequence are log-normally distributed (that is, the logarithms of bed thickness are distributed about some mean value), or nearly so, and typically there is a strong positive correlation between maximum grain size and bed thickness. Some graywacke sequences that are called "proximal turbidites," meaning that they accumulate near the source area, largely lack this parallelism and regular interstratification.

The most widespread internal structure of graywackes is graded bedding (Figure 1), although some sequences display it poorly. Sets of cross strata more than three centimetres (1.28 inches) thick are very rare, but thinner sets are very common. Parallel lamination is very common, and convolute bedding is usually present. These internal structures are arranged within graywacke beds in a regular sequence. They appear to result from the action of a single current flow and are related to changes in the hydraulics of the depositing current. In some beds, the upper part of the sequence of structures is missing, presumably because of erosion or nondeposition. In others, the lower part is missing. This has been attributed to change in the hydraulic properties of the depositing current as it moves away from its source and its velocity decreases to the point at which the first sediment deposited is laminated, rather than massive and graded as is the case closer to the source.

The most typical external structures of graywacke beds

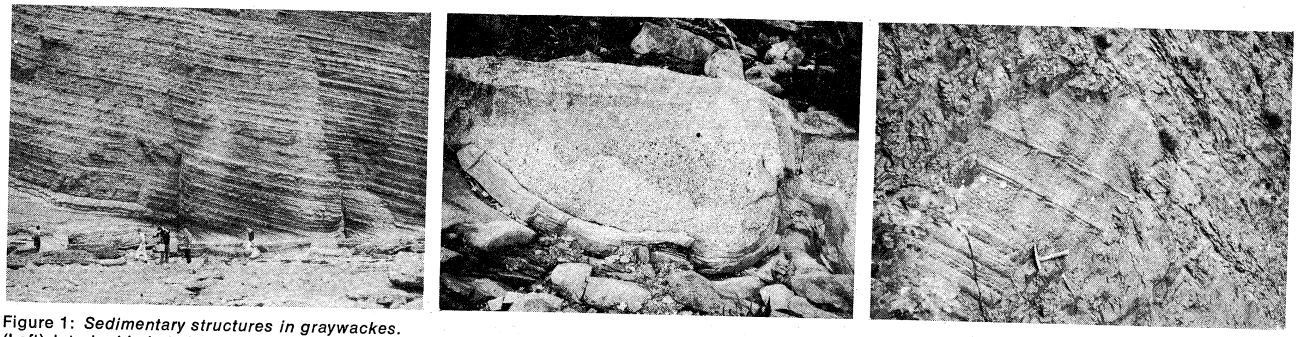


Figure 1: *Sedimentary structures in graywackes.*
(Left) Interbedded shales and graywackes, Lower Silurian Aberystwyth Grits, Wales. (Centre) Bed of very coarse Upper Cambrian graywacke, showing graded bedding and load casting, Denison Range, Tasmania. (Right) Groove molds on underside of graywacke bed, Middle Silurian Denbigh Grits, Wales.
By courtesy of K.A.W. Crook, the Australian National University, Canberra

Sole marks, casts, and fossil remains

are sole markings, which occur on their undersurfaces. Flute and groove molds (Figure 1) are the most characteristic, but many other structures have been recorded. Sole markings are not readily visible in all graywacke sequences, particularly where the beds are flat lying or the shale interbeds are well indurated (hardened). Most sole markings originate as infillings of sculptures on the surface of the underlying mud bed. Many of the sculptures are produced by the scouring action of the current before it commences to deposit and are infilled by sand during later stages of the same current. Many molds are elongated in the direction of movement of the current. They are therefore widely used as indicators of the initial direction of current flow and, by inference, the direction of the local paleoslope and location of source areas.

Usually the sculptures become somewhat distorted during their infilling because of the effects of compaction of the soft mud. In vertical sections this is expressed by characteristic "overhangs" of shale, termed load casts, along the irregular sand–shale contact.

The upper surfaces of graywacke beds are less well characterized by sedimentary structures. The most typical are current lineation and various worm tracks, particularly the highly sinuous form *Nereites*. Apart from these trace fossils, graywackes are usually sparsely fossiliferous. Where fossils occur they are generally free-floating organisms (graptolites, forams) that have settled to the bottom, or bottom-living (benthic), shallow-water organisms displaced into deeper water as part of sediment mass.

Colour. Fresh graywackes are, as their name suggests, coloured various shades of gray, ranging from neutral gray through greenish- to bluish-gray. The intensity of the colour varies from dark to light throughout.

Within this range, volcanic lithic graywackes are predominantly greenish-gray; graywackes with intermediate amounts of quartz are greenish- or bluish-gray; and quartzose graywackes are neutral gray. Plagioclase graywackes, and some more quartz-rich varieties of predominantly granitic derivation, are generally light greenish-gray to light gray.

Mineralogy. The mineralogical composition of graywackes varies widely (Table 1). Three groups can be recognized in terms of the content of free quartz grains of sand size (Figure 2). The first contains less than 15 percent free quartz and is derived from igneous rocks. It comprises two types of volcanic derivation, plagioclase graywacke (Figure 3), volcanic lithic graywacke, and serpentine graywacke, which is derived from serpentinite, a subcrustal rock. Of these, volcanic lithic graywacke is by far the most abundant. Sequences in which it occurs usually contain some beds of plagioclase graywacke. Serpentine graywacke is rare but is important because it reflects a deep rupture of the earth's crust.

Role of quartz in composition of graywackes

Quartz-poor graywackes are widespread in eastern Australia and around the margins of the Pacific and Caribbean but also occur within continents—for example, in the Urals. Serpentine graywacke is recorded from Colombia and the Solomon Islands.

Graywackes with from 15 to 65 percent free quartz form the second group. They are of varied composition and origin. Some are almost devoid of feldspar and are derived from sedimentary and low-grade metamorphic rocks. Others, richer in feldspar, are from heterogenous source areas containing volcanic, plutonic, metamorphic, and sedimentary rocks. Still others come from sources in which granite is widespread. The typical graywackes of North America and Europe are representatives of this group. Representatives also occur in New Zealand and eastern Australia.

The third group of graywackes (the subgraywackes of some authors) contains from 65 to 95 percent free quartz. Many of these would not be termed graywacke by those workers who adopt a mineralogical definition of the term, for such definitions commonly permit not more than 75 percent free quartz in the rock. These quartz-rich varieties exhibit the same dark colour and sedimentary structures as those poorer in quartz. They occur in thick geosynclinal sequences (deposits in great subsiding troughs), either as the dominant rock type or as a minor component with more common graywackes of the second group. Their high quartz content reflects source areas dominated by quartz-rich igneous, sedimentary, and metamorphic rocks or more heterogenous source areas in

Table 1: Mineralogical Composition of Graywackes
(percent)

	A*	B†	C‡	D§	E‖	F¶	G♀	Hδ
Quartz	8.8	0.4	—	29.16	19.57	26.93	48.94	80.2
K-feldspar	—	—	—	2.06	2.34	9.18	4.71	9.4
Plagioclase	67.5	12.7	2.0	2.92	21.05	18.36	4.71	1.0
Serpentinite	—	—	29.0	—	—	—	—	—
Granite and schist	—	—	13.0	—	14.22	9.55	2.76	1.0
Volcanic RF	2.5	63.6	13.5	—	14.22	11.02	2.76	—
Sedimentary RF	—	—	6.0	20.63	5.12	6.12	2.76	—
Low grade metam RF	—	—	12.0	9.96	22.45	□	2.76	1.0
Muscovite	—	—	—	—	0.75	—	5.65	1.0
Biotite	—	—	—	—	2.14	2.45	1.59	1.0
Chlorite	1.7	—	—	—	7.67	0.37	—	—
Pyroxene and hornblende	—	1.5	—	—	0.07	0.37	—	—
Other heavy minerals	0.4	0.6	1.5	—	—	1.08	—	5.2
Carbonate	1.4	0.4	—	—	—	—	—	—
Miscellaneous detritus	—	—	—	—	—	—	0.53	1.0
Matrix	—	20.3	—	35.24	□	10.31	15.41	□
Cement	17.8	0.5	23.0	—	4.63	3.99	20.12	—
Total	100.1	100.0	100.0	99.97	100.01	99.73	99.71	99.8◇

*Carboniferous plagioclase graywacke, Tamworth Trough, N.S.W. (K.A.W. Crook, 1960). †Average of 31 Devonian volcanic lithic graywackes, Tamworth Trough, N.S.W. (B.W. Chappell, 1968). ‡Serpentine graywacke, Tertiary, North Coast Basin, Colom. (W. Zimmerle, 1968). §Lithic graywacke (5 samples), Martinsburg Formation (Ordovician), Appalachians, U.S. (E.F. McBride, 1962). ‖Tanner Graywacke (Devonian), Harz, Ger. (H.G. Huckenholz, 1959). ¶Lithic graywackes, eastern Axial Facies, New Zealand Geosyncline (Upper Paleozoic and Mesozoic); W.R. Dickinson, 1969). ♀Average of 17 quartz graywackes, Tarricheskaya Formation (Triassic), Crimea (N.V. Logvinenko, *et al.*, 1961). δ92 modern deep-sea sands from Western province, North Atlantic (J.F. Hubert and W.J. Neal, 1967). □Not quoted separately. ◇Glauconite and mud pellets omitted.

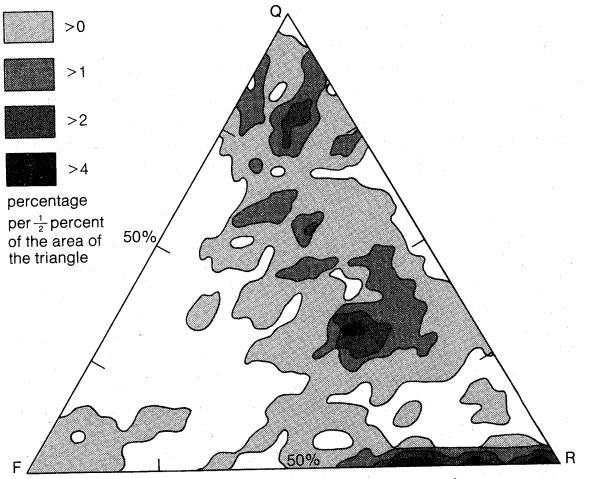


Figure 2: Modal analyses of 328 graywackes, plotted as contours in terms of percentages of free quartz grains (Q), feldspar grains (F), and rock fragments and other components (R) in the sand fraction.

Table 2: Geochemistry of Graywackes
(percent)

	A*	B†	C‡	D§	E‖	F¶	G♀
SiO_2	54.63	66.75	70.0	86.58	54.20	66.88	66.30
TiO_2	1.11	0.63	0.5	—	1.31	0.57	0.47
Al_2O_3	15.69	13.54	13.4	6.78	17.17	15.66	15.37
Fe_2O_3	1.65	1.60	1.2	1.08	3.48	1.33	1.26
FeO	7.30	3.54	2.7	—	5.49	2.59	2.85
MnO	0.16	0.12	0.1	0.07	0.15	0.07	0.07
MgO	3.70	2.15	2.1	0.36	4.36	1.57	2.05
CaO	5.01	2.54	1.4	0.97	7.92	3.52	3.97
Na_2O	4.79	2.93	3.5	0.59	3.67	3.84	3.87
K_2O	0.66	1.99	1.7	0.27	1.11	3.07	2.25
P_2O_5	0.23	0.16	0.1	—	0.28	0.21	0.13
H_2O^+	3.96	2.42	2.9	0.38	0.86	0.65	0.81
H_2O^-	0.32	0.55	2.9	2.19	—	—	0.06
CO_2	0.82	1.24	—	0.61	—	—	0.15
Total	100.03	100.16	99.6	99.88	100.00	99.90	99.61

*Average of 10 volcanic lithic graywackes, Devonian Tamworth Trough, N.S.W. (B.W. Chappell, 1968). †Average of 61 graywackes (F.J. Pettijohn, 1963). ‡Average of 10 representative Franciscan assemblage graywackes (Mesozoic; W.R. Dickinson, 1969). §Average of 3 quartz graywackes (Ordovician), Victoria (H.C. Richards, 1910). ‖Average of 49 andesites (S.R. Nockolds, 1954). ¶Average of 131 granodiorites (S.R. Nockolds, 1954). ♀Average composition of the Canadian Precambrian Shield (D.M. Shaw, *et al.*, 1967).

which intense chemical weathering eliminates the less stable minerals before the sediment is finally deposited.

Chemistry. The bulk chemistry of graywackes reflects their variable mineralogy. Four major groups can be recognized, which differ most notably in their silica (SiO_2) content. Volcanic lithic graywackes average 54 percent SiO_2 (Table 2A). Analyses of plagioclase graywackes are not available; they probably have a similar silica content. Analyses of serpentine graywackes are likewise unavailable. From their mineralogy they can be expected to form a second group, averaging about 40 percent SiO_2. The third group (Table 2B, C), with 65–70 percent SiO_2, essentially represents the graywackes of mixed origin, with from 15 to 65 percent free quartz. The fourth group, equivalent to graywackes with more than 65 percent free quartz, is poorly represented by analyses. Available data suggest an average content of approximately 85 percent SiO_2 (Table 2D).

Because graywackes contain unweathered detritus (disintegrated material), including appreciable mud, that accumulates as thick sequences in elongate troughs, the average composition of the graywackes in any sequence represents an average sample of the bulk mineralogy and chemistry of the source area. When data from several chemically similar sequences are averaged, such averages should represent typical average compositions for significant portions of the Earth's crust.

This is evident from comparison of the three chemical types of graywackes listed in Table 2 with other average rocks. The volcanic lithic graywackes strongly resemble andesites (Table 2E) but are richer in sodium. They are a sample of the crust in the vicinity of the boundary between continents and oceans. Graywackes with 65–70 percent silica, as might be expected from their location, approximate continental crust and granodiorite in composition (Table 2G, F). The high-silica graywackes have no evident analogue in standard geochemical average compositions but probably represent a fair average for the sandstone cover of continental platforms, excepting the readily soluble components such as carbonate and evaporite minerals.

Evidence of diagenesis. The almost universal strong induration of graywackes testifies to diagenetic processes; that is, to modification after burial. Their mineral content is not in equilibrium with near-surface conditions, and they accumulate in sequences of sufficient thickness to promote significant increases in pressure and temperature. Their initial permeability to fluids is low, owing to the presence of a mud matrix between the sand grains. Diagenetic reactions therefore proceed by internal chemical rearrangements rather than by the addition of material from outside.

The volcanic lithic graywackes display the most notable diagenetic (or burial metamorphic) effects. These involve the formation of hydrated calcium aluminosilicates and iron-magnesium silicates in the course of reactions that both utilize the water trapped within the sediments on burial and generate heat, thus promoting widespread and relatively uniform modification of mineral assemblages. Commonly, these reactions are not accompanied by strong deformation, so that the original fabric of the graywackes is preserved.

Mineral assemblages and mineral replacement

The newly formed mineral assemblages are in equilibrium with the temperature and pressure in the rocks. Many of the minerals are stable over only limited ranges of temperature and pressure. Thus a zonation of secondary minerals, approximately according to depth of burial, is commonly observed.

At shallowest depths, various zeolites (*q.v.*) are the char-

By courtesy of K.A.W. Crook, the Australian National University, Canberra

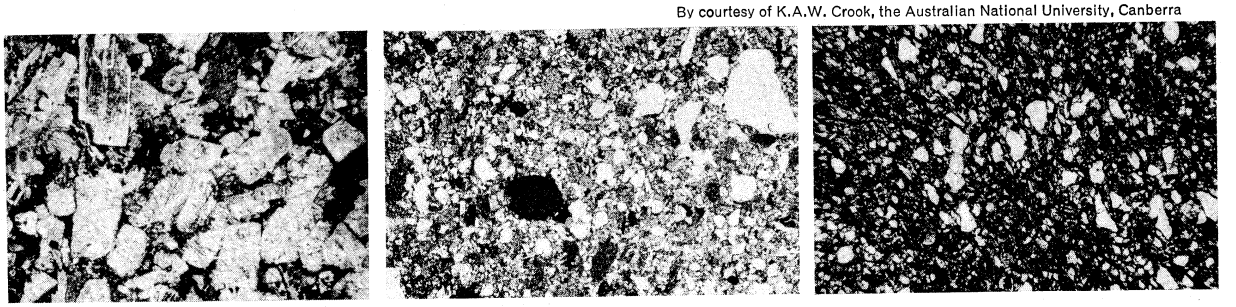


Figure 3: *Photomicrographs of graywackes.*
(Left) Lower Carboniferous plagioclase graywacke, Tamworth Trough, New South Wales, Australia. (Centre) Ordovician lithic wacke with 40 percent quartz and many clay fragments, Austin Glen, New York. (Right) Ordovician quartz-rich graywacke with abundant matrix material, Pittman Formation, Gundaroo, New South Wales, Australia. (All specimens magnified 20 ×.)

acteristic secondary calcium aluminosilicates, and montmorillonite is characteristic of the clay minerals (*q.v.*). With increasing temperature and pressure, prehnite, pumpellyite, epidote, and finally lawsonite take the place of the zeolites, and chlorite replaces montmorillonite. These minerals replace glass in rock fragments; typically, the calcium aluminosilicates also occur as replacements of feldspar and in the interstices between the sand grains. Pseudomorphic replacement of calcic plagioclase by sodic plagioclase (albite) is ubiquitous in rocks in which zeolites are no longer stable.

Graywackes rich in quartz commonly lack the highly unstable detritus present in volcanic lithic graywackes. Diagenetic modification is therefore less dramatic and probably occurs at somewhat lower temperature for a given depth of burial because exothermic chemical reactions are more limited. The principal modifications affect the sheet silicates, particularly biotite, which undergoes progressive bleaching and chloritization. The clay minerals of the matrix are also commonly reconstituted; this is largely responsible for the strong induration of these rocks.

The diagenetic mineral assemblages of graywackes provide information on the grade of metamorphism (temperature and pressure conditions) to which the rocks were subjected. Indeed, the two lowest grade facies (kinds) of metamorphism, the zeolite facies and prehnite-pumpellyite metagraywacke facies, are most typically developed in graywackes.

OCCURRENCE OF GRAYWACKES

Rock associations in geosynclinal sequences

The sedimentary rock most commonly associated with graywacke is mudstone, or shale, which in many cases is indurated to argillite or slate. Conglomerate and pebbly mudstone occur in addition in some proximal graywacke sequences. Other sequences have pelagic limestones between the graded-bedded units. In still other sequences, thick units of radiolarian chert and argillite are present that may contain occasional graywacke interbeds.

Radiolarian chert, spilite, and serpentinite form the geosynclinal rock association known as the "Steinmann Trinity." This association in fact has a fourth member, volcanic graywacke. All four rock types occur in the early formed parts of geosynclinal sequences, although the serpentinites are late intrusive bodies. The association indicates that the geosynclinal trough probably formed at depth, on a basement of oceanic-type crust.

Graywacke, together with shale, is an important component of the flysch sequence (primarily "turbides") that normally constitutes the major part of the "fill" of a geosynclinal trough. The flysch predates the deformation of the trough. In some geosynclines, sediments that are mineralogically and chemically similar to the flysch graywackes accumulate after the deformation of the trough. These are termed molasse sequences. They differ from the flysch in their sedimentary structures as a consequence of their accumulation in shallow marine or subaerial depositional environments.

Data presently available are insufficient to permit the presentation of the global distribution of the major types of graywackes. Graywacke sequences occur in almost all fold mountain belts except those dominated by limestones, such as the Canadian Rockies. Classic Paleozoic graywacke localities include the Harz Mountains, Germany; the Welsh Geosyncline; the Appalachians; and the Tasman Geosyncline of eastern Australia. Mesozoic and Cenozoic localities include the Coast Ranges of western U.S. and the New Zealand Geosyncline.

Graywacke sequences as old as Archean and as young as Pleistocene are known. Graywackes probably accumulated somewhere on earth during every geological period, but they seem to be less well represented in Proterozoic strata than in older and younger sequences.

ORIGIN AND GEOLOGICAL SIGNIFICANCE

Graywackes are commonly regarded as deep-sea sediments deposited by turbidity currents. This opinion stems from the classic investigations by P.H. Kuenen and C.I. Migliorini, who were the first marine geologists to demonstrate convincingly by experimental and field studies that dense sediment-charged currents hugging the sea floor could deposit sequences of muddy graded-bedded sands in deep water. The explanation has been given further impetus by observations on breaks in submarine cables following earthquakes. One such sequence of cable breaks to the south of the Grand Banks of Newfoundland in 1929 has been interpreted as due to a large turbidity current that reached velocities of 55 knots. Bottom samples taken over the abyssal plain to the south of the Grand Banks reveal the presence of a graded sand bed from 40 to 100 centimetres (12.7 to 39.4 inches) in thickness that is believed to occupy an area of 100,000 square kilometres (39,000 square miles). Relatively little is known about the hydrodynamics of turbidity currents. This accounts for the inadequacies of interpretation of the structures encountered within graywacke beds.

Turbidity-current and normal-current hypotheses

An alternative view of graywackes is that normal ocean currents may be responsible for their laminated and rippled portions and even for graded-bedded units. Observations in the deep sea reveal the existence of contour currents that flow parallel to the base of the continental slope and have sufficient velocity to produce ripples in sand. Several sequences are known where the paleoslope, as indicated by slump structures (slumping of bedding planes) that have scarcely moved laterally, is perpendicular to the current direction indicated by sole markings. Some of these can be explained by assuming a basin shape like that of the modern Adriatic Sea. There, much of the material introduced enters at one end and moves axially along the basin. Lesser amounts are introduced laterally, producing deposits that reflect the paleoslope of the basin sides. Such assumptions cannot, however, explain occurrences where slump structures are oriented normal to paleocurrent indicators in adjoining beds.

Many graywackes with graded beds are interstratified with and pass laterally into laminated graywacke–shale sequences. The latter deposits may require a different emplacement mechanism than the graded beds, but they may be deposits from the dilute distal parts of turbidity currents.

Some investigators consider that the high velocities claimed for turbidity currents on the basis of studies of submarine cable breaks are excessive, but no satisfactory alternative explanation for cable breaks upslope from graded sand deposits has yet been advanced.

Evidence from submarine canyons and fans

Investigations off the coast of southern California suggest that turbidity currents and submarine canyons are intimately related. Sediment is known to accumulate in the heads of the canyons by longshore drift and to be periodically flushed out. Although most people believe that the sediment moves down the canyon as a turbidity current, attempts to generate such currents artificially in the canyon heads have so far been unsuccessful. The observation of sand moving down modern canyons as grain flows and cascades with little displacement of water suggests an alternative mode of flushing. Sediment movement by this means, however, does not appear to be rapid enough to explain observed sudden losses of material from canyon heads. The role of grain flows in sedimentation has not yet been evaluated.

The canyons debouch at bathyal (deep-sea) to abyssal (sea-floor) depths and are fronted by submarine fans of sandy or coarser sediment. These fans are transected by channels with levee banks. Turbidity currents are believed to move down these channels, spreading laterally and eventually reaching the distal parts of the fans where they merge into the abyssal plain. These fans and their lateral equivalents make up the continental rise, which is believed to consist of turbidites.

The existence of submarine fans suggests that the typical geometry of a graywacke bed will be a tongue-shaped sheet. Numerous sheets of this kind may be stacked one upon the other with interbeds of finer sediment partly deposited from suspension. Ancient graywackes with this geometry are known in California, Wales, and eastern Australia.

The change in geomorphic character and current velocities across submarine fans betokens changes in sediment characteristics. Variability in the sedimentary features of graywackes has given rise to the concept of "proximality," a measure of the closeness of a graywacke bed to its source—*e.g.*, the distal end of a submarine canyon.

Some rare exceptions to the general rule of deep-sea origin for graywackes are known, although one of these apparent exceptions was re-interpreted in the early 1970s: an occurrence of graywacke with limestone in the Middle Devonian of the Hill End Trough, New South Wales, once thought to have accumulated in protected water some tens of metres deep, is now thought to have formed in deeper water. The Upper Ordovician of Girvan, Scotland, comprises a flysch sequence with abundant graywackes, the sedimentary features of which indicate a neritic or shallow marine origin. From the Tertiary flysch of the Pyrenees, bird footprints and other features suggestive of very shallow-water origin have been described. Lacustrine graywackes are known in the Tertiary of the Ridge Basin, California. Apparently the lake in which the sediments accumulated was sufficiently deep to prevent disturbance of the bottom, and bounded by two major transcurrent faults, the relief at its margins was sufficient to promote formation of turbidity currents.

Deep-sea versus shallow-marine graywackes

Although the concept of turbidity currents has proved a powerful tool for the interpretation of graywacke sequences since about 1950, the emergence of data not well explained by this concept is leading to the development of new concepts, of which grain flow and contour currents are the most significant. The wish to extend or supplant the turbidity current concept reflects both inadequate knowledge of submarine processes, including the behaviour of turbidity currents, and a facile, if preliminary, assumption that marine sedimentation below wave base is dominated by a single process. There is little doubt that the term flysch, which is used by many geologists as a synonym for graywacke sequences, embraces a range of sedimentary sequences as varied in their internal characteristics as are the subaerial and shallow-marine sequences dominated by cross stratification. Only when flysch sequences can be grouped into classes that reflect variations in deep-water environments will the relative importance of the various concepts be properly assessed. Nevertheless, it seems likely that the turbidity current concept will continue to play an important part in the interpretation of such sequences.

Depositional sites, geosynclines, and earth history

Three major depositional sites for contemporary graywackes have been recognized. The first sites, recognized early in the history of geology, are the deep-sea troughs adjoining fold mountain ranges, as, for example, those off the west coast of North and South America. These troughs have been the traditional site for ancient graywackes, especially those containing intermediate amounts of quartz. The second depositional sites, recognized rather more recently, are the troughs adjoining volcanic island arcs (*q.v.*)—for example, in Indonesia and Melanesia. These are the traditional sites of accumulation of volcanic graywackes. The third sites, recently recognized, are the continental rises along the margins of continents that lack adjacent high relief, as, for example, the east coast of North America. This latter has not been cited so far to explain any particular type of graywacke, but analyses of modern North Atlantic sediments (Table 1H) and considerations of the composition of ancient graywackes suggest that it is the typical site of accumulation of quartz-rich graywackes.

There is an intimate relationship between these depositional sites and geosynclines. The first and second sites have long been recognized as the modern analogues of ancient geosynclines. More recently the continental rise has been suggested by R.S. Dietz, an American marine geologist, as the precursor of many fold mountain belts that have been regarded as sited on old geosynclines.

Because of their association with geosynclines and their occurrence in thick sequences that have subsequently been folded, graywackes are taken to indicate zones of tectonic activity (folding and faulting) in the Earth's crust. They are generally regarded as pre-tectonic deposits, with respect to the basin in which they occur. They can also be contemporaneous with deformation of an adjoining earlier-filled basin that becomes the source of the graywacke sequence. Insofar as the Atlantic continental rise of North America is a modern graywacke accumulation, graywackes may also reflect tectonic quiescence with respect to both source and the basin of deposition.

The feature common to all modern depositional sites is that they adjoin land masses in areas of high submarine relief. The land mass may be the seismic or aseismic margins of a continent, or the margins and interstices of juvenile continental crust, the island arcs. Thus, all occurrences are related to major discontinuities in the Earth's crust. The quartz-poor graywackes include such types as serpentine graywacke; these reflect the emergence of subcrustal material at the Earth's surface, an intimate part of mountain building processes.

BIBLIOGRAPHY. A.H. BOUMA, *Sedimentology of Some Flysch Deposits* (1962), detailed descriptions of the sedimentary features of graywacke sequences, and with A. BROUWER (eds.), *Turbidites* (1964), a collection of papers on modern and ancient graywackes and related rocks; S. DZULYNSKI and E.K. WALTON, *Sedimentary Features of Flysch and Graywackes* (1965), descriptive and experimental studies of sedimentary structures in graywackes; B.W. CHAPPELL, "Volcanic Graywackes from the Upper Devonian Baldwin Formation, Tamworth-Barraba District, New South Wales," *J. Geol. Soc. Aust.*, 15:87–102 (1968), definitive work on mineralogy, petrology, and geochemistry of volcanic graywackes; D.S. COOMBS, "The Nature and Alteration of Some Triassic Sediments from Southland, New Zealand," *Trans. Roy. Soc. N.Z.* 82:65–109, (1954), the classic work on secondary alteration of graywackes; R.L. FOLK, *Petrology of Sedimentary Rocks* (1968), a discussion of the petrology of various sandstones, including graywackes; B.C. HEEZEN and G.M. EWING, "Turbidity Currents and Submarine Slumps and the 1929 Grand Banks Earthquake," *Am. J. Sci.*, 250:849–873 (1952), the first study suggesting the occurrence of turbidity currents in modern oceans; P.H. KUENEN and C.I. MIGLIORINI, "Turbidity Currents as a Cause of Graded Bedding," *J. Geol.*, 58:91–127 (1950), the classic paper that established turbidity currents as the means by which graywackes are deposited; F.J. PETTIJOHN, *Sedimentary rocks*, 2nd ed., pp. 229–242 (1957), a comprehensive treatment of graywackes. R.G. WALKER, "Turbidite Sedimentary Structures and Their Relationship to Proximal and Distal Depositional Environments," *J. Sedim. Petrol.*, 37: 25–43 (1967), paper that established the notion of proximal and distal graywacke-bearing sequences and described them.

(K.A.W.C.)

Great Barrier Reef

The largest structure ever built up by living creatures, the Great Barrier Reef is one of the most fascinating, as well as one of the most beautiful, of the natural wonders of the world. The dazzling, surf-fringed reef extends for some 1,250 miles (2,000 kilometres) off the northeastern coast of Australia, at a distance varying from ten to 100 miles offshore. Its length is thus comparable to the entire west coast of the continental United States south of the Canadian border, and, if placed off western Europe, it would extend from Portugal to Norway. It covers some 80,000 square miles, or an area larger than Syria. The reef actually consists of many thousands of separate reefs, many dry or barely awash at low tide, some with islands of coral sand, or cays, others fringing high islands or the mainland coast. In spite of this variety, the reefs share a common origin: each has been formed, over millions of years, from the skeletons and skeletal waste of a mass of living marine organisms. The "bricks" in the reef framework are formed by the calcareous remains of the tiny creatures known as coral polyps and hydrocorals, while the "cement" that binds these remains together is formed in large part by the remains of the organisms known as coralline algae and polyzoas. The interstices of this framework have been filled in by vast quantities of skeletal waste produced by the pounding of the waves and the depredations of other boring organisms. The reef as a whole has a remarkably large number of coral species (at least 350), and some regions are justly famous for

their riotous profusion of colour and form. The reef has risen on the shallow shelf fringing the Australian continent, in warm waters that have enabled the coral to flourish (it cannot exist where average temperatures fall below 70° F [21° C]). This sustaining continental shelf experiences a change of gradient at about 50 fathoms (300 feet) depth. In addition to its scientific interest, the reef has become increasingly important as a tourist attraction, while growing concern over the preservation of its magnificent natural heritage had led, by the early 1970s, to increased controls on such potentially threatening activities as drilling for petroleum resources.

Exploration and scientific study. The first physical encounter between Europeans and the Great Barrier Reef may be said to date from 1770, when the British explorer Capt. James Cook ran his ship, the "Endeavour," aground upon it. Earlier references in his journal to the various component reefs sighted in his run up the coast of what became Queensland refer to them as "banks" or "shoals," and the reef's coralline structure was only discovered when dawn broke after the grounding. The work of charting channels and passages through the maze of reefs, begun by Cook, continued during the 19th century, when many scientists, including the American marine zoologist Alexander Agassiz, made expeditions to this newfound wonder of the world. The contribution of the Great Barrier Reef Expedition of 1928–29 remains unsurpassed in the fields of coral physiology and the ecology of coral reefs, while a modern laboratory was set up on Heron Island, at latitude 23° S, in 1951. This is jointly administered by the University of Queensland and the Great Barrier Reef Committee, the latter body representing a number of individuals and institutions interested in the reef.

Discovery of the reef

Structure and hydrology. Borings have established that reefs were growing on the continental shelf as early as Miocene time, more than 25,000,000 years ago, but underlying nonmarine sediments of the Lower Tertiary, more than twice as old, indicate that the region was then above sea level. Subsidence of the continental shelf has proceeded, with some reversals, since the early Miocene: platform-like surfaces on the seabed and changes of gradient on the sides of the reefs are distinguishable at ten, 16, 20–22, and 32 fathoms and, less certainly, at 36, 56, and 80 fathoms and may well be related to periods of standstill in the sea-level fluctuations associated with the melting of Ice Age glaciers. A bench lying six to ten feet above the present sea level was probably formed during the latest postglacial maximum rise in sea level. The submarine topography of the reef area is complicated by the valleys, products of ancient periods of land erosion, that cross the continental shelf linking present river mouths with some of the deeper gaps in the outer reef barrier.

The water environment of the Great Barrier Reef is formed by the surface water layer of the southwest Pacific Ocean. The reef waters show little seasonal variations: surface-water temperature is high, ranging from 70 to 100° F (21° to 38° C), and vertical gradient, or change in temperature with depth, is small, due to the perpetual stirring action of the southeast trade winds, which pound the outer edge of the reef for nine months of the year. Average salinity is about 35 parts per 1,000; the oxygen content is high, with a 90 percent saturation most of the time. The origins of the reef waters are complex. Probable surface sources include those of the Trade Wind Drift, Southern Equatorial, and East Australian current systems, with a seasonal influx from the Arafura Sea to the north. Upwellings at the fringe of the continental slope bring in subsurface water masses from the adjacent Coral Sea, while monsoonal rains and land runoff add a further contribution. Reef tides have a daily cycle; the range is nowhere less than eight feet, and the draining waters often form foaming white falls at the reef edges. At the entrance to Torres Strait, at Australia's northern tip, and in Broad Sound on the central eastern Australian coast, tidal range reaches a massive 13 feet and 30 feet, respectively. The waters are generally crystal-clear, with submarine features clearly visible at depths of 100 feet.

Water composition

Life and resources of the reef. The marine life of the Great Barrier Reef is very rich, with some reduction in the number of species southward. The forms of life include, in addition to the corals, anemones, worms, gastropods, lobsters, crayfish, prawns, and crabs, and a great variety of fishes and birds. Many of the small fishes of the reef are strikingly colour banded and of bizarre shapes, darting or gliding in and out of complex coral structures that are muted in every colour of the rainbow. The giant clams are famous; they reach four feet in width and weigh up to 200 pounds, display a remarkable variety of colours when open, and can send spouts of water shooting upward when they close. The deadly sea wasps (the jellyfish *Chironex fleckeri*) have a more infamous reputation, but the most notorious reef animal is the crown-of-thorns starfish (*Acanthaster planci*), which has reduced the colour and attraction of many of the central reefs by eating much of the living coral. Among the reef organisms, marine algae are second only to the corals; the encrusting red algae *Lithothamnion* and *Porolithon* form the fortifying purplish-red algal rim that is one of the Great Barrier Reef's most characteristic features, while the green alga *Halimeda* flourishes almost everywhere. Above the surface, the plant life of the cays is very restricted, consisting only of some 30 to 40 species, practically all of them widely characteristic of the Indo-West Pacific biological province. Some varieties of mangrove occur only in the northern cays, although they are common on mainland mud flats along the full length of the Queensland coast.

The reef has contributed little to the fishing industry, as net fishing is precluded by the irregularities in the sea floor, although a small amount of trawling for mackerel and coral trout does take place, and tuna is taken in the adjoining Coral Sea waters. Turtles are now protected from commercial exploitation, but the pearl-shell industry is still important to the Torres Strait islanders in the north. Though world demand for mother-of-pearl has been reduced by the increased use of plastics, shells are still gathered in considerable numbers, being taken to culture farms for the making of artificial pearl. Large beds of prawns have been found on the sea floor at nine to 18 fathoms in the central section; the deadly sea wasps appear in the waters above the prawns and apparently feed on them. Saucer scallop beds have also been exploited.

The influence of man

The Great Barrier Reef also attracts an increasing number of tourists. Green Island, off Cairns in northern Queensland, has a complex underwater observatory and is but one of the high islands and cays that have been developed. Cruises set out from the coastal cities, with landings on the remoter reefs, and big-game fishing for 1,000-pound and larger black marlin, as well as swordfish and barracuda, is also popular. Other potential resources of the Great Barrier Reef include the large quantities of lime and quartz sands. Although petroleum companies have made surveys and preliminary drillings in the reef area, operations were in the doldrums in the early 1970s while a royal commission enquired into possible effects of further oil development on the complex ecology of the reef.

The controversy over oil drillings was but one facet of a growing concern for the future of the Great Barrier Reef. The advent of modern communications had opened up the region for the pursuit of both pleasure and profit, and it was becoming increasingly apparent by the 1970s that some form of regulation of man's activity would become necessary if the reef's fine natural endowment was to be preserved for future generations.

BIBLIOGRAPHY. W.G.H. MAXWELL, *Atlas of the Great Barrier Reef* (1968), a well-illustrated review with much original sedimentological work included; C.M. YONGE, *A Year on the Great Barrier Reef* (1930), a popular account of a scientific expedition; D. HILL, *"The Great Barrier Reef,"* in G.M. BADGER (ed.), *Captain Cook, Navigator and Scientist*, pp. 70–86 (1970), a chapter reviewing scientific knowledge of the Great Barrier Reef; BRITISH MUSEUM (NATURAL HISTORY), *Scientific Reports of the Great Barrier Reef Expedition 1928–29*, vol. 1 (1930–), authoritative, scientific accounts.

(D.Hi.)

Great Lakes

The Great Lakes—Superior, Michigan, Huron, Erie, and Ontario—form part of the St. Lawrence River system of east central North America and are one of the great natural features of that continent and the globe. Although Lake Baikal in the Soviet Union has a greater volume of water, the combined area of the Great Lakes—94,560 square miles (245,000 square kilometres)—represents the largest surface of fresh water in the world, covering an area larger than the United Kingdom, Uganda, or Romania. Their drainage basin of 291,100 square miles (753,950 square kilometres) extends approximately 690 miles from north to south and about 860 miles from Lake Superior in the west to Lake Ontario in the east; it covers an area larger than France or Afghanistan. Except for Lake Michigan, the lakes provide a natural border between Canada and the United States, a frontier that was stabilized by a boundary-waters treaty of 1909. It is a source of pride for both countries that there are no fortifications or warships along the boundary.

Individually, the lakes rank among the 15 largest in the world (see Table). They played a central role in the European colonization and development of North America, have continued to attract people and industry, and are now ringed with large urban concentrations. The lakes have not benefitted from this development, however, and by the 1960s they were suffering greatly from pollution. Lake Erie was pronounced "dead" in 1960, and it has been projected that Lake Michigan might well share the same fate. Concern over the fate of the lakes reached a high pitch in the 1970s, with both governments and individuals investigating plans for reversing the tragic consequences of years of misuse of the lakes' waters. (For related articles, see SAINT LAWRENCE RIVER; SAINT LAWRENCE SEAWAY; NIAGARA RIVER AND FALLS.)

Areas and Volumes of the Great Lakes

	surface area		world rank	volume		world rank
	sq mi	sq km		cu mi	cu km	
Superior	31,700	82,100	2nd	2,935	12,230	4th
Michigan	22,300	57,750	6th	1,180	4,920	6th
Huron	23,100	59,830	5th	849	3,540	7th
Erie	9,910	25,670	11th	116	483	15th
Ontario	7,550	19,550	14th	393	1,640	11th

HISTORY OF EXPLORATION AND RESEARCH

The Great Lakes have been an integral part of the exploration and development of the North American continent. A broken sword, an axe, and a shield boss found near Lake Nipigon, Ontario, and the rune stone at Kensington, Minnesota, have been cited as evidence of early Viking exploration of the region, although the authenticity of these artifacts has not been established. French exploration of the region commenced in 1535, when the explorer Jacques Cartier travelled up the St. Lawrence River to the site of modern Montreal in his search for a route to the Orient. The Huron Indians told him of the great seas lying beyond, but the upper St. Lawrence and Lake Ontario were controlled by the Iroquois, who were not friendly to the Europeans. Consequently, further exploration by another leading French explorer of North America, Samuel de Champlain, followed the course of the Ottawa River, Lake Nipissing, and the French River to Georgian Bay. He reached Lake Huron in 1615 and is credited with being the first European to see the Great Lakes. In 1634 Jean Nicolet—dispatched by Champlain to seek a route to China—led an expedition into Lake Michigan and down the length of Green Bay to the Fox River, where he encountered the Winnebago Indians. Other French explorers, including Robert Cavelier, sieur de La Salle, explored the lakes, made peace with the Indians, and established early settlements.

French exploration

Three major conflicts affected the history and development of the Great Lakes. The French and Indian War (1754–63)—a struggle between the French and British to gain control of rich fur-producing lands—concluded with the cession of Canada to England. The main consequences of the American Revolution (1775–83) to the Great Lakes region were the migration of thousands of Loyalists to New Brunswick, Nova Scotia, and Ontario and the establishment of the international boundary between the United States and Canada. During the War of 1812, Lake Erie was the site of a major naval battle.

Although several of the early explorers recorded observations of short-term fluctuations in water levels in early geographical writings, the first purely scientific expedition on the Great Lakes was not carried out until 1848. It was led by the Swiss naturalist Louis Agassiz and concentrated on studies of the north shore of Lake Superior. Water-level gauges were established in all the lakes by 1860, and all waters were charted by 1882. Studies of plant and animal life began in the 1870s, and the first study of lake currents was conducted in the early 1890s. Interest in lake fisheries was strong throughout the early 20th century, and, since the 1960s, research on the changes in plant and animal life wrought by pollution has increased considerably.

GEOLOGY

The age of the Great Lakes is still not definitely determined. Estimates range from 7,000 to 32,000 years of age. It is generally accepted that Lake Erie reached its present level about 10,000 years ago, Lake Ontario about 7,000 years ago, and Lakes Huron, Michigan, and Superior around 3,000 years ago.

The present configuration of the Great Lakes Basin is the result of the movement of massive glaciers through the midcontinent, a process that began during the Pleistocene Epoch, about 1,000,000 years ago. Studies of Lake Superior indicate that a river system and valleys formed by water erosion existed before the Ice Age. The glaciers undoubtedly scoured these valleys, widening and deepening them and changing the drainage of the area.

The last glacial stage in North America is called the Wisconsin Glaciation because it advanced southward to what is now the southern border of the state of Wisconsin. As the ice sheet melted and receded about 18,000 years ago, the first segments of the Great Lakes were created. Lake Chicago, in the southern Lake Michigan Basin, and Lake Maumee, in western Lake Erie and its adjacent lowlands, originally drained southward into the Mississippi River. As the ice retreat continued, Lake Maumee was drained into Lake Chicago through a valley that now contains the Grand River in Michigan. Eventually, drainage to the east and into the Atlantic Ocean was established, first down the Mohawk Valley and then along the course of the upper St. Lawrence River. At one high-water stage, the waters of the Superior, Huron, and Michigan Basin formed the large Lake Algonquin. At the same time, Lake Duluth, in the western Lake Superior Basin, also drained to the Mississippi.

The weight of the ice sheet exerted great pressures on the land mass. As the ice sheet retreated, low-lying areas, such as the region to the east of Georgian Bay, were exposed. About 10,000 years ago, the upper lakes evidently discharged through this area via the Ottawa River Valley, and their levels were substantially reduced. After the weight of the ice was removed, the land began to rise, closing off some outlets and changing the water levels of the lakes. The largest postglacial lake, Nipissing, occupied the basins of Huron, Michigan, and Superior. Drainage through the Ottawa River Valley ceased, and outflow from the upper lakes was established by way of the St. Clair and Detroit rivers into Lake Erie. Uplift continues at about one foot in 100 years; this is evidenced by the drowned river mouths of western Lake Erie.

Retreat of the ice

A wide range of rock types and deposits are found in the Great Lakes because of their great area and glacial origin. The ancient rocks of the Canadian Shield are part of the Superior and Huron basins, while younger sedimentary rocks make up the remainder of the basins. There are limestone outcrops and large deposits of sand and gravel, usually near shore. Glacial clays and organic sediments occur in the deep areas.

THE LACUSTRINE SYSTEM

The lakes drain roughly from west to east, emptying into the Atlantic Ocean. Except for Lakes Michigan and Huron, their altitudes drop with each lake, usually causing a progressively increasing rate of flow.

Lake Superior, bordered by Ontario, Minnesota, Wisconsin, and upper Michigan, is the northernmost and westernmost lake and can be considered the headwater of the system. It is the deepest lake (mean depth 487 feet), lies at an altitude of 600 feet above sea level, and discharges into Lake Huron through the St. Marys River at an average rate of 74,200 feet per second.

Lake Michigan lies directly south of Lake Superior and is bordered by upper and lower Michigan, Wisconsin, Illinois, and Indiana. It has a mean depth of 276 feet. The average water level is 579 feet above sea level, and its waters flow northward into Lake Huron through the Straits of Mackinac at 56,000 cubic feet per second.

Lake Huron lies at the same altitude as and is slightly larger than Lake Michigan. Its mean depth, however, is only 195 feet. It is bounded by Ontario and Michigan. The average outflow is 177,500 cubic feet per second through the St. Clair River, the shallow basin of Lake St. Clair, and the Detroit River to Lake Erie.

Lake Erie is bordered by Ontario, lower Michigan, Ohio, Pennsylvania, and New York. It is the shallowest of the Great Lakes, with a mean depth of 58 feet. The basin slopes from west to east with depths of 24 feet and 210 feet, respectively. It lies at an altitude of 570 feet, and its waters discharge at an average flow of 194,300 cubic feet per second. The course of the outflow is along the Niagara River and includes a rapid plunge over Niagara Falls before the waters reach Lake Ontario.

Lake Ontario is the smallest of the system. It has, however, the second greatest mean depth, 283 feet. It lies between Ontario and New York, at an altitude of 245 feet, and discharges into the St. Lawrence River at a rate of 233,000 cubic feet per second. It flows for about 750 miles until it empties through the Strait of Gaspé (Détroit de Gaspé) into the Gulf of St. Lawrence.

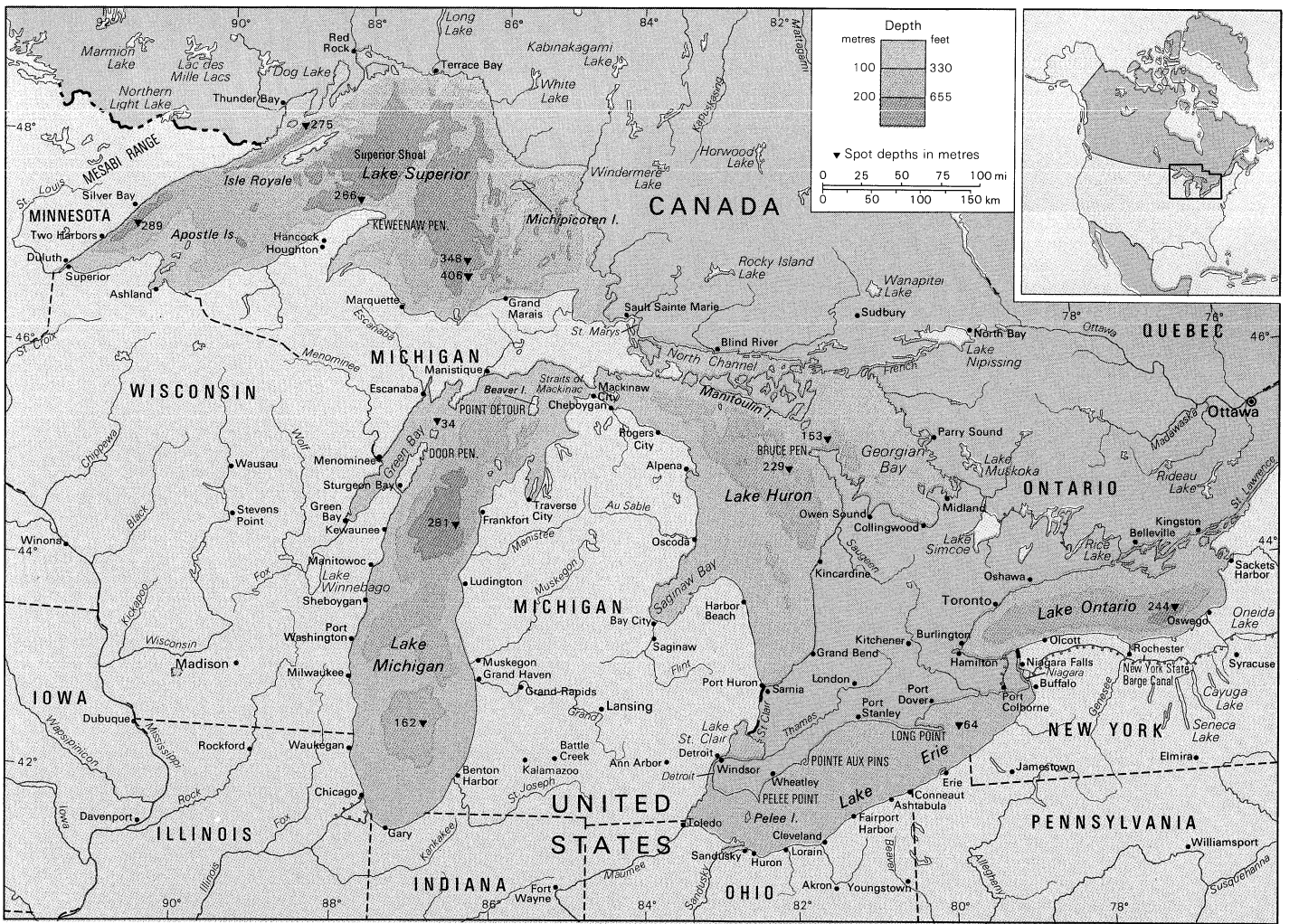
PHYSICAL AND CHEMICAL CHARACTERISTICS

The lakes ultimately receive their water supply from precipitation, which increases from west to east. The average annual precipitation in the Lake Superior Basin is 29 inches, in Lakes Huron and Michigan it is 31 inches, and in Lakes Erie and Ontario it is 34 inches. About two-thirds of the annual precipitation is lost by evaporation: 22 inches on Lake Superior; 35 inches on Lake Erie (Erie receives most of its water from Lake Huron); and 26 inches on Lakes Huron, Michigan, and Ontario. Some water enters Lake Superior from the Hudson Bay drainage system via the Long Lake–Ogoki River diversion, while water is taken out of Lake Michigan by the Chicago Sanitary and Ship Canal.

Effects on climate

The lakes greatly modify the climate of the surrounding region. They absorb large quantities of heat in the warmer months, which are lost to the atmosphere during the colder months. This causes cooler summers and warmer winters. Precipitation is substantially higher along the eastern shores of the lakes, creating the snow belt that afflicts Erie, Pennsylvania; Buffalo, New York; and similarly situated cities.

Lake levels vary about one to two feet throughout the year, the highest levels occurring in summer and the lowest in late winter and early spring. There are small tides of around two inches, but they are relatively unimportant. Seiches—harmonic oscillations of the lakes—


Depth
metres
feet
100
330
200
655
▼ Spot depths in metres
0 25 50 75 100 mi
0 50 100 150 km
Lake Superior
Lake Michigan
Lake Huron
Lake Erie
Lake Ontario
CANADA
UNITED STATES
MINNESOTA
WISCONSIN
MICHIGAN
ILLINOIS
INDIANA
OHIO
IOWA
ONTARIO
QUEBEC
NEW YORK
PENNSYLVANIA
Georgian Bay
Lake Nipissing
Lake Simcoe
Lake Muskoka
Rice Lake
Lake St. Clair
Pelee I.
Michipicoten I.
Isle Royale
Apostle Is
Superior Shoal
KEWEENAW PEN.
DOOR PEN.
POINT DETOUR
BRUCE PEN.
POINTE AUX PINS
PELEE POINT
LONG POINT
Straits of Mackinac
North Channel
Manitoulin I.
Beaver I.
Rocky Island Lake
Wanapitei Lake
Windermere Lake
Horwood Lake
Kabinakagami Lake
White Lake
Dog Lake
Lac des Mille Lacs
Marmion Lake
Northern Light Lake
Long Lake
Lake Winnebago
Green Bay
Saginaw Bay
Georgian Bay
Thunder Bay
Terrace Bay
Red Rock
Sault Sainte Marie
Sudbury
North Bay
Blind River
Parry Sound
Owen Sound
Collingwood
Midland
Kincardine
Oshawa
Toronto
Hamilton
Niagara Falls
Buffalo
Rochester
Syracuse
Oswego
Kingston
Belleville
Sackets Harbor
Ottawa
Madawaska
Rideau Lake
St. Lawrence
Oneida Lake
Cayuga Lake
Seneca Lake
New York State Barge Canal
Port Colborne
Port Dover
Port Stanley
Port Huron
Sarnia
London
Grand Bend
Kitchener
Burlington
Olcott
Jamestown
Elmira
Williamsport
Conneaut
Ashtabula
Fairport Harbor
Cleveland
Lorain
Toledo
Sandusky
Huron
Akron
Youngstown
Erie
Windsor
Wheatley
Detroit
Ann Arbor
Battle Creek
Kalamazoo
Benton Harbor
Chicago
Gary
Fort Wayne
Milwaukee
Madison
Port Washington
Sheboygan
Manitowoc
Green Bay
Sturgeon Bay
Menominee
Escanaba
Marquette
Grand Marais
Houghton
Hancock
Calumet
Ironwood
Duluth
Superior
Two Harbors
Silver Bay
Ashland
Wausau
Stevens Point
Kewaunee
Manistique
Mackinaw City
Cheboygan
Rogers City
Alpena
Oscoda
Au Sable
Traverse City
Frankfort
Manistee
Muskegon
Grand Haven
Grand Rapids
Lansing
Saginaw
Bay City
Harbor Beach
Ludington
Escanaba
Waukegan
Rockford
Dubuque
Davenport
Rock Island
Moline
Peoria
Joliet
Gary
Fort Wayne
Maumee
Sandusky
Lake Erie
Allegheny
Susquehanna
Beaver
Thames
Detroit
St. Clair
St. Joseph
Kankakee
Illinois
Fox
Wisconsin
Rock
Iowa
Mississippi
Wapsipinicon
Chippewa
Black
Menominee
Escanaba
Manistique
Wolf
Kickapoo
Pigeon
Sturgeon
Nipigon
Long Lake
Ogoki Reservoir
Matagami
Lake Abitibi
Lake Timiskaming
Lake Nipissing
French
Ottawa
Madawaska
St. Lawrence
Rideau Lake
Cayuga Lake
Seneca Lake
Oneida Lake
Susquehanna
Allegheny
Beaver
Sandusky
Maumee
Thames
Detroit
St. Clair
Lake Erie
Lake Ontario
Lake Huron
Lake Michigan
Lake Superior
Lake Simcoe
Lake Muskoka
Rice Lake
Georgian Bay
Saginaw Bay
Green Bay
Lake Winnebago
Lake Nipissing
Lake Abitibi
Lake Timiskaming
Long Lake
Ogoki Reservoir
Matagami
Nipigon
Sturgeon
Pigeon
Kickapoo
Wolf
Black
Chippewa
Menominee
Escanaba
Manistique
Wisconsin
Rock
Iowa
Mississippi
Wapsipinicon
Illinois
Fox
Kankakee
Joliet
Peoria
Moline
Rock Island
Davenport
Dubuque
Waukegan
Rockford

are caused by such atmospheric disturbances as winds or differences in barometric pressure. They have resulted in a piling up of water on one side, or end, of the lakes and have caused differences in the water level between Buffalo and Toledo, Ohio, on Lake Erie, as high as 13.5 feet. Currents are highly variable; they respond quickly to wind changes; their direction is determined by the rotation of the earth and the shape of the lake basins.

The Great Lakes have bicarbonate waters, the alkalinity of which ranges from 46 parts per million of carbonates in Lake Superior to 110 parts per million in Lake Michigan. Because Lake Huron is fed by both Superior and Michigan, its chemical content lies in the middle of the range. Alkalinity then increases as the waters flow into and through Lake Erie (95 parts per million) and into Lake Ontario (102 parts per million).

The overall chemical composition of the lakes does not differ greatly from that of other large bodies of fresh water. Limestone in the Lake Michigan Basin supplies large amounts of calcium and magnesium to the system, while sodium concentrations are greater than those of magnesium in Erie and Ontario. Although chemical distribution is relatively uniform in any one lake, concentrations of phosphorus and nitrogen are greatest along the shores, in bays and harbours, and especially near urban centres.

During the 20th century, concentrations of most chemicals have increased significantly in all the lakes except Superior. Chloride, sodium, and sulfate have increased significantly in Lakes Erie, Michigan, and Ontario. Chloride concentrations have increased almost four times over levels reported in 1900, and limited data for Lake Erie indicate that nitrogen concentrations increased five times and phosphorus threefold in 30 years. Especial importance is attached to these nutrients because they stimulate growth of algae.

PLANT AND ANIMAL LIFE

Algae and crustaceans. Diatoms—microscopic algae with glasslike shells of silica—are the major forms of algae, although green and blue-green algae are abundant during the summer in Lakes Erie, Ontario, and Michigan.

Microscopic life

Copepods and cladocerans, microscopic crustaceans, are important in the animal forms of plankton. Most abundant during the spring months in the upper lakes, plankton reaches two peaks of abundance—spring and fall—in the lower lakes and in the more productive waters of the upper lakes.

The organisms living on the bottom in shallow waters are the same kinds of snails, clams, worms, mayflies, and caddis flies found in most small lakes. The deep waters, however, are the realm of some organisms that are found only in the deep, cold lakes of the northern latitudes. These include the delicate opossum shrimp, the deepwater scud (a crustacean), two types of copepods, and the deepwater sculpin (a spiny, large-headed fish).

Fish. The fish community of the lakes and tributaries includes representatives of most families of North American fishes. Lake trout, whitefish, and lake herring have always been important in the lakes, while perch, pike perch, bass, and catfish are abundant in the shallow, warmer waters. The fish populations have changed drastically over the past century. By 1880, the damming and pollution of tributaries had caused the elimination of the Atlantic salmon and the restriction of the whitefish. At the same time, the lake sturgeon was purposely overfished, and the carp was introduced. Smelt entered Lake Michigan in 1927 and in a short time had spread throughout the lakes. The predatory sea lamprey (an eel) migrated into the upper lakes and established spawning populations in the 1930s, causing the collapse of the lake trout populations in Lakes Huron and Michigan by the early 1950s. Other large predators were also drastically reduced by the sea lamprey. Consequently, when the alewife migrated into the upper lakes between 1931 and 1954, it met little competition and predation and soon became the most abundant species. Alewives are of little commercial value, however, and cause costly sanitary operations when, periodically, millions of them die and are washed up to rot on the beaches. During the 1960s, lake trout were reintroduced when it appeared that the sea lamprey was under control. Coho and chinook salmon were also introduced, both to provide a sports fishery and to control the alewife.

The changing fish population

Bird life. Herring and ring-billed gulls and terns are the most common birds; small islands are important nesting grounds for them. The lakes are important as wintering areas for ducks such as the scaups and the old-squaw duck, and a diversity of shorebirds and songbirds migrate through the region during the spring and fall. Various places along the shoreline, such as Pelee Point in Lake Erie, are favoured locations for birdwatching.

RESOURCES AND THEIR DEVELOPMENT

Early interest in the lakes was stimulated by the easy transportation route that they offered into the heartland of the continent. The value of the extensive forests and fertile land in the region was soon realized, and lumbering and agriculture became important. Large coalfields and deposits of iron, copper, limestone, and other minerals were found along or near the extensive shorelines. The combination of these vast resources with a plentiful water supply naturally favoured the development of huge industries and large metropolitan areas around the Great Lakes. The population of these increased from around 300,000 at the start of the 19th century to around 37,000,000 by 1970. Present population growth supports the speculation that a large megalopolis ultimately will extend from Milwaukee and Chicago around southern Lake Michigan, across the state of Michigan to Detroit, and along the southern shore of Lake Erie and will include the Toronto–Hamilton area on Lake Ontario.

Transportation. About 246,000,000 net tons of shipping were moved annually by Great Lakes vessels in the late 1960s. Most of the total included iron ore, coal, and grain for lake ports, but about 14,000,000 net tons were shipped overseas through the St. Lawrence Seaway. The Welland Canal allows passage around Niagara Falls from Lake Ontario into Lake Erie, and the channels and locks in the St. Marys River have made Lake Superior accessible to ships up to 730 feet in length.

Forestry and agriculture. Although the virgin pine forests were felled by 1910, the growing of timber is important and is supported by both federal and state governments. The counties bordering the lakes in the United States have about half of their lands in farms, and over 5 percent of the value of U.S. agriculture is produced in the lakes region. Bordering Canadian counties have roughly 30 percent of their lands under agriculture, producing about 10 percent of the total agricultural value.

Minerals. The iron ranges around Lake Superior—such as the Mesabi in Minnesota—were once a major source of iron for the United States. Peak production occurred in 1953, when almost 100,000,000 net tons were produced. The large deposits of rich ores have since been depleted, but low-grade taconite ores can now be efficiently processed. Lake Superior's Keweenaw Peninsula is a major source of copper, although sources outside of the lakes are relatively more important.

Industry and power. The industry of the lakes region is highly diversified. Perhaps the more important are the large steel mills in Illinois, Indiana, Ohio, and Ontario and the automobile industry centred in the Detroit area.

Commercial fishing was once a major industry on the lakes, but the decline of the more desirable species led to its collapse. Emphasis has switched to developing major sports fisheries based on coho and chinook salmon, lake trout, and rainbow trout.

Fisheries

The value of the lakes for a broad spectrum of recreational activities is inestimable. Powerboating and sailing have become major activities. Many miles of sandy beaches stretch along the lake shores. State, federal, and county lands offer many camping, picnicking, and park areas for a thriving tourist industry.

Of major importance is the water supply that the lakes provide for industries and for about 240 municipalities, which use almost 16,000,000,000 gallons per day. Hydroelectric generating stations on the St. Marys, Niagara,

and St. Lawrence rivers have a total capacity of about 8,000,000 kilowatts. Over 30 large thermal-power plants around the lakes use their waters for cooling. The total capacity of these plants is about 20,000,000 kilowatts.

THE FUTURE

The multiple uses of the lakes are often conflicting. The shipping and hydroelectric-power industries favour higher water levels, but shoreline-property owners find that high levels increase erosion of the shoreline. Conservationists believe that diversion of treated sewage away from the lakes, such as is now done at Chicago, is the best solution for maintaining the quality of the lake waters. Regardless of the approach that is taken, it is obvious that changes have occurred in the plant and animal life of the lakes and that these changes are related to increases in the chemical content, depletion of dissolved oxygen in some areas, and the accumulation of sewage sludge. Significant changes started around 1900 and paralleled the buildup in human population. Misuse of the lakes can be seen at a number of beaches closed due to pollution, but the less obvious deterioration in water quality is more dangerous. It affects either directly or indirectly all uses—raising the cost of water treatment, killing valuable fish species, and adversely affecting the tourist industry and the value of shoreline property. Unless every effort is made to control pollution, more areas will follow the deterioration of Lake Erie, and a major resource of North America will be spoiled.

BIBLIOGRAPHY. A.M. BEETON and D.C. CHANDLER, "The St. Lawrence Great Lakes," in D.G. FREY (ed.), *Limnology in North America*, pp. 535–558 (1963), a summary of the physical, chemical, and biological characteristics of the Great Lakes and a historical account of early research up to about 1960; H.H. HATCHER and E.A. WALTER, *A Pictorial History of the Great Lakes* (1963), an account of early exploration and settlement of the Great Lakes region, includes an interesting section on shipping; J.L. HOUGH, *Geology of the Great Lakes* (1958), detailed information on the origin of the lakes.

(A.M.B.)

Great Plains

A major physiographic province of North America, the Great Plains lie between the Rio Grande on the south and the delta where the Mackenzie River empties into the Arctic Ocean on the north, between the Central Lowland of the United States and the Canadian Shield on the east and the Rocky Mountains on the west. Their length is some 3,000 miles, their width from 300 to 700 miles, and their area approximately 1,125,000 square miles (2,900,000 square kilometres), roughly equivalent to one-third of the United States. Parts of ten states of the United States (Montana, North Dakota, South Dakota, Wyoming, Nebraska, Kansas, Colorado, Oklahoma, Texas, and New Mexico) and the three Prairie Provinces of Canada (Manitoba, Saskatchewan, and Alberta), and portions of the Northwest Territories are within the Great Plains proper. Some writers have used the 100th meridian as the eastern boundary, but a more precise one is an eastward-facing escarpment that runs from Texas to North Dakota, generally somewhat east of the 100th meridian. In the Canadian portion the line dividing the Great Plains from the Canadian Shield runs east of the Red River of the North; cuts through Lake Winnipeg; then curves northwestward, crossing Lake Athabasca, Great Slave Lake, and Great Bear Lake to reach the Arctic Ocean east of the Mackenzie Delta.

Landscape. Once known as the Great American Desert, the Great Plains are a vast high plateau of semi-arid grassland. Their altitude at the base of the Rockies in the United States is between 5,000 and 6,000 feet above sea level; this decreases to 1,500 feet at their eastern boundary. The altitudes of the Canadian portion are lower, and near the Arctic Ocean the surface is a little above sea level. Some sections, such as the Staked Plains in the Texas Panhandle, are extremely flat; elsewhere, tree-covered mountains—the Black Hills of South Dakota and the Bearpaw, Big Snowy, and Judith mountains of Montana—rise 1,500 to 2,000 feet above the general level of the Plains. In the United States the Great Plains are drained by the Missouri River and its great tributaries (the Yellowstone, Platte, and Kansas) and the Arkansas, which flow eastward from the Rockies in broad, steep-sided, shallow valleys.

Surface features, soils, and climates

The soil groups of the Great Plains are correlated with rainfall and grass cover. In the more humid region with heavier grass cover, deep, black soils with much organic matter developed. Sections with less moisture have lighter, shallower soils with less organic matter, while in the most arid regions the soils are even thinner, lighter in colour, and less organic in composition.

The Great Plains have a continental climate. Over much of their expanse, cold winters and warm summers prevail, with low precipitation and humidity, much wind, and sudden variations of temperature. The major source of moisture is the Gulf of Mexico, and the amount falls off both to the north and west. Thus, the southern Plains have 15 to 25 inches of rain annually, the northern Plains 12 to 15 inches, the eastern margin in Nebraska 25 inches, and the western margin in Montana less than 15 inches. The southern parts of the Canadian Plains receive 10 to 20 inches and have a growing season of 70 to 110 days. The growing season averages 240 days in Texas, 120 days at the United States–Canadian boundary, and from 70 to 110 days in the Prairie Provinces. Grasses are the dominant natural plant life, with trees generally confined to river valleys.

Before settlement, the Plains were the home of the great grazing herd mammals—the buffalo and the pronghorn antelope. North of the 54th parallel the grasslands give way to forest, where the moose, woodland caribou, Canada lynx, and timber wolf make their homes.

Human settlement. European immigrants played an important role in settling the Plains. By 1910 foreign white stock (foreign-born and their children) made up 43 percent of the population of the six northern Plains states (Montana, North Dakota, South Dakota, Wyoming, Nebraska, and Kansas), with the British, Germans (many of them from Russia), and Scandinavians the leading ethnic groups. On the southern Plains, peoples of a pre-Columbian stock with Spanish surnames are also important. The Prairie Provinces were settled by British, German Russians (many of them Mennonites), Ukrainians, and Scandinavians.

Many of the immigrants were religious, thrifty, hard-working people with an attachment to the land. Kinship and nationality ties drew the plainsmen together, and they would travel long distances to visit and exchange work. Class differentiation was less and the status ladder shorter than in older parts of North America or Europe.

By 1970 the population of the Plains proper was about 5,000,000 in the United States and about 3,300,0000 in Canada. Although there were few large cities, some 60 percent of the population was urban. The largest cities were Edmonton (440,000 in 1971) and Calgary (400,000) in Alberta; Denver, Colorado (515,000 in 1970); and Lubbock, Texas (150,000). The rural population was also sparse, about 4 per square mile in the United States and in the settled portion of the Canadian section.

Life-styles of the Plains

Ranchers enjoyed their remoteness and looked upon their rangeland as the last remaining trace of the Old West, with its vast expanse of plains and untamed wilderness. Not generally a gregarious kind of people, they were highly individualistic in politics. Farmers, more inclined to social interaction, made economic cooperatives strong on the Plains. In recent years, ranchers and farmers alike have valued horsemanship and rodeos as symbols of a tradition and style of life that evolved from the natural habitat.

The need for larger farms and ranches to produce viable economic units has led to a heavy migration from the rural areas. This exodus has been demoralizing for the farmers, ranchers, and businessmen who remained, and it has made survival more difficult for churches, schools, and the rural trading centres. The low density of population has concentrated services more and more in a few centres, necessitating long trips to attend school and church, to do business, and to find recreation. To meet

these difficulties, some farmers have moved into town to live and commuted back to their land to work, a revival of a centuries-old pattern.

Economic development. Indians on horseback had long exploited the buffalo herds, but in the 1870s cattle replaced the buffalo, and cowboys replaced the Indians. In the 1880s and 1890s farmers began to crowd the ranchers, and wheat began to replace cattle. Settlement came in years of good rains, so the Plains were overpopulated in the first rush; and a heavy emigration followed. Many grain farmers left because their farms were too small and more vulnerable to drought than the cattle ranches. Those who stayed built up the size of their holdings, saved against hard times, and added livestock to grain farming. From 1930 to 1970 the outmigration brought a decline of 31 percent in the rural population of the United States portion of the Great Plains proper, but urban population grew by 166 percent.

Agricultural and mineral bases of the economy

The Great Plains remain basically an agricultural area producing wheat, cotton, sorghum, and hay and raising cattle and sheep. In 1971 seven of the ten leading U.S. wheat states (Kansas, North Dakota, Montana, Nebraska, Colorado, Oklahoma, and South Dakota) lay within the Great Plains; and the Prairie Provinces were leading wheat producers in Canada. Livestock brought the largest cash farm receipts in nine of the ten Plains states. Only in North Dakota did crop income exceed that from livestock.

The Great Plains states also produce much mineral wealth, with Texas leading the nation and three other Plains states (Oklahoma, New Mexico, and Wyoming) ranking in the top twelve. Four of the Great Plains states have the largest coal reserves in the nation (Wyoming, North Dakota, Montana, and Colorado) but rank low in actual production. Texas leads the United States in production of petroleum and natural gas, and eight other Plains states are substantial producers. Alberta leads Canada in petroleum and natural gas.

Prospects. In the 1960s, the urban population gains in the United States portion of the Great Plains proper barely equalled rural losses, and the total Plains population grew only minutely. Population growth in the Prairie Provinces, although far greater in terms of percentage, was also slowing down. In both Canadian and United States portions of the Plains such other trends as urbanization, emphasis on soil and water conservation, larger and fewer farms, town residence for farmers, and expansion of Plains communities will probably continue. Planners for county reorganization, for example, believe that North Dakota could best be served by 8 counties, each with a principal trade centre, instead of the present 53; Nebraska and Saskatchewan each need only 15. By expanding the communities and integrating rural and urban areas, such reorganization, the planners argue, would reduce costs and improve the services of government, provide more specialized business and professional services, and increase cultural opportunities for the widely scattered population.

BIBLIOGRAPHY. The first important account of the region was WALTER PRESCOTT WEBB, *The Great Plains* (1931). Webb's brilliant interpretation on cultural adaptation was extended by CARL F. KRAENZEL in *The Great Plains in Transition* (1955). Indian culture is well described by ROBERT H. LOWIE in *Indians of the Plains* (1954). The classic on white pioneering is MARI SANDOZ, *Old Jules* (1935). The anthropologist JOHN W. BENNETT analyzes present-day trends for a Canadian portion in *Northern Plainsmen: Adaptive Strategy and Agrarian Life* (1969).

(E.B.R.)

Great Salt Lake

Its waters so dense that swimmers are buoyed above the surface, the Great Salt Lake, in northern Utah, is the largest body of saline water in the Western Hemisphere and among the most saline inland bodies of water in the world. Like the Dead Sea, it exists within an arid environment and has chemical characteristics similar to that of the oceans. It has a much greater salinity than the oceans, however, since natural evaporation exceeds the supply of water from the rivers feeding it. This evaporation and changes in the flow of the rivers feeding it have varied its area drastically—from 2,400 square miles (6,200 square kilometres) in 1873 to 950 square miles in 1963 and some 1,600 square miles in the early 1970s. It is generally less than 15 feet (4.5 metres) deep, with a maximum depth of 35 feet.

Character and location

Surrounded by great stretches of sand, salt land, and marsh, the Great Salt Lake remains strangely isolated from the nearby cities, towns, and other human habitations, though in recent years means have been found to turn its apparent sterility to a profit in both economic and recreational terms. It has become important not only as a source of minerals but also as a beach and watersports attraction and a wildlife preserve.

Geological and human history. The Great Salt Lake is the largest of the saline-lake remnants of prehistoric freshwater Lake Bonneville, the others being Bear Lake, on the Utah–Idaho border, and Utah Lake, west of Provo, Utah. Originating some 1,000,000 years ago in the Pleistocene Epoch, Lake Bonneville covered almost 20,000 square miles of western Utah, extending also into present-day Nevada and Idaho. During succeeding glacial periods, large quantities of freshwater entered this intermontane basin and drained out through the Snake River—ultimately into the Columbia River and the Pacific Ocean. During the interglacial and postglacial periods, however, water levels decreased and outflows were cut off. Water, thus, could escape only through evaporation, and the mineral salts from the inflowing rivers remained trapped in the lake.

Exploration and settlement

The lake appeared on 18th-century maps of the continent through reports of explorer-trappers and Indian tales, a semilegendary body alternately named Timpanogos or Buenaventura, depending on the source. The first Westerners whose accounts are fully credited were the trappers Étienne Prevost and Jim Bridger, who came upon the lake independently in 1824–25. More detailed investigations were made by Capt. John C. Frémont in 1843 and 1845. The Mormons' settlement in 1847 of their "promised land," on the nearby site of Salt Lake City, brought the region more fully into national awareness. The lake was surveyed in 1850, and in 1869 the last spike of America's first transcontinental railroad was driven near the lake's northeastern shore. The study of the Great Basin region by the U.S. Geological Survey in 1890 was an important source of information about the lake, and later studies have been led by that agency.

Surface features and chemistry. The lake's basin is defined by the foothills of the Wasatch Range to the north, east, and south and by the Great Salt Lake Desert, a remnant of the bed of Lake Bonneville, to the west. The part of this desert known as the Bonneville Salt Flats has become an automobile raceway, the site of many trials for world land-speed records. The varying shoreline comprises beaches, marshes, and mudflats. The 30-mile-long Lucin Cutoff, a causeway laid down for a rail line in 1959, connects the cities of Ogden and Lucin, splits the lake, and affects the water level. Because the main tributaries enter from the south, the water level of the southern section is several inches higher than that of the northern part. Several islands, the largest of which are Antelope and Fremont (36 and 14 square miles, respectively), lie south of the cutoff.

Saline components

Freshwater enters the lake from the Bear, Weber, and Jordan rivers, which carry in over 1,100,000 tons of salts annually. The total accumulation in the lake basin is 6,000,000,000 tons, mainly sodium chloride, though sulfate, magnesium, and potassium are abundant. Table salt and potash production dates from the 19th century, while magnesium production from brines was begun on a large scale only in 1971. Extraction of other minerals is anticipated.

Natural life. The high salt content makes the lake unviable for all but a few forms of life. A small crustacean, the brine shrimp *Artemia*, which lives in saline lakes throughout the world, is abundant. The larvae of two kinds of salt flies and a few species of protozoans, bacteria, and algae are found. The marshes, mudflats, and islands, however, attract much waterfowl, including peli-

cans, herons, cormorants, terns, and gulls, while Antelope Island has been made a refuge for the bison.

BIBLIOGRAPHY. W.H. BRADLEY, "Paleolimnology," in D.G. FREY (ed.), *Limnology in North America*, pp. 621–652 (1963), considerable information on Lake Bonneville and the Great Salt Lake; C.L. HUBBS and R.R. MILLER, "The Zoological Evidence," in a symposium, "The Great Basin with Emphasis on Glacial and Post-Glacial Times," in *Bull. Univ. Utah, Biol. Ser.*, 10:18–166 (1948), a good source for the zoogeography of the Great Basin, but limited information on the Great Salt Lake; A.J. EARDLEY, V. GVOSDETSKY, and R.E. MARSELL, "Hydrology of Lake Bonneville and Sediments and Soils of Its Basin," *Bull. Geol. Soc. Am.*, 68:1141–1201 (1957), information on the size and outflow of Lake Bonneville; G.K. GILBERT, *Lake Bonneville* (1891), an excellent treatise on the geology of the lake; G.E. HUTCHINSON, *A Treatise on Limnology*, vol. 1 (1957), considerable material on Great Salt Lake scattered throughout, with a good discussion of the chemistry of closed basins; W.B. LANGBEIN, "The Salinity and Hydrology of Closed Lakes," *Prof. Pap. U.S. Geol. Surv.*, no. 412 (1961), information on a number of closed basins; H.C. WHITEHEAD and J.H. FETH, "Recent Chemical Analyses of Waters from Several Closed-Basin Lakes and Their Tributaries in the Western United States," *Bull. Geol. Soc. Am.*, 72:1421–1425 (1961), a good source for the chemistry of the Great Salt Lake.

(A.M.B.)

Greco, El

El Greco was a unique genius whose true greatness as a religious painter and portraitist came first to be recognized in the early part of the 20th century.

Early life and works. El Greco was born in 1541 at Candia, on the island of Crete. He never forgot he was of Greek descent and usually signed his paintings in Greek letters with his full name, Doménikos Theotokópoulos. He is, nevertheless, generally known as El Greco (the Greek), a name he acquired when he lived in Italy, where the custom of identifying a man by designating country or city of origin was a common practice. The curious form of the article (El), however, may be the Venetian dialect or more likely from the Spanish.

Whether El Greco first learned his trade as an icon painter in the Byzantine tradition is disputed. The icon of "St. Luke Painting the Portrait of the Madonna" in the Benaki Museum at Athens is signed in Greek "By the Hand of Domenikos" (Cheìr Doménikou). Greek writers believe that this painter was Theotokópoulos, known as El Greco, but it cannot be assumed that only he among Greek artists of the 16th century had the name Doménikos.

Archivo Mas, Barcelona



"Burial of the Conde de Orgaz," canvas by El Greco, 1586–88. In the church of Santo Tomé, Toledo, Spain. 4.8 m × 3.6 m.

Study in Venice

Since Crete was then a Venetian possession and the young artist a Venetian citizen, he decided to go to Venice to study. The exact year in which this took place is not known; but speculation has placed the date anywhere from 1560, when he was 19, to 1566. In Venice El Greco entered the studio of Titian, who was the greatest painter of the day.

Knowledge of El Greco's years in Italy is limited. A letter of November 16, 1570, written by Giulio Clovio, an illuminator in the service of Cardinal Alessandro Farnese, requested lodging in the Palazzo Farnese for "a young man from Candia, a pupil of Titian." On July 8, 1572, "the Greek painter" is mentioned in a letter sent from Rome by a Farnese official to the Cardinal. Shortly thereafter, on September 18, 1572, "Dominico Greco" paid his dues to the guild of St. Luke in Rome. Just how long the young artist remained in Rome is unknown, since he may have returned to Venice, *c.* 1575–76, before he set forth for Spain. An anecdote told by Guilio Mancini, author of a treatise on artists in Rome, relates that El Greco suggested that if Michelangelo's "Last Judgment" in the Sistine Chapel were destroyed he (El Greco) "would have made it with honesty and decorum." Mancini reported that El Greco was forced to leave Rome because of this criticism. Such picturesque gossip is characteristic of writings of the period, however, and there is no reason to accept it literally. Besides, Michelangelo's fresco was under general attack at this time because of the nudity of the figures.

Another anecdote, first published in 1923, proved to be a literary forgery. It holds that Giulio Clovio recorded in his diary a visit to El Greco's studio in Rome, where he found the artist on a sunny day with the curtains drawn. The young Greek is quoted as saying that the light of day destroyed his inner light. No diary of Giulio Clovio exists in the library at Split, in Yugoslavia, where is was falsely claimed to have been found.

To the artist's first years in Italy has been attributed a small, portable altar of three panels in the museum at Modena, Italy, usually referred to as the "Modena Triptych." Signed "Hand of," in the manner widely used by all Greek artists of the 16th century, it also has only a first name, Doménikos. Finally, El Greco's authorship is dubious because the altar is clearly a Veneto-Cretan work that combines Greek and Venetian elements in a way frequently practiced by many artists who had been exposed to both artistic traditions.

The certain works painted by El Greco in Italy are completely in the Venetian Renaissance style of the 16th century. They show no effect of his Byzantine heritage except possibly in the faces of old men; for example, in the "Christ Healing the Blind." The placing of figures in deep space and the emphasis on an architectural setting in High Renaissance style are particularly significant in his early pictures, such as "Christ Cleansing the Temple"; a slightly later version is in the Minneapolis (Minnesota) Institute of Arts. In his use of rich colour he reflects the traditions of Titian, but he had also clearly studied the works of Paolo Veronese, Jacopo Bassano, and Tintoretto. During his sojourn in Rome, El Greco indubitably admired the painting and sculpture of Michelangelo, as demonstrated by his "Pietà," in which the composition is based upon Michelangelo's sculpture of the same subject, now in the cathedral at Florence.

The first evidence of El Greco's extraordinary gifts as a portraitist appears in Italy in a portrait of a friend, "Giulio Clovio." The arrangement of a half-length figure silhouetted against a plain background that has an open window revealing a distant landscape is known in the works of Titian and other Venetians. In the portrait of Vincentio Anastagi, El Greco employed the full-length

figure, with brilliantly painted armour against a red curtain. Again, this type of composition has ample precedent in the work of Titian and north Italian masters.

Middle years. El Greco first appeared in Spain in the spring of 1577, initially at Madrid, later in Toledo. One of his main reasons for seeking a new career in Spain must have been knowledge of Philip II's great project, the building of the monastery of San Lorenzo at El Escorial, some 40 miles northwest of Madrid. Moreover, the Greek must have met important Spanish churchmen in Rome through Fulvio Orsini, a humanist and librarian of the Palazzo Farnese. It is known that at least one Spanish ecclesiastic who spent some time in Rome at this period—Luis de Castilla—became El Greco's intimate friend and was eventually named one of the two executors of his last testament. Luis' brother, Diego de Castilla, gave El Greco his first commission in Spain, which possibly had been promised before the artist left Italy.

Move to Spain

In 1578 Jorge Manuel, the painter's only son, was born at Toledo, the offspring of Doña Jerónima de Las Cuevas. She appears to have outlived El Greco, and, although he acknowledged both her and his son, he never married her. That fact has puzzled all writers, since he mentioned her in various documents, including his last testament. It may be that El Greco had married unhappily in his youth in Crete or Italy and therefore could not legalize another attachment.

For the rest of his life El Greco continued to live in Toledo, busily engaged on commissions for the churches and monasteries there and in the province. He became a close friend of the leading Humanists, scholars, and churchmen. Antonio de Covarrubias, a classical scholar and son of the architect Alonso de Covarrubias, was a friend whose portrait he painted. Fray Hortensio Paravicino, the head of the Trinitarian order in Spain and a favourite preacher of Philip II of Spain, dedicated four sonnets to El Greco, one of them recording his own portrait by the artist. Luis de Góngora y Argote one of the major literary figures of the late 16th century, composed a sonnet to the tomb of the painter. Another writer, Don Pedro de Salazar de Mendoza, figured among the most intimate circle of El Greco's entourage.

The inventories compiled after his death confirm the fact that he was a man of extraordinary culture—a true Renaissance Humanist. His library, which gives some idea of the breadth and range of his interests, included works of the major Greek authors in Greek, numerous books in Latin, and others in Italian and in Spanish: Plutarch's *Lives*, Petrarch's poetry, Ariosto's *Orlando Furioso*, the Bible in Greek, the proceedings of the Council of Trent, and architectural treatises by Vitruvius, Vignola, Alberti, Palladio, and Serlio. El Greco himself prepared an edition of Vitruvius, accompanied by drawings, but the manuscript is lost.

In 1585 and thereafter El Greco lived in the large, late-medieval palace of the Marqués de Villena. Although it is near the site of the now-destroyed Villena Palace, the museum in Toledo called the Casa y Museo del Greco was never his residence. It can be assumed that he needed space for his atelier more than for luxurious living. In 1605 the palace was listed by the historian Francisco de Pisa as one of the handsomest in the city; it was not a miserable ruined structure, as some romantic writers have presumed. El Greco surely lived in considerable comfort, even though he did not leave a large estate at his death.

First commission in Spain

El Greco's first commission in Spain (given to him by Diego de Castilla) was for the high altar and the two lateral altars in the conventual church of Santo Domingo el Antiguo at Toledo (1577–79). Never before had the artist had a commission of such importance and scope. Even the architectural design of the altar frames, reminiscent of the style of the Venetian architect Andrea Palladio, was prepared by El Greco. The painting for the high altar, "Assumption of the Virgin," also marked a new period in the artist's life, revealing the full extent of his genius. Though in a general way inspired by Titian's "Assumption" (1516–18; Sta. Maria dei Frari, Venice), it is majestic and solemn rather than triumphant. The figures are brought close into the foreground, and in the Apostles a new brilliance of colour is achieved. The technique remains Venetian in the laying on of the paint and in the liberal use of white highlights; yet the intensity of the colours and the manipulation of contrasts, verging on dissonance, is distinctly El Greco. For the first time the importance of his assimilation of the art of Michelangelo comes to the fore, particularly in the painting of the "Trinity," in the upper part of the high altar (now in the Prado, Madrid), where the powerful sculpturesque body of the nude Christ leaves no doubt of the ultimate source of inspiration. In the lateral altar painting of the "Resurrection," the poses of the standing soldiers and the contrapposto (a position in which the upper and lower parts of the body are contrasted in direction) of those asleep are also clearly Michelangelesque in inspiration.

At the same time, El Greco created another masterpiece of extraordinary originality—the "Espolio" ("Disrobing of Christ"). In designing the composition vertically and compactly in the foreground he seems to have been motivated by the desire to show the oppression of Christ by his cruel tormentors. He chose a method of space elimination that is common to middle- and late-16th-century Italian painters known as Mannerists, and at the same time he probably recalled late Byzantine paintings in which the superposition of heads row upon row is employed to suggest a crowd. The original altar of gilded wood that El Greco designed for the painting has been destroyed, but his small sculptured group of the "Miracle of St. Ildefonso" still survives on the lower centre of the frame.

Elongation of the human figure

El Greco's tendency to elongate the human figure becomes more notable at this time; for example, in the handsome and unrestored "St. Sebastian," which was inspired in the pose by Michelangelo's statue of "Victory" (1532–34; Palazzo Vecchio, Florence). The same extreme elongation of body is also present in Michelangelo's work, in the painting of the Venetians Tintoretto and Paolo Veronese, and in the art of the leading Mannerist painters. Writers unfamiliar with Italian art of the 16th century have proposed erroneously that El Greco elongated the human figure because he had astigmatism. The increased slenderness of Christ's long body against the dramatic clouds in "Crucifixion with Donors" foreshadows the artist's late style.

El Greco's connection with the court of Philip II was brief and unsuccessful, consisting first of the "Allegory of the Holy League" ("Dream of Philip II") in the Escorial (1578–79) and second of the "Martyrdom of St. Maurice" (1580–82). The latter painting did not please the King, who promptly (1584) ordered another work of the same subject to replace it. Thus ended the great artist's connection with the Spanish court.

The "Martyrdom of St. Maurice" is one of the artist's most notable paintings, despite the fact that Philip II rejected it. The King may have been troubled by the almost shocking brilliance of the yellows as contrasted to the ultramarine in the costumes of the main group, which includes St. Maurice in the centre. On the other hand, to the modern eye El Greco's daring use of colour is particularly appealing. The brushwork remains Venetian in the way the colour suggests form and in the free illusionistic and atmospheric creation of space. This essentially optical method of painting, preferably called illusionism rather than impression, was a creation of Titian in his mature years, and thereafter it was adopted by the Venetian school in general.

The "Burial of Conde de Orgaz"

The "Burial of the Conde de Orgaz" (1586–88) is universally regarded as El Greco's masterpiece. The supernatural vision of Gloria (Heaven) above and the impressive array of portraits represent all aspects of this extraordinary genius' art. El Greco clearly distinguished between heaven and earth: above, heaven is evoked by swirling icy clouds, semi-abstract in their shape, and the saints are tall and phantomlike; below, all is normal in the scale and proportions of the figures. According to the legend, SS. Augustine and Stephen appeared miraculously to lay the Conde de Orgaz in his tomb as a reward for his generosity to their church. In golden and red vestments

they bend reverently over the body of the Count, who is clad in magnificent armour that reflects the yellow and reds of the other figures. The young boy at the left is El Greco's son, Jorge Manuel; on a handkerchief in his pocket is inscribed the artist's signature and the date 1578, the year of the boy's birth. The men in contemporary 16th-century dress who attend the funeral are unmistakably prominent members of Toledan society.

El Greco's Mannerist method of composition is nowhere more clearly expressed than here, where all of the action takes place in the frontal plane. In contrast, Raphael, a classical High Renaissance master, disposed his figures throughout a series of parallel planes in deep space, as, for example, in the "Disputa" and the "School of Athens" (1508–11; Vatican, Rome).

Later life and works. *Religious paintings.* From 1590 until his death in 1614 El Greco's output was prodigious. His pictures for the churches and convents of the Toledan region include the "Holy Family with the Magdalen" and the "Holy Family with St. Anne." He repeated several times the "Agony in the Garden," in which a supernatural world is evoked through strange shapes and brilliant, cold, clashing colours. The devotional theme of "Christ Carrying the Cross" is known in 11 originals, many copies, and some modern forgeries. El Greco depicted most of the major saints, often repeating the same composition: St. Dominic, Mary Magdalen, St. Jerome as cardinal, St. Jerome in penitence, and St. Peter in tears. St. Francis, however, was by far the saint most favoured by the artist; it is no wonder that Francisco Pacheco, the Spanish painter and art theorist who visited El Greco in 1611, declared him the best interpreter of the founder of the Franciscan order. About 25 originals representing St. Francis survive and in addition more than 100 pieces by followers and as many modern forgeries. The most popular of several types was "St. Francis and Brother Leo Meditating on Death," of which there is a fine example in the National Gallery of Canada at Ottawa.

Two major series ("Apostolados") survive representing Christ and the Twelve Apostles in 13 canvases: one in the sacristy of Toledo Cathedral (1605–10) and another, unfinished set (1612–14) in the Casa y Museo del Greco at Toledo. The frontal pose of the Christ blessing in this series suggests a medieval Byzantine figure, although the colour and brushwork are El Greco's personal handling of Venetian technique.

El Greco's artistic individuality

In these works the devotional intensity of mood reflects the religious spirit of Catholic Spain in the period of the Counter-Reformation. Although Greek by descent and Italian by artistic preparation, the artist became so immersed in the religious environment of Spain that he became the most vital visual representative of Spanish mysticism. Yet, because of the combination of these three cultures, he developed into an artist so individual that he belongs to no conventional school but is a lonely genius of unprecedented emotional power and imagination.

Several major commissions came El Greco's way in the last 15 years of his life: three altars for the Chapel of San José, Toledo (1597–99); three paintings (1596–1600) for the Colegio de Doña María de Aragon, an Augustinian monastery in Madrid; the high altar, four lateral altars, and the painting "St. Ildefonso" for the Hospital de la Caridad at Illescas (1603–05). A litigation over payment showed the hospital authorities to be malicious and deceptive, but the documents provide valuable data about the artist's life.

Extreme distortion of body characterizes El Greco's last works; for example, the "Adoration of the Shepherds," painted in 1612–14 for his own burial chapel. The brilliant, dissonant colours and the strange shapes and poses create a sense of wonder and ecstasy, as the shepherd and angels celebrate the miracle of the newly born child. In the unfinished "Fifth Seal of the Apocolypse," El Greco's imagination led him to disregard the laws of nature even more. The gigantic swaying figure of St. John the Evangelist, in abstractly painted icy-blue garments, reveals the souls of the martyrs who cry out for deliverance. In like manner, the figure of the Madonna in the "Immaculate Conception," originally in the church of San Vicente, floats heavenward in a paroxism of ectasy supported by long, distorted angels. The fantastic view of Toledo below, abstractly rendered, is dazzling in its ghostly moonlit brilliance; and the clusters of roses and lilies, symbols of the Virgin's purity, are unalloyed in their sheer beauty.

Landscapes. In his three surviving landscapes El Greco demonstrated his characteristic tendency to dramatize rather than to describe. The "View of Toledo" (*c.* 1595) renders a city stormy, sinister, and impassioned with the same dark, foreboding clouds that appear in the background of his earlier "Crucifixion with Donors." Painting in his studio, he rearranged the buildings depicted in the picture to suit his compositional purpose. "View and Plan of Toledo" (1610–14) is almost like a vision, all of the buildings painted glistening white. An inscription by the artist on the canvas explains quite fancifully that he had placed the Hospital of San Juan Bautista on a cloud in the foreground so that it could better be seen and that the map in the picture shows the streets of the city. At the left, a river god represents the Tagus, which flows around Toledo, a city built on rocky heights. Although El Greco had lived in Italy and in Rome itself, he rarely used such classical Roman motives.

Views of Toledo

The one picture with a mythological subject, so dear to most Renaissance artists, is the "Laocoön." For ancient Troy he substituted a view of Toledo, similar to the one just discussed, and he displayed little regard for classical tradition in painting the highly expressive but great, sprawling body of the priest.

Portraits. Although El Greco was primarily a painter of religious subjects, his portraits, though less numerous, are equally high in quality. Two of his finest late works are the portraits of "Fray Felix Hortensio Paravicino" (1609) and "Cardinal Don Fernando Niño de Guevara" (*c.* 1600). Both are seated, as was customary after the time of Raphael in portraits presenting important ecclesiastics. Paravicino, a Trinitarian monk and a famous orator and poet, is depicted as a sensitive, intelligent man. The pose is essentially frontal; and the white habit and black cloak provide highly effective pictorial contrasts. The Cardinal, in crimson robes, is almost electrical in his inherent energy, a man accustomed to command. El Greco was able to project the domineering personality of this powerful politician, archbishop, and grand inquisitor.

His portrait of "Jeronimo de Cevallos" (1605–1610), on the other hand, is most sympathetic. The work is half-length, painted very thinly and limited to black and white. The huge ruff collar, then in fashion, enframes the kindly face. By such simple means, the artist created a memorable characterization that places him in the highest rank as a portraitist, along with Titian and Rembrandt.

El Greco died on April 7, 1614, and was buried at Toledo in the chapel in Santo Domingo el Antiguo. Because of a dispute with the nuns, his son, Jorge Manuel, transferred the family vault to San Torcuato. Since the destruction of that church a century ago, no sign of the artist's burial place survives.

No followers of any consequence remained in Toledo after El Greco's death. Only his son and a few unknown painters produced weak copies of the master's work. His art was so personally and so highly individual that it could not survive his passing. Moreover, the new Baroque style of Caravaggio and of the Carracci soon supplanted the last surviving traits of 16th-century Mannerism.

MAJOR WORKS

ITALIAN PERIOD: "Christ Cleansing the Temple" (signed, 1560–65; National Gallery of Art, Washington, D.C.); same subject (signed, 1570–75; Minneapolis Institute of Arts, Minnesota); "Boy Lighting a Candle" (1570–75; Museo e Gallerie Nazionali di Capodimonte, Naples); "Christ Healing the Blind" (*c.* 1570; Galleria Nazionale, Parma, Italy); "Giulio Clovio" (signed, *c.* 1570; Museo e Gallerie Nazionali di Capodimonte, Naples); "Pietà" (1574–76; Hispanic Society of America, New York); "Vincentio Anastagi" (signed, *c.* 1575; Frick Collection, New York); "Annunciation" (1575–76; Contini Bonacossi Collection, Florence).

SPANISH PERIOD (RELIGIOUS PAINTINGS, 1577–1644): "Assumption of the Virgin" (signed and dated 1577; Art Institute

of Chicago); "Trinity" (1577–79; Prado, Madrid); "Espolio" (1577–79; Cathedral, Toledo, Spain); "St. Sebastian" (signed, 1577–78; Cathedral, Palencia, Spain); "Crucifixion with Donors" (signed, *c.* 1580; Louvre, Paris); "Martyrdom of St. Maurice" (signed, 1580–82; Museos Nuevos, Escorial); "Christ Carrying the Cross" (1585–90; Metropolitan Museum of Art, New York); "St. Louis of France" (1585–90; Louvre, Paris); "Burial of the Conde de Orgaz" (signed, 1586–88; Santo Tomé, Toledo, Spain); "Holy Family with the Magdalen" (1590–95; Cleveland Museum of Art); "Holy Family with St. Anne" (1590–95; Hospital of San Juan Bautista, Toledo, Spain); "Christ at Gethsemane (1590–95; Toledo Museum of Art, Ohio); "St. Francis Kneeling in Meditation" (1595–1600; M.H. de Young Memorial Museum, San Francisco); "St. Jerome as Cardinal" (1595–1600, Frick Collection, New York); Chapel of San José, Toledo, Spain (1597–99), all signed: High altar, "St. Joseph and the Infant Christ," lateral altars, "St. Martin and the Beggar" and "The Virgin with St. Agnes and St. Tecla" (these two now in the National Gallery of Art, Washington, D.C.). High altar, Colegio de Doña María de Aragon, Madrid (all signed, 1596–1600): "Annunciation" (Museo Balaguer, Villanueva y Geltrú, Spain); "Adoration of the Shepherds" (Art Museum of the Socialist Republic of Romania, Bucharest); "Baptism" (Prado, Madrid). "St. Ildefonso" (1600–03; Hospital de la Caridad, Illescas); "St. Dominic" (1600–05; Cathedral, Toledo, Spain); "St. Francis and Brother Leo Meditating on Death" (1600–05; National Gallery of Canada, Ottawa); Illescas, Hospital de la Caridad (1603–05), altars, sculpture, and paintings; "Agony in the Garden" (1605–10; Santa María, Andújar); "Immaculate Conception" (from San Vicente; 1607–14; Museo de Santa Cruz, Toledo, Spain); "The Fifth Seal of the Apocolypse" (1608–14; Metropolitan Museum of Art, New York), original intended in altar in San Juan Bautista, Toledo, Spain; "Adoration of the Shepherds" (1612–14; Prado, Madrid). (LANDSCAPES): "View of Toledo" (*c.* 1595; Metropolitan Museum of Art, New York); "View and Plan of Toledo" (1610–14); Casa y Museò del Greco, Toledo, Spain); "Laocoön" (1610–14; National Gallery of Art, Washington, D.C.). (PORTRAITS): "Knight Taking an Oath" (1578–80; Prado, Madrid); "Cardinal Don Fernando Niño de Guevara" (*c.* 1600; Metropolitan Museum of Art, New York); "Jerónimo de Cevallos" (1605–10; Prado, Madrid); "Fray Feliz Hortensio Paravicino" (1609; Museum of Fine Arts, Boston).

BIBLIOGRAPHY. MANUEL B. COSSIO, *El Greco*, 2 vol. (1908), the first monograph on the artist, now important as a source; A.L. MAYER, *Dominico Theotocopuli El Greco* (1926), a catalog, but undiscriminating in attributions; ANTONINA VALLENTIN, *El Greco* (1954), an interesting biography; HAROLD E. WETHEY, *El Greco and His School*, 2 vol. (1962), a biography and catalog that separates the master's own work from school pieces, copies, and forgeries; FRANCISCO DE BORJA DE SAN ROMAN, *El Greco en Toledo* (1910), the artist's testament, inventories, and contracts; EDOARDO ARSLAN, "El Greco," *Encyclopedia of World Art*, vol. 6, col. 835–845 (1962), with good bibliography.

(H.E.W.)

Greco-Persian Wars

The Greco-Persian Wars, fought intermittently for a hundred years (*c.* 546–*c.* 448 BC), reached a climax in the period 490–479, when the Persians twice invaded mainland Greece. At that time the empire ruled by the Achaemenian house of Persia (Iran) was at the height of its power. The Greek states beyond its western borders seemed insignificant by comparison; yet a number of them combined so well and fought so skillfully that they not only drove the Persians out of Europe but also liberated the Greek states in Anatolia from Persian rule. Finally they extracted from Persia a treaty of nonaggression, known as the Peace of Callias, about 448. Greek success was of cardinal importance. Independent states survived in Europe and created the cultural forms and the political systems of Greece and Rome, which continued to flourish long after the fall of the Persian Empire was complete.

Expansion of the Persian Empire (559-500 BC). Within 30 years the Persian kings Cyrus and Cambyses created an empire that extended from the Indus Valley to the Aegean Sea and from the Caucasus to Arabia. After the Persians defeated Croesus of Lydia about 546, the small Greek states situated on the Asiatic coast were reduced piecemeal.

In the meantime, Sparta, the strongest state on the Greek mainland, did nothing more than lodge diplomatic protests. Darius, who reigned from 522 to 486, consolidated and extended the Persian Empire. From his capital, far inland at Susa, the royal roads led to about 20 provinces, called satrapies, which were governed by satraps possessing full military and civil powers. The conquered peoples owed tribute and military service to the king. So long as they fulfilled their obligations they were generously treated, being permitted to practice their own religion and manage their internal affairs; but disobedience was harshly punished by massacre or deportation. The imperial army consisted of the Median and Persian cavalry, archers and spearmen, and of the best troops of the subject peoples; the navy was drawn from the Greek states of Anatolia and from the lands of Phoenicia, Cyprus, and Egypt.

Supreme authority in war and peace was vested in the Persian monarch, whose absolute powers were tempered only by the custom of consulting his Persian officials. Darius was described as "the great king, king of kings, king of the countries possessing all kinds of peoples, king of this great earth far and wide." As the vicegerent of the Persian god, Ahura Mazdā, Darius laid claim to world rule.

Darius' first European expedition

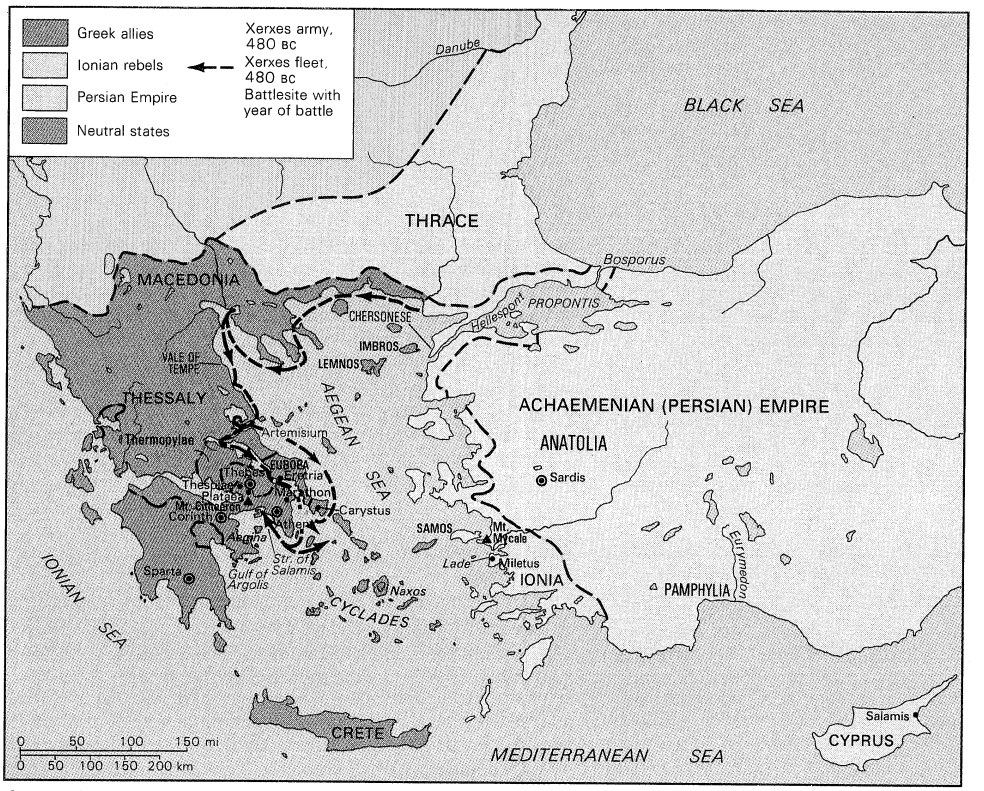
In 514 Darius prepared to conquer Europe. Having made a reconnaissance of Greece and Scythia, he decided to attack Scythia first and instructed a Samian engineer to build a pontoon bridge across the Bosporus. The imperial army overran eastern Thrace, crossed the Danube (where the navy of Greek contingents had made a pontoon bridge), and advanced far into what is now the Ukraine, probably in 513. The Scythians retreated until Darius outran his lines of supply and then harassed his forces when he turned back. The Greek commanders in the Persian navy, although asked by the Scythians to cut the bridge over the Danube, remained loyal to Darius, but some Greek states bordering the Bosporus and the Hellespont rose in rebellion at the news of his discomfiture.

These operations convinced Darius that a strong bridgehead in Europe was necessary. His generals punished the Greek rebels, established in southern Thrace a satrapy that cut off the Scythians from their Spartan allies, and received the submission of the king of Macedonia. Meanwhile, the Persian navy reduced Lemnos and Imbros (Imroz), and a Persian force was ready in 500 to attack Naxos, the strongest island in the Cyclades. This expedition was probably intended to pave the way for an invasion of Greece.

Ionian revolt (499–493 BC). When the Naxos expedition failed in 499, its promoter, Aristagoras of Miletus, instigated the Ionian states on the coast of Anatolia to rebel and obtained help from Athens and Eretria. In 498 their forces took and burned Sardis, capital of the satrapy; they were then joined by the Greek states of Cyprus, the Bosporus, and the Hellespont and by the Carians, while Athens withdrew its support. The Persians subdued Cyprus in 496, and the Bosporus and Hellespont by 495. Their naval victory over the Ionians at Lade in the same year, and the capture of Miletus in 493, virtually ended the revolt.

The Ionian revolt was of great value to the Greek cause. It postponed the Persian attack on Greece for a decade and showed the need of close cooperation and strong leadership. The Ionians had indeed created a council of deputies drawn from the individual states and had entrusted to it the direction of strategy, but they had failed to include in the council the Greeks of the Bosporus, Hellespont, and Cyprus, and they had not appointed a commander in chief of the allied forces until the eve of the battle of Lade, when it was already too late.

Persian attack on Eretria and Athens (490 BC). Darius punished the ringleaders in Asia by execution or deportation, but he made a liberal settlement with the states. Democratic governments were permitted, a moderate rate of tribute was imposed, and the states were required to submit their disputes to arbitration. This was politic,



Greece during the Greco-Persian Wars.
Adapted from W. Shepherd, *Historical Atlas;* Barnes & Noble Books, New York

since Darius hoped to use the Ionian fleet against Greece. In Europe, his son-in-law Mardonius reestablished Persian rule in Thrace and reduced Macedonia, while his envoys visited the free Greek states to ask for "earth and water," the tokens of submission to Persia. In 491 Eretria and Athens were well informed that an attack by sea was impending.

Before the Ionian revolt Sparta and Athens had been at war, but the Persian threat brought them closer together. In 491 the Spartans tried to prevent their ally Aegina from joining Persia, and in 490 Athens attacked Aegina (the chronology is disputed). Athenian policy toward Persia had vacillated before and during the Ionian revolt, but the will to resist was now strengthened by the Persian support of the exiled tyrant Hippias and by the advice of Miltiades (*q.v.*), previously a ruler in the Chersonese, who had returned home with knowledge of Persian tactics. Even so, Sparta and Athens had no plans for united action when the Persian force, perhaps comprising about 25,000 fighting men, sailed across the Aegean Sea, landed on Euboea, and captured Carystus and Eretria. In September 490 the Persian army landed unopposed on the Plain of Marathon in northeast Attica, whence the lines of supply with Euboea and the east were easy and secure. The speed and the initiative of the Persians found Athens still isolated.

Battle of Marathon

The Athenian army was at Athens, prepared to repel any landing in its vicinity, and the small fleet was ready to attack any Persian convoy heading for Aegina. When news came from Marathon, a runner was dispatched to inform Sparta and the assembly decided on the proposal of Miltiades to send the heavy-armed hoplite army (infantry) to the foothills above Marathon. The decision was wise; for the alternative, to stay and defend Athens, would have cut Athens off from Sparta by land and sea, exposing the city to blockade. At Marathon, however, the ten generals (of whom each held operational command for one day, according to Herodotus) and Callimachus, the "polemarch" or nominal commander in chief, had to choose between attack and delay. Miltiades' advice prevailed: to attack as soon as opportunity offered. First the Athenians advanced their position to within a mile of the enemy by felling trees and making obstacles against the dreaded Persian cavalry. Then the opportunity came. Before dawn some Ionian deserters reported "The cavalry are away." Miltiades, who chanced to hold the operational command, attacked at dawn. With a thin centre and strengthened wings the line of about 10,000 Athenians and 1,000 Plataeans charged the enemy infantry before the cavalry could return. The Greek wings defeated the Persians and wheeled inward to attack the Persian centre, which had driven the Greek centre back. The longer spear and heavier armour of the bronze-clad Greek infantryman prevailed over the Persian with his short spear, wicker shield, and padded clothing. The rout was complete. According to Herodotus, the Greeks lost 192 men, the Persians 6,400. The majority escaped to the fleet, which sailed at once, hoping to surpise Athens, but the Athenians—by a forced march—arrived that evening to defend the city. The Persians then departed. A Spartan force, which had been delayed by religious observances at Sparta, arrived at Marathon too late to take part in the battle.

Expedition of Xerxes (480–479 BC). The Persian failure was followed by a full-scale invasion. It was delayed by a revolt in Egypt and the death of Darius until 480, when Xerxes crossed the Hellespont in late spring with a vast army and a large fleet. The advance was slow and the fleet had to provision the army. The Greeks therefore had ample time to make preparations. The problem of uniting the 30 states that had the will to resist Persia was solved by Sparta, which held a congress of delegates and formed a general alliance. The states agreed to stop all wars among themselves and conferred the command by land and sea on Sparta. The congress met regularly, each state having one vote, and decisions by the majority were binding on all members. It possessed recruiting, diplomatic, and judicial powers. In the field the commander in chief, who was nominated by Sparta, consulted the commanders of the national contingents but made his own decisions.

Thus the Greek congress was a highly centralized and efficient organization for allied action. Its chief strength on land lay in the Spartans and their allies; at sea it lay in the Athenians, ably led by Themistocles, who had increased the fleet to 200 ships. A coordinated defense on

both elements was now possible. It lay with Sparta to choose the time and place for applying the relatively small but excellent forces of the Greek congress.

The first decision, to hold the narrow Vale of Tempe between Macedonia and Thessaly, was abandoned when it was realized that the position could easily be turned. On news of this the Athenians voted, on the proposal of Themistocles, to entrust themselves to "the wooden wall" of their ships in accordance with an utterance of the Delphic oracle, and plans were made for evacuating the noncombatants. Next the Greeks occupied the still narrower pass of Thermopylae with 6,000 or 7,000 hoplites and stationed 271 ships at Artemisium in northern Euboea. The positions were linked by communication between the Spartan commanders, King Leonidas at Thermopylae and Eurybiades at Artemisium, who intended to halt and damage the Persian forces. Meanwhile, Xerxes was advancing slowly. He made no use of separate columns, and his fleet suffered heavy losses in a storm when it was convoying supply ships along the coast. It was already August when Xerxes began the operations, which extended over three days.

Thermopylae

On the first day he sent a detachment of 200 ships, unseen by the Greeks, to sail round Euboea and close the narrows of the Euboean Channel; and he also attacked with his best infantry at Thermopylae, where the Greeks inflicted heavy casualties. During the afternoon the Greek fleet, having learned about the Persian detachment from a deserter, engaged the main Persian fleet with some success. The Greeks intended to sail south that night and destroy the detachment next day, but a tremendous storm kept the Greeks at Artemisium and wrecked the 200 Persian ships off south Euboea. On the second day news of the Persian disaster was brought up by a reinforcing squadron of 53 Athenian ships. Xerxes attacked again with no success at Thermopylae, and the Greeks sank some Cilician vessels off Artemisium. That evening a Greek traitor, Ephialtes, offered to guide the Persians along a mountain path and turn the position at Thermopylae. The Persians' best infantry, called the "Immortals," were entrusted to him. At dawn on the third day they began to descend toward the plain behind the Greek position. Leonidas retained the troops of Sparta, Thespiae, and Thebes and sent the remainder south. He then advanced. Except for the Thebans, who surrendered, he and his men fought to the death. Meanwhile the Persian fleet attacked at noon. Both sides suffered heavy losses and the Greeks realized that they could only succeed in narrower waters. That evening, when the fall of Thermopylae was known, the Greek fleet withdrew down the Euboean Channel and took station in the narrow Strait of Salamis.

In September, Xerxes, joined by many Greeks north of Attica, burned Athens. The city was almost deserted, for the evacuation had been completed. The Greek congress decided to fortify the isthmus and keep the fleet forward at Salamis. This decision caused dissension among the ship captains. Many wished to retire to the Argolic Gulf. As a stratagem, Themistocles informed Xerxes of their desire; Xerxes, who saw the end of the campaigning season close at hand, sent 200 ships that night to cut the Greek line of retreat and posted the main fleet, numbering probably 1,207 ships, off the eastern exit of the Straits of Salamis. During the night the Greeks learned of his dispositions and intentions. Putting to sea at dawn they feigned a retreat, actually sending a detachment northward to look out for the 200 Persian ships, and their manoeuvres led the enemy to advance incautiously into the narrow waters where superior numbers were of little effect. Within the narrows the Greek ships, stoutly built for ramming, had room to manoeuvre against the congested stream of Persian ships, which, designed for boarding tactics, proved less handy under oar and fell foul of one another. The result was a complete triumph for Greek seamanship. The Persians fled in confusion. Soon afterward their fleet, still superior in numbers but not in morale, set sail for Asia.

That winter, while Xerxes departed to Asia, a large army wintered in Thessaly under the command of Mardonius. By skillful diplomacy he drew the Greeks forward in the summer of 479 to the northern foothills of Mt. Cithaeron near Plataea, where difficulties of supply forced the Greek army of 110,000 men to withdraw during the night. The withdrawal was disorderly and dawn found the army scattered. Mardonius at once attacked a group of 11,500 Spartan and Tegean hoplites who had halted on hilly ground. Their commander, Pausanias, undismayed by the swarms of Persian infantry, led his men downhill in close formation, charged at the double, and overwhelmed the enemy. When the Athenians came up after defeating the Thebans, the Greeks stormed the camp and the survivors of the Persian army fled. Meanwhile, the Greek fleet had passed to the offensive at Mycale on the Asiatic coast opposite Samos. The Persians refused battle, beached their ships, and joined a large supporting army, but the Spartan king Leotychidas landed his men farther north and attacked with complete success. The victories of Plataea and Mycale ended the Persian invasion.

Plataea and Mycale

Greek offensive (478–448 BC). The Greek triumph was due to Spartan leadership, Athenian loyalty, and Greek fighting power. The Spartans, however, had no desire to campaign in Asia, whereas the Athenians were ready to deploy their fleet in support of the Ionians. Hence arose the Delian League, formed by Athens as executive leader and by many Greek states on the islands and Asiatic coast, to defend Greek liberty and exact retribution from Persia. A series of successful operations culminated *c.* 466 in victory at the Eurymedon River in Pamphylia, where an allied force of 300 ships defeated a Persian army and navy. In 460 the Athenians and their allies supported Egypt in a successful revolt. But the Persian army returned to the attack; Egypt made a separate peace, and the Greeks, overconfident in their sea power, were trapped on the Nile and annihilated in 454. By this time the Athenians were at war with Sparta, but a truce on the Greek mainland enabled them to launch successful attacks on Cyprus in 450–449. A treaty of peace was concluded, probably in 448, by the Athenians, their allies, and Artaxerxes I of Persia that recognized the liberty of the Greek states in Europe and Asia and kept the Persian fleet out of the Aegean Sea.

BIBLIOGRAPHY

Translation: AUBREY DE SELINCOURT, *Herodotus: The Histories* (1954).

General accounts: The Cambridge Ancient History, vol. 4, *The Persian Empire and the West*, ch. 1, 7–10 (1926) and vol. 5, *Athens*, ch. 2–3 (1927), somewhat dated and lacks references to ancient sources; ANDREW R. BURN, *Persia and the Greeks: The Defence of the West, c. 546–478 B.C.* (1962), the most detailed, up-to-date account; N.G.L. HAMMOND, *A History of Greece to 322 B.C.*, 2nd ed., pp. 176–286 (1967), gives references to ancient sources.

Special topics: GEORGE B. GRUNDY, *The Great Persian War and Its Preliminaries* (1901, reprinted 1969), an original and important work; CHARLES HIGNETT, *Xerxes' Invasion of Greece* (1963), critical of ancient literary sources, not strong in matters of topography; P.A. BRUNT, "The Hellenic League Against Persia," *Historia*, 2:135–163 (1953); N.G.L. HAMMOND, *Studies in Greek History* (1973), includes discussions of the battles of Marathon and Salamis and of the Athenian Alliance of 478–477 BC; B.D. MERITT, H.T. WADE-GERY, and M.F. MCGREGOR, *The Athenian Tribute Lists*, 4 vol. (1939–53).

(N.G.L.H.)

Greece

A constitutional monarchy until 1974—though the reigning monarch fled the country in 1967 as part of a series of events that continues to mold the nation—Greece is a European republic stretching across the tip of the Balkan Peninsula from the Ionian Sea to the Aegean and including islands in both seas. Its area—50,960 square miles (131,986 square kilometres)—is about that of Czechoslovakia or North Korea, and its neighbours (clockwise from the northwest) are Albania, Yugoslavia, Bulgaria, and Turkey; its insular possessions stretch as far south as the major Mediterranean island of Crete (Kríti). Its capital is Athens (Athínai).

There is about Greece a vitality, all too often undisciplined, that makes many another European country seem tame, even dull, by comparison. This is evident immediately on sailing into a Greek port or crossing a Greek border: sounds, smells, movements, colours—the very tempo of things—conspire to heighten sensibilities and intensify expectations. But alongside all this is the serene coolness, even aloofness, of what remains in Greece from classical antiquity, the visible monuments of which constantly stand as a challenge to (if not even a rebuke of) contemporary endeavours.

The vitality of Greece and Greeks can be said to stem from the heady mixture over centuries of many peoples and ways of life. That land, now the home of about 9,000,000 persons, has long been "at the crossroads"—at that place in the Balkans where "three roads meet," where Europe, Africa, and the East converge, and at that point in time where the ancient, the medieval, and the modern coexist and conflict. The Greek is familiar with the Middle East: his language and food, to say nothing of his religion and history, are marked by exotic and even oppressive elements from Turkey, the Holy Land, Egypt, and beyond. But he is familiar as well with the more sober West, with the Europe to which the young go for training and for work, with the Americas and Australia to which so many have gone for a new home. Greeks, like the Jews whom they resemble in so many ways, have long been able to adjust themselves as merchants in many climes and to many ways of life. But, also like the Jews, they have preserved in their heart of hearts a vital memory of the homeland to which they yearn to return.

The crossroads

This homeland is as much a state of mind as it is a place to be found on maps. The yearning to "return," then, is almost as strong among those who have never left Greek soil as among those who find themselves abroad. Perhaps it is a yearning to attain that which has never been but which has always been aspired to. It is a yearning evident in the melancholy of Greek music, the nostalgia that can be heard even in the lively tunes and ballads sung on festive occasions. It is a yearning that can be heard as well in interminable conversations, especially those with which Greeks refresh themselves through the long cool nights that follow blistering summer days. And it is a yearning that can be seen in the faces and deeds of the Greeks, a yearning that makes it impossible for them "to leave well enough alone." A perpetual restlessness, much like that which was said to characterize the political life of ancient Athens, is evident, a restlessness that can continue subterraneanly despite surface conformity to the tyranny of the moment.

Indeed, there have been many tyrannies in Greece, tyrannies that are as much a part of the much-discussed "Greek experience" as (if not even the most frequent result of) their volatile democracies. Perhaps, it might even be said, memories of tyrannies remind Greeks of the unpredictability of human things, of the disaster that can follow upon prosperity, of the trials that even the most successful encounter from time to time. Life can be expected to be as hard, as unyielding, and as toughening as the soil and the sea from which Greeks have for centuries wrested their livelihood. But it can also be as enriching and as exciting as the landscape and the light for which Greece has always been celebrated and that can be seen, if not at this moment or place, then surely in a little while or down the road a few kilometres.

The following article surveys contemporary Greece. Additional information may be found in the articles on AEGEAN CIVILIZATIONS; AEGEAN SEA; ATHENS; BALKANS, HISTORY OF THE; BYZANTINE EMPIRE; CRETE; GREEK CIVILIZATION, ANCIENT; and the appropriate sections of VISUAL ARTS, WESTERN, and LITERATURE, WESTERN.

THE LAND

Topography. The Greek landscape is conspicuous not only for its beauty but also for its complexity and its variety. The dominant influence—as noted by Strabo, the great geographer of classical antiquity, and confirmed by a glance at the map—is the sea. An ever-present factor, the sea presses deep into the land in a host of arms and inlets, which are often separated by the rocky spines of peninsulas that thrust back into the sea and are continued in the arcs and clusters of beautiful islands across its surface. Only a small, wedge-shaped portion of the interior of modern Greece is more than 50 miles from the sea. Mountains are the second major element in Greek topography. They cover about three-quarters of the country's surface, forming a ribbed, interlaced network, trending generally from northwest to southeast, and enclosing numerous small basins. The basins—together with narrow valleys, small plains (spreading more extensively about river mouths), and a thin, broken coastal strip—form the third element in the relief, the lowland.

The three elements in the topography

In their combination and interaction, these three elements have been of immense significance in shaping national development: the rich soils of the basins nurtured agriculture and the first stirrings of civilization, but the mountains (while often serving as a barrier against invaders) constricted this social development to small, fiercely independent areas and impeded communications between them. The sea—as both history and the ancient stories attest—made for contact between the separate communities and stimulated contacts farther afield, although it also exposed the coastal regions to external attack. For the geographer these three elements, in local combination, are a convenient basis for a division of the contemporary Greek landscape into six natural regions.

The Pindus Mountains. The core region in Greek topography is unquestionably the rugged Pindus Mountains (Píndhos Óros) area of the northwest interior. Following the general northwest–southeast trend of the mountains of the Balkan Peninsula, the Pindus sweep down in a series of rugged, roughly parallel ridges from the Albanian and Yugoslavian frontier and are structurally a part of the Dinaric mountain system of those countries. This system of young fold mountains was created relatively late in geological time, and earthquakes continue to afflict the region as the mountain structures settle down. The highest point in this region is Smólikas Óros (*óros*, plural *óri*, "mountain"), 8,651 feet (2,637 metres) high. The mountain scenery, with jagged granitic peaks, wild gorges, and a succession of magnificent views glimpsed from winding roads, is justly famed.

Macedonia and Thrace. A number of topographic regions surround this mountainous core and are often penetrated by extensions of it. The northernmost division, roughly the regions of Macedonia and Thrace (Thráki), extends in a long, narrow east–west band between the north Aegean coast and the frontier with Yugoslavia and Bulgaria. It is bounded on the east by the Maritsa River (Évros Potamós; *potamós*, "river"), which marks the Turkish frontier, and consists of hills and forest-clad mountains interpenetrated by valleys, river basins, and alluvial plains. Along the Bulgarian border and beyond it rise the ancient crystalline rock structures of the Rhodope Mountains (Rodhópi Óri), against which the newer mountains of the Pindus were crushed during their formation. In the west, the three-pronged peninsula of Khalkidhikí (Chalcidice) forms a distinctive feature as it thrusts out into the Aegean. On the easternmost prong, Áyion Óros (Holy Mountain), is located Mt. Áthos, the site of the famous monastic community.

The Khalkidhikí peninsula and Mt. Áthos

The peninsula is separated from the rest of the coastal region by a fault line of structural weakness, marked by the lakes Korónia and Vólvi. Just to the west extends a large plain drained by the Vardar (Axiós) and Aliákmon rivers, whose swampy deltas are slowly pushing out into the nearby Thermaïkós Kólpos (*kólpos*, "gulf"). The forested Vérmion Óros and, beyond it, the barren inland basins around lakes Vegorrítis and Kastorías mark the boundary with the Pindus proper. Farther east is a succession of plains, often swampy; that of Sérrai, around the lower Struma (Strimón), and the deltaic plain of the lower Néstos are most significant. Inland basins of structural origin include that of the Pedhiás Drámas.

Central Greece. Central Greece lies to the south of Macedonia and Thrace and is lent character by four spurs that thrust out from the main Pindus mass, following the northwest–southeast trend of that region. A number of

distinctive basins and plains lie amidst these upland ribs. The northernmost, a rather broken spur called the Kamvoúnia Óri, runs along the coast of the Thermaïkós Kólpos and continues south to form the peninsula bounding one side of the Órmos Vólou (*órmos*, "bay"). One of its peaks is Mt. Olympus (Óros Ólimbos)—the mythical seat of the gods, whose often cloud-topped summit rises to 9,570 feet (2,917 metres), the highest point in Greece—and also the equally fine peaks of Óssa and Pílion (Pelion). The next spur on the west is the Óros Óthris range, which continues across the narrow Dhíavlos Oreón (Oreón Channel) in the northern sector of the long, narrow island of Évvoia (Euboea). Between the two spurs lie the ancient basins (formerly the site of lakes) of Thessalía (Thessaly), Tríkkala (Tríkala), and Lárisa, drained by the Piniós. Just to their south, the basin of Almirós, of similar origin, lies around the Órmos Vólou.

To the southwest, the third spur leaving the Pindus is that of the Oíti, continued in the Ókhi Óros of southern Évvoia. Just before the Oíti reaches the sea, near the head of the Maliakós Kólpos, lies the pass of Thermopylae (Thermopílai), scene of the famous battle of antiquity. The last (and perhaps the most important) of the four spurs thrusting down into central Greece is that curving away to the southeast through the twin-peaked mass of Mt. Parnassus (Parnassós). This mountain rises to 8,061 feet (2,457 metres) and was held to be the home of the muses. The view from its summit at sunrise, with a broad expanse of the heart of Greece gradually unfolding, is regarded as one of the finest in the world. The range continues as the backbone of the peninsula lying between the Vórios Evvoïkós Kólpos and the Gulf of Corinth (Korinthiakós Kólpos), and it reaches as far as Párnis Óros, just to the north of Athens. To its north lie the plains of Fokís (Phocis) and Voiotía (Boeotia) and around its southern tip lie the depressions of Attica (Attikí), hotter and more arid but with a strategic importance that helps to explain the rise of Athens.

Mt. Parnassus and its regional context

The Pelopónnisos. The entire southern portion of mainland Greece forms a peninsula lying to the south of the Gulf of Corinth. Technically, this region, the Pelopónnisos, or Peloponnese, also known as the Morea, is now an island, for the 3.9-mile (6.3-kilometre) Dhiórix Korínthou (Corinth Canal) cuts across the narrow neck of land formerly separating the Gulf of Corinth from that of Aíyina (Aegina). The Pelopónnisos consists of an oval-shaped mass with mountains rising to 7,800 feet and four peninsular prongs pushing out southward toward Crete. The limestone mass of the plateau of Arkadhía (Arcadia), where streams disappear underground in the soluble rocks, forms the heart of this mass, with the barren land of Taïyetos Óros (rising 7,897 feet, or 2,407 metres) extending southward to form the backbone of one of the subsidiary peninsulas. This mountainous core is surrounded by a thin fringe of fertile coastal plain in the north and west and by the larger alluvial depressions of Lakonía (Laconia), Messinía (Messenia), and Árgos, which head the inlets between the peninsulas to the south. The coast is indented and offers some fine harbours, and the whole region is noted for its wild beauty.

The western uplands and islands. The western side of the Greek mainland north of the Gulf of Corinth to the Albanian frontier and the offshore islands (the Ionian Islands; Iónioi (Nísoi) possess their own distinctive topography and regional character. This effect has been enhanced by the fact that the mountainous barrier of the Pindus lying inland and the ameliorating climatic influences from the west have led to historic isolation from the rest of Greece. Fertile basins are not as well developed as in some other parts of Greece because they are constricted by the parallel ranges of coastal mountains, but the mountain regions themselves, being adequately supplied with rainfall, are not so barren as elsewhere. Kérkira (Corfu), the island lying opposite the Albanian frontier, is the northernmost of the seven major Ionian Islands. It is fertile and well watered. The other islands, Paxoí (Paxos), Levkás (Leucas), Skorpiós, Ithákí (Ithaca), Kefallinía (Cephalonia), and Zákinthos (Zacynthus), lie farther south. Lack of rainfall accentuates their gaunt, broken limestone relief, although Levkás and Zákinthos have sheltered eastern plains.

The Aegean Islands (Nísoi Aiyaíou). Just as ridgelike extensions of the Pindus interpenetrate the basins and coastal plains of mainland Greece, island groups, which are often further extensions of the same mountain chains, form distinct regional clusters in the Aegean Sea. In the northeast, the island region forms the most extensive—and, visually, perhaps the most attractive—of the physiographic divisions of Greece. In the north, off Thrace (Thráki), lie Thásos (an oval block of ancient mineral rocks similar in composition to neighbouring blocks on the mainland) and harbourless Samothráki (Samothrace), an island of volcanic origin. Límnos, situated midway between Asia Minor and Áyion Óros, is almost cut in two by the northern Órmos Pourniá and the deep southern harbour afforded by the Kólpos Moúdhrou.

Regional clusters of islands

To the southeast, the rocky but sheltered islands of Lésvos (Lesbos), Khíos (Chios), and Sámos lie close to the Turkish coast and are extensions of peninsulas on the coast of Asia Minor. Across the central Aegean, near northern Évvoia (Euboea), lie the Northern Sporades (Voríai Sporádhes), or "scattered" islands; their crystalline rocks are similar to those of the Greek mainland. Farther south, in the heart of the Aegean, lie the Cyclades (Kikládhes), "islands in a circle." These roughly centre on Dhílos (Delos) and represent the tips of drowned mountain ridges continuing the structural trends of Évvoia and the region around Athens.

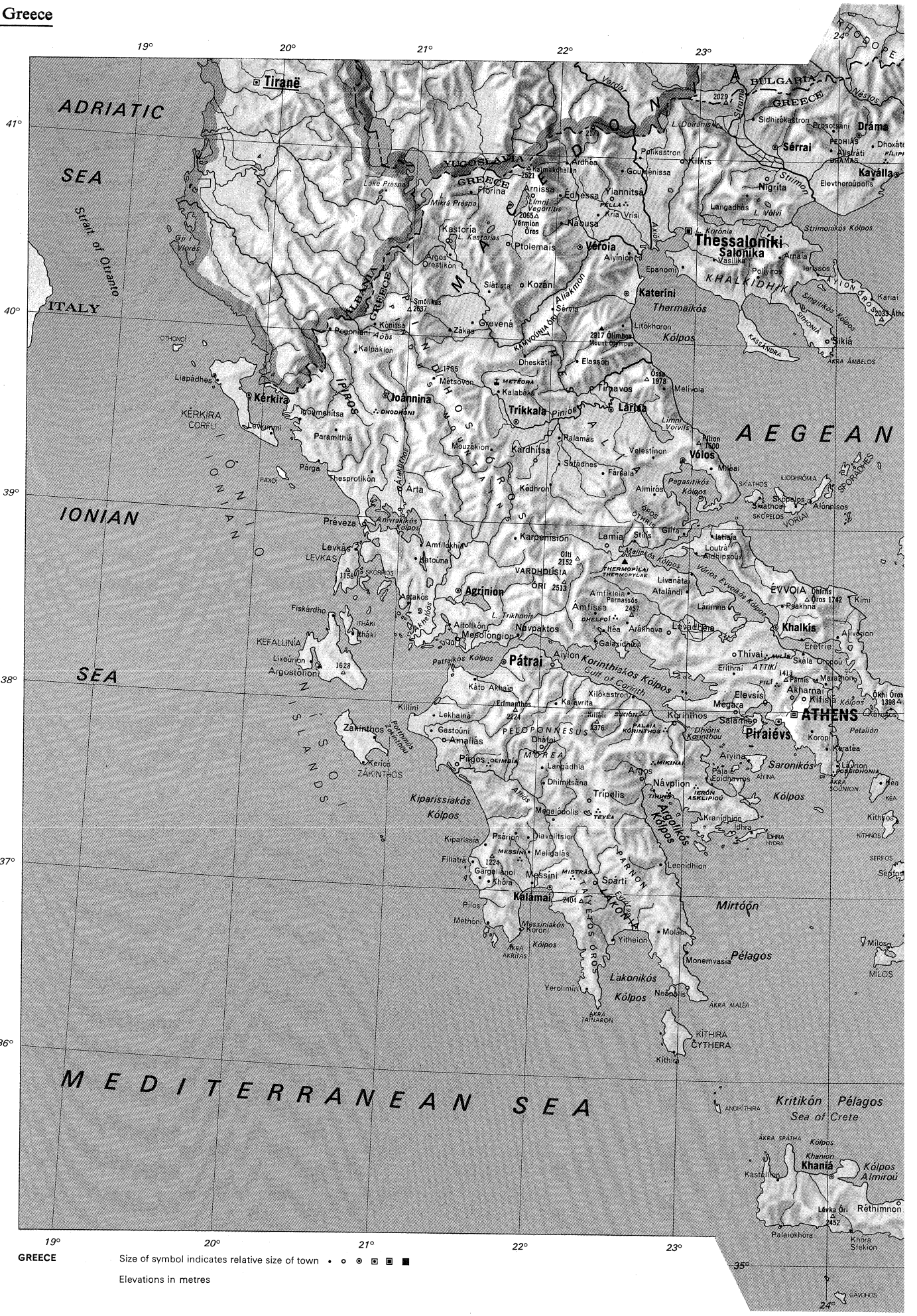
Between the Cyclades and the Turkish coast, the Dodecanese (Dhodhekánisos) group, with Rhodes (Ródhos) the largest of a dozen major islands, have a varied geological structure ranging from the gray limestones of Kálimnos, Símí, and Khálki to the complete ancient volcanic cone that forms Nísiros. Finally, the long, narrow shape of Crete stands at the entrance of the Aegean in the extreme southern portion of Greek territory. Its harsh, rugged landscapes mark yet another extension of the fold mountains of the Balkan Peninsula.

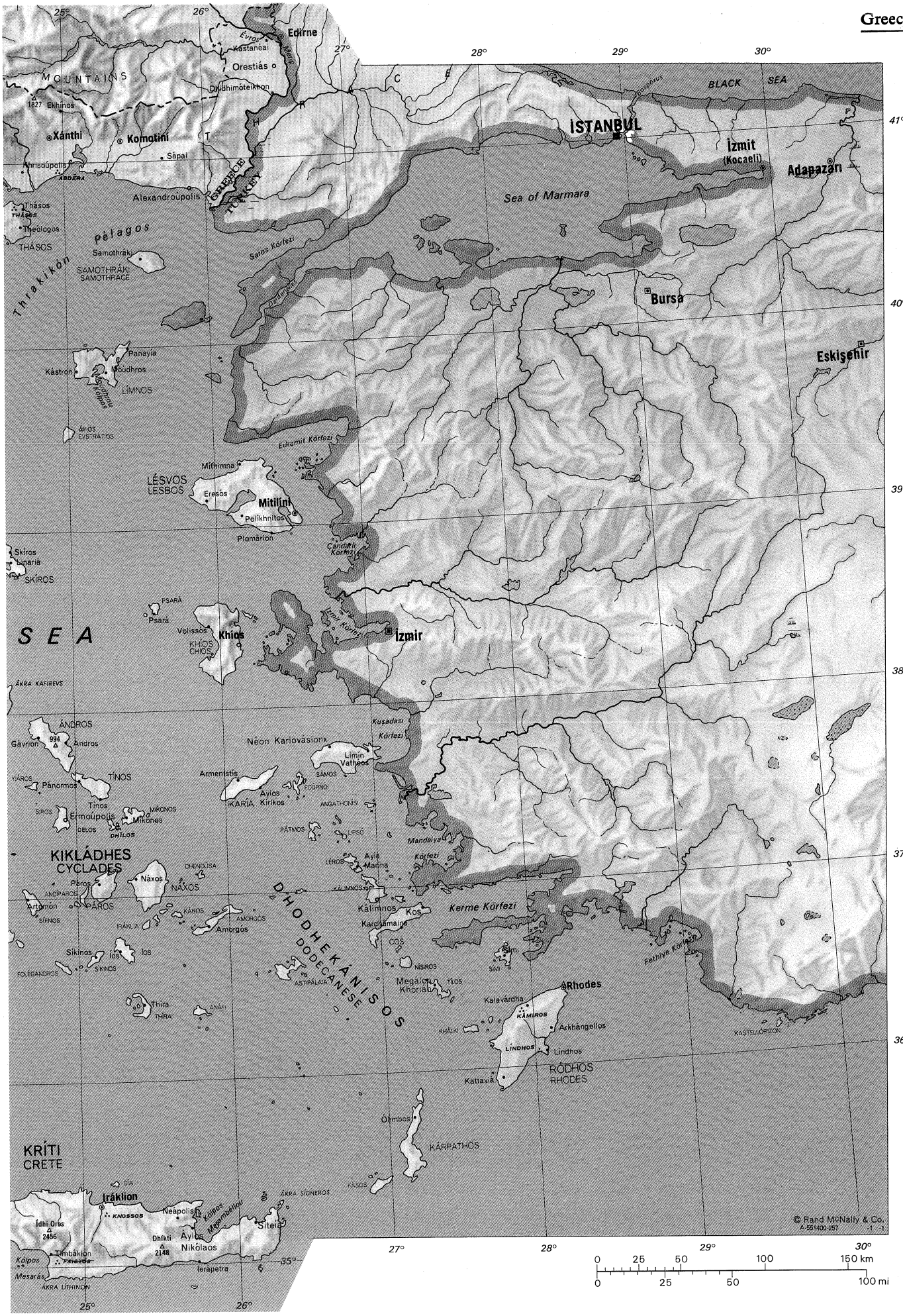
Climate and drainage. The basically Mediterranean climate of Greece is subject to a number of regional and even local variations occasioned by the country's physical diversity. In winter the belt of low pressure disturbances moving in from the North Atlantic shifts southward, bringing with it warm, moist, westerly winds. Squalls and spells of rain ruffle the Aegean, but sunshine often breaks through the clouds. As the low pressure areas enter the Aegean region, they may draw in cold air from those eastern regions of the Balkans that, sheltered by the Dinaric mountain system from western influences, are open to climatic extremes emanating from the heart of Eurasia. This icy wind is known as the boreas. Partly as a result, Thessaloníki (Salonika) has an average January temperature of 43° F (6° C), while Athens has 50° F (10° C). *Shilok*, or warm winds, are similarly drawn in from the south. The western influences bring plentiful rain to the Ionian coast and the mountains behind it; winter rain also starts early, and snow lingers into spring. At Kérkira (Corfu), January temperatures average 50° F (10° C), and the island's average annual rainfall is 52 inches (1,300 millimetres), compared with the Athens total of 16 inches (400 millimetres).

In summer, when the low pressure belt swings away again, the climate is hot and dry almost everywhere, with the average July sea-level temperature approaching 80° F (27° C), although heat waves can push the temperature up over the 100° F mark for a day or so. Topography is again a modifying factor: the interior northern mountains continue to experience some rainfall, while all along the winding coast the afternoon heat is eased slightly by sea breezes. In other regions the hot, dry summers are accentuated by the parching etesian winds, which become drier and drier as they are drawn southward.

In all seasons—perhaps especially in summer—the quality of the light is one of Greece's greatest treasures. Although the larger cities have not escaped the pernicious effects of industrial and vehicular pollution, the Greek atmosphere is generally pure and clear. The interplay of light and varied landscape is remarkable. The harsh white

The quality of Greek light


ADRIATIC SEA
IONIAN SEA
AEGEAN SEA
MEDITERRANEAN SEA
Kritikón Pélagos
Sea of Crete
ITALY
Strait of Otranto
ALBANIA
GREECE
YUGOSLAVIA
BULGARIA
Tiranë
Kérkira
CORFU
Ioánnina
Tríkkala
Lárisa
Kozáni
Kastoria
Véroia
Thessaloníki
Salonika
Kateríni
Thermaikós Kólpos
Sérrai
Drama
Kaválla
Vólos
Lamía
Agrínion
Préveza
Levkás
LEVKAS
Pátrai
Korinthiakós Kólpos
Gulf of Corinth
ATHENS
Piraiévs
Khalkís
EVVOIA
Kalamai
Messini
Pílos
Argolikós Kólpos
Kiparissiakós Kólpos
Návplion
Tripolis
Sparti
Mirtóon Pélagos
Lakonikós Kólpos
Monemvasia
Yithion
Neápolis
Khaniá
Kólpos Almirou
Réthimnon
Iráklion
Sitía
Ierápetra
Ágios Nikólaos
Mália
Kólpos Mirabéllou
Kríti
CRETE
Pindos Mountains
Rodopi Mountains
Olympus
Mount Olympus
Parnassós
Parnitha
Pilion
Ossa
Pelion
Peloponnesus
Mainland
Ionian Islands
Aegean Islands
Sporades
Cyclades
Dodecanese
Rhodope
Strymon
Nestos
Vardar
Axios
Aliákmon
Pinios
Acheloos
Lake Prespa
Mikrá Préspa
Limni Vegorítis
Limni Vólvis
Limni Kastorías
Gulf of Corinth
Kólpos
Akra
Ákra
Cape
Island
Mount
Mountains
Gulf
Bay
Strait
Lake
Limni
Nisí
Óros
Akrotírion
Kólpos
Pélagos
Ákra
Kríti
CRETE
KÍTHIRA
CYTHERA
ANDIKÍTHIRA
KÍTHNOS
SÍFNOS
SÉRIFOS
MÍLOS
KÉA
SKÓPELOS
SKÍATHOS
ÁLONNISOS
LÉSVOS
CHÍOS
SÁMOS
KÁRPATHOS
RÓDOS
KÁSOS
GÁVDHOS
KÁSTELLORIZO
ÍOS
SÍKINOS
FÓLEGANDROS
SÝROS
MÝKONOS
TÍNOS
ANDROS
KÉA
KÝTHNOS
SÉRIFOS
SÍFNOS
MÍLOS
KÍMOLOS
PÁROS
NÁXOS
AMORGÓS
SKÝROS
LÍMNOS
THÁSOS
SAMOTHRÁKI
LESBOS
CHÍOS
SÁMOS
ÍKARIA
PÁTMOS
LÉROS
KÁLIMNOS
KÓS
NÍSYROS
TÍLOS
SÝMI
KÁRPATHOS
KÁSOS
GÁVDHOS
KÁSTELLORIZO
Size of symbol indicates relative size of town
Elevations in metres
GREECE
19°
20°
21°
22°
23°
24°
41°
40°
39°
38°
37°
36°
35°



Edirne
Kastanéai
Orestiás
Dhidhimóteikhon
MOUNTAINS
1827 Ekhínos
Xánthi
Komotiní
Sápai
Khrisoúpolis
ABDÉRA
Alexandroúpolis
GREECE
TURKEY
Thásos
THÁSOS
Theólogos
THÁSOS
Pélagos
Samothráki
SAMOTHRÁKI
SAMOTHRACE
Thrakikón
Saros Körfezi
Dardanelles
Panayía
Kástron
Moúdhros
LÍMNOS
ÁYIOS EVSTRÁTIOS
Edremit Körfezi
Mithimna
LÉSVOS
LESBOS
Eresós
Mitilíni
Polikhnitos
Plomárion
Skíros
Linaria
SKÍROS
PSARÁ
Psará
Volissós
Khíos
KHÍOS
CHIOS
SEA
ÁKRA KAFIREVS
ÁNDROS
Gávrion
994
Ándros
TÍNOS
Pánormos
Tínos
Ermoúpolis
MÍKONOS
Mikónos
DHÍLOS
KIKLÁDHES
CYCLADES
Páros
Náxos
NÁXOS
Artemón
PÁROS
Sikinos
Íos
Íos
Amorgós
Thíra
THÍRA
ANÁFI
DHODHEKÁNISOS
DODECANESE
Kuşadası Körfezi
Néon Karlovásionx
Limín Vathéos
SÁMOS
Armenistís
Ayios Kírikos
IKARÍA
FOÚRNOI
ANGATHONÍSI
PÁTMOS
LIPSÓ
Mandalya Körfezi
LÉROS
Ayía Marina
KÁLIMNOS
Kálimnos
Kos
Kardhámaina
COS
Kerme Körfezi
NÍSIROS
Megálon Khoriáb
TÍLOS
SÍMI
Rhodes
Kalavárdha
KÁMIROS
Arkhángellos
LÍNDHOS
Lindhos
RÓDHOS
RHODES
Kattaviá
Fethiye Körfezi
KASTELLÓRIZON
Ólimbos
KÁRPATHOS
KÁSOS
KRÍTI
CRETE
Iráklion
KNOSSOS
Neápolis
Kólpos Mesambéllou
Síteia
Ídhi Oros
2456
Dhíkti
2148
Áyios Nikólaos
Timbákion
Kólpos
Mesarás
ÁKRA LÍTHINON
Ierápetra
ÁKRA SÍDHEROS
BLACK SEA
Bosporus
İSTANBUL
İzmit
(Kocaeli)
Adapazarı
Sea of Marmara
Bursa
Eskişehir
İzmir Körfezi
İzmir
25°
26°
27°
28°
29°
30°
41°
40°
39°
38°
37°
36°
35°
0 25 50 100 150 km
0 25 50 100 mi
© Rand McNally & Co.
A-551400-257

MAP INDEX

Cities and towns

Agrínion....38·37n 21·24e
Aitolikón....38·27n 21·22e
Aíyina....37·44n 23·27e
Aiyínion....40·30n 22·33e
Aíyion....38·15n 22·05e
Akhainaí....38·05n 23·44e
Alexandroúpolis..40·50n 25·52e
Alístráti....41·04n 23·57e
Alivérion....38·24n 24·02e
Almirós....39·11n 22·46e
Alónnisos....39·08n 23·50e
Amaliás....37·49n 21·23e
Amfíkleia....38·38n 22·35e
Amfilokhía....38·51n 21·10e
Ámfissa....38·31n 22·24e
Amorgós....36·50n 25·54e
Ándros....37·50n 24·57e
Arákhova....38·29n 22·35e
Ardhea....40·59n 22·03e
Árgos....37·39n 22·44e
Árgos Orestikón..40·28n 21·16e
Argostólion....38·10n 20·30e
Arkhángellos....36·12n 28·08e
Armenistís....37·36n 26·08e
Arnaía....40·29n 23·35e
Árnissa....40·48n 21·50e
Árta....39·09n 20·59e
Artemón....36·59n 24·43e
Astakós....38·32n 21·05e
Atalándi....38·39n 23·00e
Athens (Athínai)..37·58n 23·43e
Ayía Marína....37·09n 26·52e
Áyios Kírikos....37·37n 26·14e
Ayios Nikólaos...35·11n 25·42e
Canea, see Khaniá
Chalcis, see Khalkís
Dháfni....37·48n 22·01e
Dheskáti....39·55n 21·49e
Dhidhimóteikhon....41·21n 26·30e
Dhimitsána....37·37n 22·03e
Dhoxáton....41·05n 24·14e
Diavolitsion....37·18n 21·58e
Dráma....41·09n 24·08e
Édhessa....40·48n 22·03e
Ekhínos....41·17n 24·59e
Elassón....39·54n 22·11e
Elevsís....38·02n 23·32e
Elevtheroúpolis..40·55n 24·16e
Epanomí....40·26n 22·56e
Erétrie....38·23n 23·48e
Eresós....39·18n 25·51e
Erithraí....38·13n 23·19e
Ermoúpolis....37·26n 24·56e
Fársala....39·18n 22·23e
Filiatrá....37·10n 21·35e
Fiskárdho....38·27n 20·35e
Flórina....40·47n 21·24e
Galaxídhion....38·22n 22·23e
Gargaliánoi....37·04n 21·39e
Gastoúni....37·51n 21·16e
Gávrion....37·52n 24·46e
Glífa....38·57n 22·58e
Goumenissa....40·57n 22·27e
Grevená....40·05n 21·25e
Ídhra....37·20n 23·29e
Ierápetra....35·00n 25·45e
Ierissós....40·24n 23·52e
Igoumenítsa....39·30n 20·16e
Ioánnia....39·40n 20·50e
Ios....36·44n 25·17e
Iráklion....38·04n 23·46e
Istiaía....38·57n 23·09e
Itéa....38·26n 22·24e
Itháki....38·23n 20·42e
Kalabáka....39·42n 21·43e
Kalámai....37·04n 22·07e
Kalavárdha....36·20n 27·57e
Kalavrita....38·01n 22·06e
Kálimnos....36·57n 26·59e
Kalpákion....39·55n 20·20e
Kardhamaina....36·47n 27·09e
Kardhítsa....39·21n 21·55e
Kariaí....40·16n 24·15e
Káristos....38·00n 24·24e
Karpenísion....38·55n 21·40e
Kastaneai....41·38n 26·28e
Kastéllion....35·30n 23·38e
Kastoría....40·31n 21·15e
Kástron....39·52n 25·04e
Kateríni....40·16n 22·30e
Káto Akhaïa....38·09n 21·32e
Katoúna....38·47n 21·07e
Kattaviá....35·57n 27·46e
Kaválla....40·56n 24·25e
Kéa....37·38n 24·21e
Kédhron....39·13n 22·03e
Keratéa....37·48n 23·59e
Keríon....37·40n 20·48e
Kérkira (Corfu)...39·36n 19·56e
Khalkís (Chalcis)..38·28n 23·36e
Khaniá (Canea)..35·31n 24·02e
Khíos....38·22n 26·08e
Khora....37·04n 21·43e
Khóra Sfakíon...35·12n 24·09e
Khrisoúpolis....40·58n 24·42e
Kifisiá....38·04n 23·48e
Kilkís....41·00n 22·53e
Killíni....37·55n 21·09e
Kími....38·37n 24·06e
Kiparissía....37·14n 21·40e
Kíthira....36·02n 22·58e
Kíthnos....37·26n 24·26e
Komotiní....41·08n 25·25e
Kónitsa....40·02n 20·45e
Korinthos....37·56n 22·56e
Koróni....36·48n 21·56e
Koropí....37·54n 23·53e
Kos....36·53n 27·18e
Kozáni....40·18n 21·47e
Kranídhion....37·22n 23·10e
Kría Vrísi....40·41n 22·18e
Lamía....38·54n 22·26e
Langadhás....40·45n 23·04e
Langádhia....37·41n 22·02e
Lárimna....38·34n 23·18e
Lárisa....39·38n 22·25e
Lávrion....37·44n 24·04e
Lekhaina....37·56n 21·17e
Leonídhion....37·10n 22·52e
Levadheia....38·25n 22·54e
Levkás....38·50n 20·41e
Levkímmi....39·25n 20·04e
Liapádhes....39·40n 19·44e
Limín Vathéos...37·45n 27·00e
Linariá....37·24n 24·57e
Líndhos....36·06n 28·04e
Litókhoron....40·06n 22·30e
Livanatai....38·42n 23·03e
Lixoúrion....38·12n 20·26e
Loutrá Aidhipsoúx....38·51n 23·02e
Marathón....38·10n 23·58e
Megálon Khorián....36·27n 27·21e
Megalópolis....37·24n 22·08e
Mégara....38·01n 23·21e
Meligalás....37·13n 21·59e
Melívoia....39·45n 22·48e
Mesolóngion....38·21n 21·17e
Messíni....37·04n 22·00e
Methóni....36·50n 21·43e
Métsovon....39·46n 21·11e
Mikonos....37·26n 25·20e
Miléai....39·20n 23·09e
Milos....36·45n 24·27e
Mithimna....39·22n 26·10e
Mitilíni....39·06n 26·32e
Moláoi....36·48n 22·52e
Monemvasía....36·41n 23·03e
Moúdhros....39·52n 25·16e
Mouzákion....39·26n 21·40e
Náousa....40·37n 22·05e
Návpaktos....38·23n 21·50e
Navplion....37·34n 22·48e
Náxos....37·06n 25·23e
Neápolis....35·15n 25·37e
Neápolis....36·30n 23·04e
Néon Karlovásionx....37·48n 26·44e
Nigríta....40·55n 23·30e
Ólimbos....35·44n 27·11e
Orestiás....41·30n 23·40e
Palaiá Epídhavros....37·38n 23·09e
Palaiokhóra....35·14n 23·41e
Palamás....39·28n 22·05e
Panayía....39·56n 25·20e
Pánormos....37·38n 25·02e
Paramithiá....39·28n 20·30e
Párga....39·17n 20·23e
Páros....37·04n 25·08e
Pátrai....38·15n 21·44e
Pílos....36·55n 21·43e
Piraiévs (Piraeus)....37·57n 23·38e
Pirgos....37·41n 21·28e
Plomárion....38·59n 36·22e
Pogonianí....40·00n 20·25e
Políkastron....41·00n 22·34e
Políkhnitos....39·05n 26·11e
Políyiros....40·23n 23·27e
Préveza....38·57n 20·44e
Prosotsáni....41·10n 23·59e
Psakhná....38·35n 23·38e
Psará....38·46n 25·36e
Psárion....37·20n 21·51e
Ptolemaís....40·31n 21·41e
Rethimnon....35·22n 24·29e
Ródhos (Rhodes)....36·26n 28·13e
Salamís....37·59n 23·28e
Salonika, see Thessaloníki
Samothráki....40·28n 25·31e
Sápai....41·02n 25·41e
Sérifos....37·09n 24·31e
Sérrai....41·05n 23·32e
Sérvia....40·11n 22·00e
Siátista....40·16n 21·33e
Sidhirókastron...41·14n 23·22e
Sikiá....40·02n 23·56e
Síkinos....25·07n 26·43e
Símí....36·36n 27·50e
Sitía....35·12n 26·07e
Skála Oropoú....38·20n 23·46e
Skíathos....39·10n 23·29e
Skíros....38·53n 24·33e
Skópelos....39·07n 23·43e
Sofádhes....39·20n 22·06e
Spárti (Sparta)...37·05n 22·27e
Stilís....38·55n 22·36e
Thásos....40·47n 24·42e
Theólogos....40·39n 24·41e
Thesprotikón....39·15n 20·47e
Thessaloníki (Salonika)....40·38n 22·56e
Thíra....36·25n 25·26e
Thívai....38·21n 23·19e
Timbákion....35·04n 24·46e
Tínos....37·32n 25·10e
Tirnavos....39·45n 22·17e
Trikkala....39·34n 21·46e
Trípolis....37·31n 22·21e
Vasiliká....40·28n 23·08e
Velestínon....39·23n 22·45e
Véroia....40·31n 22·12e
Volissós....38·29n 25·58e
Vólos....39·21n 22·56e
Xánthi....41·08n 24·53e
Xilókastron....38·05n 22·38e
Yerolimín....36·28n 22·24e
Yiannitsá....40·48n 22·25e
Yítheion....36·45n 22·34e
Zákas....40·02n 21·16e
Zákinthos....37·47n 20·53e

Physical features and points of interest

Abdēra, *ruins*....40·59n 24·58e
Aegean Sea....38·30n 25·00e
Aíyina, *island*....37·46n 23·26e
Akhelóös, *river*...38·36n 21·14e
Akrítas, Ákra, *cape*....36·43n 21·54e
Alfiós, *river*....37·40n 21·33e
Aliakmon, *river*...40·30n 22·36e
Almiroú, Kólpos, *bay*....35·23n 24·20e
Ámbelos, Ákra, *cape*....39·56n 23·55e
Amorgós, *island*..36·50n 25·59e
Amvrakikós Kólpos, *bay*....39·00n 21·00e
Anáfi, *island*....36·21n 25·50e
Andikíthira, *island*....35·52n 23·18e
Andíparos, *island*....37·00n 25·03e
Ándros, *island*...37·45n 24·42e
Angathonísi, *island*....37·28n 27·00e
Aóös, *river*....40·05n 20·37e
Árakhthos, *river*..39·01n 21·03e
Argolikós Kólpos, *bay*....37·33n 22·45e
Astipálaia, *island*....36·35n 26·25e
Áthos, *mountain*.40·09n 24·19e
Áthos, see Áyion Óros
Attikí, *historic region*....38·00n 23·30e
Axiós (Vardar), *river*....40·35n 22·50e
Áyion Óros (Áthos), *peninsula*....40·15n 24·15e
Ayios Evastrátios, *island*....39·30n 25·04e
Chios, see Khíos
Corfu, see Kérkira
Corinth, Gulf of, see Korinthiakos Kólpos
Cos, see Kos
Crete, see Kríti
Crete, Sea of, see Kritikón Pélagos
Cyclades, see Kikládhes
Cythera, see Kírthira
Delos, *island*....37·22n 25·16e
Delphi, see Dhírfis Óros
Dhelfoí, *ruins*....38·30n 22·29e
Dhenoúsa, *island*....37·06n 25·50e
Dhíkti, *mountains*....35·08n 25·22e
Dhílos, *island*....37·22n 25·16e
Dhírfis Óros (Delphi), *mountain*....38·38n 23·49e
Dhodhekánisos (Dodecanese, Sporádhes), *islands*....36·35n 27·10e
Dhodhoni, *ruins*..39·34n 20·47e
Diá, *island*....35·28n 25·14e
Dodecanese, see Dhodhekánisos
Doïránis, Lake...41·13n 22·44e
Drámas, Pedhiás, *plain*....41·04n 24·00e
Epirus, see Ípiros
Erímanthos, *mountain*....37·59n 21·51e
Évros, *river*....40·52n 26·12e
Evrótas, *river*....36·48n 22·40e
Evvoia (Euboea), *island*....38·34n 23·50e
Faistós, *ruins*....35·01n 24·48e
Filí, *ruins*....38·10n 23·40e
Filippoi, *ruins*....41·00n 24·16e
Folégrandros, *island*....36·36n 24·56e
Foúrnoi, *island*...37·34n 26·30e
Gávdhos, *island*..34·50n 24·06e
Hydra, see Ídhra
Ídhi Oros, *mountain*....35·18n 24·43e
Ídhra (Hydra), *island*....37·20n 23·32e
Ierón Asklipioú, *ruins*....37·37n 23·02e
Ikaría, *island*....37·41n 26·20e
Iliodhrómia, *island*....39·14n 23·54e
Ionian Sea....38·30n 19·00e
Iónio Nísoi (Ionian Islands)..38·30n 20·30e
Ios, *island*....36·42n 25·24e
Ípiros (Epirus), *historic region*...39·40n 20·50e
Iráklia, *island*....36·50n 25·26e
Itháki, *island*....38·24n 20·42e
Kafirévs, Ákra, *cape*....38·09n 24·36e
Kálimnos, *island*.37·00n 27·00e
Kalmakchalan, *mountain*....40·55n 21·47e
Kámiros, *ruins*...36·19n 27·57e
Kamvoúnia Óri, *mountains*....40·00n 21·52e
Káros, *island*....36·52n 25·40e
Kárpathos, *island*....35·40n 27·10e
Kásos, *island*....35·22n 26·56e
Kassándra, *peninsula*....40·06n 23·22e
Kastorias, Límni, *lake*....40·30n 21·17e
Kéa, *island*....37·34n 24·22e
Kefallinía, *island*.38·15n 20·35e
Kérkira (Corfu), *island*....39·40n 19·42e
Khálki, *island*....36·17n 27·35e
Khalkidhikí, *historic region*...40·25n 23·27e
Khanían, Kólpos, *bay*....35·34n 23·48e
Khíos (Chios), *island*....38·22n 26·00e
Kikládhes (Cyclades), *islands*....37·00n 25·00e
Killíni, *mountain*.37·57n 22·23e
Kiparissiakós Kólpos, *bay*....37·37n 21·24e
Kíthira (Cythera), *island*....36·20n 22·58e
Kíthnos, *island*...37·25n 24·28e
Knossos, *ruins*...35·20n 25·10e
Korinthiakos Kólpos (Gulf of Corinth)....38·19n 22·04e
Korínthou, Dhiórix, *canal*...37·57n 22·56e
Korónia, Límni, *lake*....40·41n 23·05e
Kos (Cos), *island*.36·50n 27·10e
Kríti (Crete), *island*....35·29n 24·42e
Kritikón Pélagos (Sea of Crete)...35·46n 23·54e
Lakonia, *historic region*....37·00n 22·30e
Lakonikós Kólpos, *bay*....36·25n 22·37e
Lamia, Gulf of, see Maliakós Kólpos
Léros, *island*....37·08n 26·52e
Lésvos (Lesbos), *island*....39·10n 26·20e
Lévka Óri, *mountains*....35·18n 24·01e
Levkas, *island*....38·39n 20·27e
Límnos, *island*...39·54n 25·21e
Líndhos, *ruins*...36·06n 28·05e
Lipsó, *island*....37·20n 26·45e
Líthinon, Ákra, *cape*....34·55n 24·44e
Macedonia (Makedhonía) *historic region*...41·00n 22·00e

MAP INDEX (continued)

Maléa, Ákra, *cape*.....36·26n 23·12e
Maliakós Kólpos (Gulf of Lamia)..38·52n 22·38e
Mediterranean Sea.....36·00n 21·00e
Merambéllou, Kólpos, *bay*.....35·14n 25·47e
Mesarás, Kólpos, *bay*.....34·58n 24·36e
Messíni, *ruins*....37·11n 21·57e
Messiniakós Kólpos (Gulf of Messíni), *bay*....36·58n 22·00e
Metéora, *monastery*.....39·46n 21·36e
Mikínai, *ruins*....37·44n 22·45e
Míkonos, *island*..37·29n 25·25e
Mikrá Préspa, Lake.....40·46n 21·04e
Mílos, *island*.....36·41n 24·15e
Mirtóön Pélagos, *sea*.....36·51n 23·18e
Mistrás, *ruins*....37·04n 22·21e
Morea, see Peloponnesus
Moúdhrou, Kólpos (Moúdhros), *gulf*.....39·49n 25·14e
Mount Olympus, see Ólimbos
Náxos, *island*....37·02n 25·35e
Néstos, *river*.....40·41n 24·44e
Nísiros, *island*....36·35n 27·10e
Oíti, *mountain*....38·49n 22·17e
Ókhi Óros, *mountain*.....38·05n 24·25e
Olimbía, *ruins*....37·38n 21·41e
Ólimbos (Mount Olympus), *mountain*.....40·05n 22·21e
Óssa, *mountain*..39·49n 22·42e
Othonoí, *island*...39·50n 19·26e
Óthris, Óros, *mountains*.....39·05n 22·45e
Pagasitikós Kólpos (Gulf of Volos), *bay*.....39·15n 22·51e
Palaia Kórinthos, *ruins*.....37·54n 22·56e
Parnassós, *mountain*.....38·32n 22·35e
Párnis, *mountain*.....38·11n 23·42e
Parnon, *mountains*.....37·18n 22·35e
Páros, *island*.....37·08n 25·12e
Pátmos, *island*...37·20n 26·33e
Patraikós Kólpos (Gulf of Patras), *bay*.....38·14n 21·15e
Paxoí, *island*.....39·12n 20·12e
Pélla, *ruins*.....40·45n 22·33e
Peloponnesus (Morea), *historic region*...37·30n 22·00e
Petalión, Kólpos, *bay*.....37·59n 24·02e
Pílion, *mountain*..39·28n 23·02e
Píndhos Óros (Pindus Mountains).....39·49n 21·14e
Piniós, *river*.....39·54n 22·45e
Prespa, Lake.....40·50n 21·02e
Psará, *island*.....38·35n 25·37e
Rhodes, see Ródhos
Rhodope *mountains*.....41·30n 24·30e
Ródhos (Rhodes), *island*.....36·10n 28·00e
Saloniki, Gulf of, see Thermaïkos Kólpos
Sámos, *island*....37·48n 26·44e
Samothráki (Samothrace), *island*.....40·30n 25·32e
Saronikós Kólpos, *bay*.....37·54n 23·12e
Sérifos, *island*....37·11n 24·31e
Sídheros, Ákra, *cape*.....35·19n 26·19e
Sífnos, *island*....36·59n 24·40e
Sikión, *ruins*.....37·59n 22·44e
Síkinos, *island*...36·39n 25·06e
Simi, *island*.....36·35n 27·52e
Singitikós Kólpos, *bay*.....40·12n 24·03e
Síros, *island*.....37·26n 24·54e
Sithoniá, *peninsula*.....35·12n 26·07e
Skíathos, *island*..39·12n 23·28e
Skópelos, *island*..39·10n 23·40e
Skórpios, *island*..38·41n 20·45e
Smólikas, *mountain*.....40·06n 20·52e
Soúnion, Ákra, *cape*.....37·39n 24·02e
Spatha, Ákra, *cape*.....35·42n 23·44e
Sporádhes, see Dhodhekánisos
Strimón (Struma), *river*..40·47n 23·51e
Strimonikós Kólpos, *bay*.....40·40n 23·50e
Taínaron, Akra, *cape*.....36·22n 22·30e
Taíyetos Óros, *mountains*.....37·16n 22·12e
Teyéa, *ruins*.....37·29n 22·24e
Thásos, *island*...40·41n 24·47e
Thásos, *ruins*....40·46n 24·33e
Thermaïkós Kóplos (Gulf of Saloniki), *bay*...40·23n 22·47e
Thermopílai (Thermopylae), *battlefield*.....38·48n 22·33e
Thessalía (Thessaly), *historic region*...39·30n 22·00e
Thíra, *island*.....36·24n 25·29e
Thrace (Thráki), *historic region*...41·15n 26·15e
Thrakikón Pélagos, *sea*....40·15n 24·28e
Tílos, *island*.....36·25n 27·25e
Tínos, *island*.....37·38n 25·10e
Tirins, *ruins*.....37·36n 22·48e
Trikhonis, Límni, *lake*.....38·34n 21·28e
Vardar, see Axiós
Vardhoúsia Óri, *mountains*.....38·44n 22·07e
Vegorrítis, Límni, *lake*.....40·41n 21·44e
Vérmion Óros, *mountain*.....40·39n 21·53e
Voïviis, Límni, *lake*.....39·32n 22·45e
Volos, Gulf of, see Pagasitikós Kólpos
Vólvi, Límni, *lake*.....40·41n 23·23e
Voríai Sporádhes, *islands*.....39·17n 23·23e
Vórios Evvoïkós Kólpos, *bay*.....38·40n 23·15e
Yiáros, *island*....37·38n 24·44e
Zákinthos (Zante), *island*...37·52n 20·44e
Zakínthou, Porthmós, *strait*.....37·50n 21·00e

limestone crags of the islands, contrasting with the deep blue of the Aegean waters and an equally powerful sky; the dusty green olive groves, the burnt-orange tiled roofs, and dazzling whitewashed walls of coastal communities; and the ever-present weathered stones of the country's great number of ancient monuments all add their own tones.

The drainage pattern of Greece is significantly influenced by the porosity and solubility of the rocks of the limestone regions; hence, seasonal downpours are often immediately lost through seepage and runoff. Much rainfall is also lost in rugged terrain of the geologically young northern mountains, where there is a tortured network of rushing mountain streams, often falling into narrow, spectacular gorges. Finally, the irregular, deeply penetrating coastline makes for short river courses. The overall effect is to produce short rivers with an erratic seasonal flow, virtually useless for navigation and limited for irrigation purposes. The Vardar, Struma, and the Néstos, which crosses Macedonia and Thráki to enter the northern Aegean, are the major rivers—but only because they drain large regions beyond the Greek frontier. A host of small and medium-sized rivers drain the rest of the country: the Aliákmon, the Piniós (running east across the main peninsula), and the Evrótas of the Pelopónnisos are noteworthy.

Plant and animal life. Like other Balkan countries, Greece is open to influences from several major biogeographic zones, with the major Mediterranean influences supplemented by plants and animals stemming from the central European interior. Hence, local topographic and climatic conditions also occasion great variety. On the mountain flanks, and in the north generally, the central European types of vegetation prevail. In central and southern regions and in narrow belts along the valleys of the mountains, about half the land is under scrub of various kinds; and maquis, the classic Mediterranean scrub complex—with oleander, bay, evergreen oak, olive, and juniper—is particularly well developed in the Pelopónnisos. Evergreen trees and shrubs and herbaceous plants are found in the lowlands, with the flowers offering brilliant patterns in springtime. Pines, planes, and poplars line the rivers, the higher slopes, and the coastal plains. Oak, chestnut, and other deciduous trees are found in the north, giving way at higher altitudes to coniferous forests dominated by the Grecian fir, in which clearings are carpeted in spring and summer with irises, crocuses, and tulips. Forests and scrub are found at the highest levels: the black-pine forests coating Mt. Olympus are particularly noteworthy.

The forested zones, especially in the north, harbour such European animals as wildcat, martin, brown bear, roe deer, and, more rarely, wolf, wild boar, and lynx. Animals of the Mediterranean regions include jackals, wild goats, and porcupines, all adapted to lack of moisture and to the heat. Birds include pelicans, storks, and herons, while many varieties winter in Greece from farther north. Reptile and fish life is rich and varied.

The human imprint. The large number of monuments dotted across Greece are testimony of the antiquity of man's attempts to wrest a living from this sometimes harsh environment. Contemporary patterns of settlement, especially in the rural areas, bear the marks of long centuries of development.

Village life remains a powerful influence. It nevertheless has a cosmopolitan nature—seen especially in the village-square discussions—and the modern cities have something of the village in their character. The rural communities themselves range from the little communities of the northern mountain interior, reminiscent of central Europe, to the sun-beaten villages of Crete, which are almost African in appearance. Although rural settlement patterns reflect the vagaries of physical geography and the changing currents of history in a particular region, the tiled roofs, low, whitewashed walls, long, narrow windows, the central church, and the traces of fortifications are frequent features of village communities. Between villages, the ancient stone walls and winding roads add their own character to the landscape.

Yet town and city life is becoming increasingly important, and more than half of the people were classified as urban in the early 1970s. The metropolitan sprawl centred on Athens (home of over 2,500,000 people) is one of the great urban complexes of the Mediterranean, complete with industrial and port facilities. The port of Thessaloníki, with a metropolitan population close to 550,000, is the second major urban centre. There is then something of a gap, for the next half dozen or so major centres—often ports—have populations generally falling between 50,000 and 100,000. Urbanization and the modernization of the economy—especially improvements aimed at the influx of tourists—are also leaving their mark on the traditional Greek landscape.

Greece, Area and Population

	area		population	
	sq mi	sq km	1961 census	1971 census
Regions (*dhiamerísmata*)				
Aegean Islands				
Departments (*nomoí*)				
Cyclades	993	2,572	100,000	86,000
Dodecanese	1,044	2,705	123,000	121,000
Khíos	349	904	62,000	54,000
Lésvos	832	2,154	140,000	115,000
Sámos	300	778	52,000	42,000
Central Greece and Évvoia				
Departments				
Aitolía and Akarnanía	2,103	5,447	238,000	229,000
Attikí*	1,303	3,375	205,000	258,000
Evritanía	790	2,045	40,000	30,000
Évvoia	1,509	3,908	166,000	165,000
Fokís	819	2,121	48,000	41,000
Fthiótis	1,686	4,368	160,000	155,000
Voiotía	1,240	3,211	114,000	115,000
Crete				
Departments				
Iráklion	1,020	2,641	208,000	210,000
Khaniá	917	2,376	131,000	120,000
Lasíthi	702	1,818	74,000	66,000
Rethímni	578	1,496	70,000	61,000
Ípiros				
Departments				
Árta	622	1,612	83,000	78,000
Ioánnina	1,927	4,990	155,000	135,000
Préveza	419	1,086	63,000	57,000
Thesprotía	585	1,515	52,000	41,000
Greater Athens†	167	433	1,853,000	2,540,000
Ionian Islands				
Departments				
Kefallinía	361	935	46,000	37,000
Kérkira	247	641	102,000	93,000
Levkás	125	325	29,000	25,000
Zákinthos	157	406	36,000	30,000
Macedonia				
Departments				
Dráma	1,339	3,468	121,000	91,000
Flórina	719	1,863	67,000	52,000
Grevená	903	2,338	43,000	35,000
Imathía	656	1,699	115,000	118,000
Kastoría	651	1,685	47,000	46,000
Kavála	814	2,109	141,000	122,000
Khalkidhikí‡	1,267	3,281	83,000	75,000
Kilkís	1,003	2,597	103,000	84,000
Kozáni	1,375	3,562	153,000	136,000
Pélla	968	2,506	133,000	126,000
Piería	598	1,548	98,000	92,000
Sérrai	1,539	3,987	248,000	203,000
Thessaloníki	1,375	3,560	544,000	710,000
Pelopónnisos				
Departments				
Akhaía	1,239	3,209	239,000	240,000
Argolís	855	2,214	90,000	89,000
Arkadhía	1,706	4,419	135,000	111,000
Ilía	1,035	2,681	189,000	165,000
Korinthía	884	2,289	113,000	113,000
Lakonía	1,404	3,636	119,000	96,000
Messinía	1,155	2,991	212,000	173,000
Thessalía				
Departments				
Kardhítsa	995	2,576	153,000	134,000
Lárisa	2,067	5,354	231,000	232,000
Magnisía	1,018	2,636	164,000	161,000
Tríkala	1,289	3,338	143,000	133,000
Thráki				
Departments				
Évros	1,638	4,242	158,000	139,000
Rodhópi	982	2,543	109,000	108,000
Xánthi	692	1,793	90,000	83,000
Total Greece	50,960§	131,986§	8,389,000§	8,769,000§

*Attikí Department excludes area and population of Greater Athens, shown separately. †Constitutes part of Attikí Department. ‡Includes area and population of Áyion Óros (Mt. Áthos), an autonomous administration. §Detail does not add to total given because of rounding.
Source: Official government figures.

THE PEOPLE

Linguistic, ethnic, and religious background. Despite the great variety of influences that have shaped modern Greece, and the marked differences among the many isolated regions, a sense of community binds the Greeks together, especially in a national emergency, such as World War II. A common religion, a great heritage, and a common popular tongue (variations in regional dialects notwithstanding) tend to make all Greeks feel that they are somehow one people equal to each other in important respects. Combined with this sense of equality is the interest in, and appetite for, political discussion.

The sense of community

All but about 5 percent of the populace adheres to the Greek Orthodox Church (see EASTERN ORTHODOXY). This body appoints its own ecclesiastical hierarchy and is headed by a synod of 12 metropolitans under the presidency of the archbishop of Athens. The Greek Church has links in dogma with the other Orthodox churches. The Muslim minority, just over 1 percent of the populace, is mainly Turkish and is concentrated in western Thráki and the Dodecanese. Roman and Greek Catholics (concentrated in Athens and the western islands formerly under Italian sway) account for less than 0.5 percent, and there are a few thousand adherents of Protestant churches, the Gregorian Rite of the Eastern Church (mostly Armenians), and Judaism, the last named being much reduced in numbers by the German genocide of World War II.

In terms of ethnic composition, Greeks again make up all but 5 percent of the total, the remainder being composed of Macedonians, Turks, Albanians, and Romanians. Except in Cyprus, southern Albania, and Turkey, there are no major enclaves of Greeks in nearby foreign countries, although Greek communities play a distinctive role in Europe, the Western Hemisphere, and Australia.

Demographic trends. The Greek population has never displayed the high rates of growth attributed to it by some analysts, although—despite grievous losses in a succession of wars and constant emigration as a result of poor economic conditions—it has usually shown a regular increase since the first census, in 1828. Most of its growth in the years since it gained its independence from the Turks resulted from two factors—annexations of surrounding areas (the Ionian Islands; Thessalía [Thessaly] and Árta; Ípiros [Epirus], Macedonia, and Crete; Thráki [Thrace]; and the Dodecanese) and the influx of more than 1,000,000 Greek refugees from Asia Minor in the 1920s. Emigration continues to be a limiting factor, the most active periods having been 1911–15 when nearly 130,000 persons left, 1956–60 (160,000), and the decade of the 1960s (830,000). The commonest destinations of the emigrants are the United States, Canada, Australia, and, most recently, West Germany, which has also attracted the largest number of the 300,000 Greeks working temporarily in western Europe. Vital statistics are comparable to those of the developed countries of Europe —deaths are 8.5 and births 16.1 per 1,000 annually, giving a natural increase of 7.6 per 1,000. Higher birth rates after World War II, however, have produced a youthful population: about 25 percent are less than 15 years of age and 45 percent less than 30. Along with population growth, urbanization has been a most important factor, especially since World War II. The rural component has shrunk to about one-third of the total. All these factors have had important social and political implications. The average density of population is about 175 per square mile (68 per square kilometre), although the variation in natural conditions makes the average rather meaningless. It is perhaps more significant to note that more than eight out of every 10 Greeks live on the main peninsula, and in 1971 more than one Greek in four lived in the Greater Athens area. (G.An.)

THE NATIONAL ECONOMY

Although Greece is a Balkan country, its principal economic links are with the United States, the European Economic Community (EEC; Common Market), of which it is an associate member, and the Organization for Economic Cooperation and Development (OECD), of which it is a member. Trade with the Socialist bloc and with Yu-

goslavia has increased, however, and by 1974 about 16 percent of Greece's annual exports were going to those countries; they account, however, for only 5 percent of imports into Greece. With a population of only about 9,000,000 and a relatively low annual income per capita, Greece is not an important country industrially or in international trade. In the world of international shipping, however, Greek owners occupy an important place. If tonnage actually controlled by Greek nationals (as opposed to tonnage flying the Greek flag) is considered, the Greek fleet of about 4,380 ships, aggregating about 43,630,000 tons gross by 1976, was the second largest in the world (after Liberia). The fleet flying the Greek flag ranks fifth in the world, with more than 2,740 vessels aggregating 22,530,000 tons.

The extent and distribution of resources. *Mineral resources.* Greece's total mineral and quarry deposits are estimated at between 5,000,000 and 10,000,000 tons. There are important reserves of bauxite, lignite, and chromiferous iron, while the main quarry products are ceramic clays, gypsum, asbestos, and the famous Greek marble. Oil has been drilled for in several parts of the country, but commercially exploitable fields have not been found except in the northern Aegean Sea, and these have led to disputes with Turkey.

Greece is one of the main bauxite-producing countries in Europe. Proved reserves in mainland Greece amount to 40,000,000 tons, and probable deposits are estimated at more than twice as much. Output (mainly for export) doubled in the 1960s to approach 2,000,000 tons a year. A substantial and increasingly larger portion is used in the local production of alumina and aluminum metal. Total lignite (brown coal) reserves are estimated at 1,000,000,000 tons. About three-quarters of the lignite mined is used by power stations, the remainder for the manufacture of chemical fertilizers and domestic fuels. Ordinary iron ores and ores with a chrome and nickel content are present, but only the latter are exploited.

Biological resources. About 30 percent of the total area of Greece is arable land, about 40 percent is rough pasture land, 20 percent is forest, and the remainder is either unsuitable for cultivation or is unexploited.

In spite of the natural poverty of the soil and a system that leads to excessive fragmentation of the land (the average parcel is only about 1.4 acres, or 0.56 hectare), Greece is still essentially an agricultural country, about half the working population being directly engaged in farming and related activities. Agricultural products account for more than half the total exports, though Greece also imports substantial amounts of food.

Principal crops

The chief crops are wheat, cotton, tobacco, currants and seedless raisins, grapes, olives, and citrus fruit, production of which grew rapidly in the 1960s, partly owing to United States aid. Attempts are being made to improve dairy and meat production. Forestry is not important, though progress has been made in restoring the forests that were severely depleted during World War II. The very long coastline and numerous islands help nourish an important fishing industry. With the modernization of the fishing fleet and the extension of refrigeration and processing facilities, output has increased; the once-important sponge-diving industry, on the other hand, has declined in the face of competition from synthetic products.

Power resources. Waterpower and lignite are the most important power resources in Greece, about 70 percent of the electric power being derived from them. A power station at Megalópolis in the Pelopónnisos successfully utilizes lignite of remarkably low calorific value and high moisture content. This plant has interested specialists in other countries dependent upon low-grade lignite deposits. Surveys are still being carried out (with United Nations aid) with the aim of developing more hydroelectric stations, and the establishment of a nuclear generating plant has been under consideration.

Sources of national income. *Agriculture, forestry, and fishing.* The agriculture, forestry, and fishing sector accounted for 31 percent of the gross domestic product (GDP) in 1958, but thereafter it declined steadily. The decline was caused by the drift of workers from the land, either to other employment in Greece or to work abroad. In the period from 1961 to 1971 employment in agriculture declined at an average annual rate of 1.4 percent. Productivity, nevertheless, has been growing faster in this sector than in industry and in the services sector. As a result, agricultural production actually increased on average by nearly 3 percent a year in the period from 1965 to 1970, notwithstanding its decline as a proportion of total national output. Despite the comparatively more rapid growth of productivity in agriculture, however, its level of productivity is still a good deal lower than that of the average in other sectors—a factor experienced by the agricultural workers themselves as poverty. In an attempt to alleviate this, the government has tried to shift the emphasis away from excess production of wheat and tobacco to more profitable crops, notably cotton. Wheat subsidies were abolished in favour of a system of intervention price that sets a minimum selling price guaranteed by the government, which will buy wheat if prices fall below it and sell such purchases at free market prices. Farmers have also been compensated for loss of income by the waiving of debts and by grants to help pay for seeds, fertilizers, and agricultural machinery.

Government intervention in agriculture

Mining and quarrying. The mining and quarrying sector employs about 0.6 percent of the total labour force and accounts for just over 1 percent of the GDP. According to the industrial census of 1969, the number of mines and quarries (including quarries of building materials) and of solar-evaporation salt plants was 1,604, and average annual employment was 22,600 workers. In 1973 the value of the mining output was equivalent to 5.6 percent of total industrial output. In volume terms, more than one-half of the total output is accounted for by lignite production. The mining of bauxite is playing an increasingly important part in the economy, though if production of aluminum reaches the level planned for the mid-1970s, bauxite may have to be imported.

Manufacturing. Manufacturing output has tended to expand at a rate of about 10 percent per annum, but, largely because of the rapid growth of the services sector, manufacturing output as a proportion of the GDP remained around the level of 16 percent during the early 1960s; by the mid-1970s, however, it had risen to more than 25 percent of the GDP. Employment in the industrial sector as a whole (including building, mining, power production, and manufacturing) grew in the 1960s at an average annual rate of just over 4 percent, and over the latter half of this period productivity grew at the rate of 4.3 percent a year.

The main branches of industry are food processing, textiles, chemicals, steel, aluminum, and handicrafts. Athens and Thessaloníki are the main industrial centres, but government policy is to encourage industry to develop in other areas as well. The main change in the pattern of industry, however, has been the establishment of large industrial complexes, notably the huge Esso–Pappas complex at Dhiavatá, near Thessaloníki, which started production in 1966. Steel production is about 600,000 tons a year. The French-owned Pechiney complex at Dhírfis (Delphi) produces large quantities of aluminum for export. Another important industry, located at Elevsís (Eleusis), near Athens, is shipbuilding and ship refitting.

Products of industry

Energy. Electricity and gas production accounts for 2 percent of the GDP. In 1950, when 665,000,000 kilowatt-hours of electricity were produced, only 823 towns and villages in Greece were supplied with electricity; the Athens area accounted for 84 percent of the country's total consumption. In that year was founded the Public Power Corporation, on the initiative and with the aid of the United States; its first plant started operating in 1953. By 1970 a further 6,657 remote villages were receiving electricity, and the output was 9,000,000,000 kilowatt-hours; Athens then accounted for only 40 percent of consumption. Greece is still heavily dependent on imports for fuel supplies.

Financial services. The central bank is the Bank of Greece, founded in 1928. The bank is privately owned, but the governor is appointed by the government, and the state has a share in the profits. The Bank of Greece

is also the sole bank of issue, and it supervises all banking operations in the country, implementing the decisions made by the Currency Committee. Consisting of the governor of the bank and the ministers of coordination, finance, agriculture, and industry and commerce, this body formulates all monetary, credit, and foreign-exchange policy.

There are eight commercial banks, which are private competitive institutions. In addition to undertaking short-term financing, they extend long-term loans to industry and trade and sometimes take up shares in the equity capital of industrial firms. Two of the commercial banks have also set up special investment banks. There are also an agricultural and a mortgage bank. The latter, besides granting housing loans, arranges for loans to public corporations and to the tourist sector of the economy. Long-term industrial development is also promoted by the Hellenic Industrial Development Bank, which is wholly state owned and endowed with government funds. The capital market is still relatively underdeveloped, though a number of measures have been introduced to improve matters. There is one stock exchange, located in Athens, in which there is active trading in about 30 banking and commercial shares.

Foreign trade. Over the 1960s exports more than doubled, as did imports. Exports also increased at a somewhat faster rate than imports; however, because imports were generally about three times as large as exports, the trade gap grew steadily throughout the decade, until it was equivalent to more than 18 percent of national income. An encouraging feature of trade trends in the second half of the 1960s was the growth in exports of industrial and handicraft products; exports of aluminum were particularly important. Exports of tobacco, once Greece's largest export earner, declined, and the market appears to be saturated. Items of capital equipment and fuel accounted for an increasing share of imports, but imports of manufactured consumer goods (and also food) also expanded, in spite of government attempts at restriction. The resulting inflation contributed to the fall of the military junta in 1974.

The widening trade gap

Although the Greeks have expressed dissatisfaction from time to time at the effects of their association (as associate member) with the EEC, trade with this area has grown steadily since 1964, when EEC countries absorbed 38 percent of Greek exports and supplied 42 percent of imports. West Germany is Greece's main trading partner, taking about 19 percent of the country's exports and supplying 20 percent of its imports.

Although Greece has a persistent and increasing trade deficit, the balance of payments is helped by a steady surplus on invisible services, notably shipping, foreign tourism, and funds sent back home by workers who have emigrated to other countries.

Management of the economy. *Private enterprise and the role of the government.* The Greek economy remains predominantly one of private enterprise, with the state limiting its intervention to the field of tax incentives, cheap finance, and its own or bank guarantees to overseas suppliers (and then only for large projects). With the exceptions of electric power, railways, telecommunications, and broadcasting, the state produces neither goods nor services, and its policy is to refrain from activities that can be undertaken effectively by private firms.

It has been official policy to aim at a surplus on the ordinary budget and to use this surplus to finance part of the deficit on the public investment budget, the rest of this deficit being met by domestic issues of bonds and interest-bearing treasury bills and by borrowing abroad. Direct taxes contribute about 20 percent of the total revenues of the Greek fiscal system; between 1957 and 1966 direct taxes on households rose by more than 8 percent annually. Over the same period net indirect taxes rose on average by more than 13 percent, and their share in the total gross national product at market prices rose from 9.7 percent in 1957 to nearly 14 percent 10 years later. In the 1970s policy appeared to aim at more efficient collection of existing direct taxes instead of alteration of the fiscal structure as such.

Taxation

Trade unions and employer associations. The military government dissolved most trade unions and deported many union leaders soon after it came to power in 1967. After the fall of that government, trade union activity revived. The interests of employers are promoted by the Federation of Greek Industries, founded in 1907, and by the Federation of Shipowners and Industrialists. There are chambers of commerce in Athens, Piraiévs (Piraeus), and Thessaloníki.

Economic policies. The broad aim of economic policy is to secure growth through augmenting the share of industry in national output, while keeping prices as stable as possible. Since 1967 great emphasis has been put on the latter objective, and prices were held down by administrative measures for three years. Beginning in 1970, however, prices began to edge upward. In agriculture, the aim is to raise productivity through switching to more profitable crops, and the government no longer subsidizes the growing of wheat and tobacco at uneconomic prices. Under the first Five-Year Plan of the military regime of 1967–74, investment in infrastructure, in tourist facilities, and in industry was meant to be considerably increased, and the primary source of finance for the projected investments was intended to be found in domestic budgetary savings.

At the same time, however, great efforts were made to encourage private foreign investment. This was done not only in order to supply much-needed investment capital and technical and managerial expertise but also to balance the external-payments account by meeting the current deficit with a large inflow of capital. The government also did all it could to promote tourism and shipping, which are major sources of earnings.

In October 1975 a flexible Five-Year Plan (1976–80) was outlined, aiming at a 6 to 7 percent annual growth rate during the plan period and providing for fuller use of domestic energy resources, exploitation of mineral resources, and more active solicitation of foreign investment involving export industries.

The chief economic problem for Greece is the age-old one of lack of natural resources. The structure of employment and of the national income shows Greece's great dependence on secondary economic activities: only about 25 percent of the working population is engaged in manufacturing and construction, and very nearly half of the GDP is generated in the services sector. This leads to a persistent balance-of-payments problem, and foreign indebtedness has grown steadily.

Resource scarcity

(E.I.U./Ed.)

Transportation. Not unexpectedly, the Greek landscape and seascape have had immense effects on the development of transportation patterns in the country and perhaps help to account for the pre-eminence of Greek names in the world of international shipping. The needs of tourism, military and political considerations, and the general economy have helped stimulate the development of a modern national transport system.

Sea transport. The importance of Greek (and Greek-owned) shipping fleets in national and international trade was noted above. Partly as a result, there has been strong emphasis on port development. Piraíevs (Piraeus), the port of Athens, is the major centre, followed by Pátrai, Préveza, Iráklion, Kaválla, and Vólos, among others. There is a developed steamer service to the various islands, and car ferries ply across many of the straits and inlets. Cruise ships and private vessels add to the demand for service facilities at ports.

Road transport. About 50 percent of the road network remains unpaved, and many of the smaller rural roads, especially in the mountain regions, still leave much to be desired. The Athens–Pátrai, Athens–Lamía, and Athens–Thessaloníki highways are modern roads, and the total amount of paved road tripled during the 1960s. There is also an extensive network of rural bus routes.

Railway transport. Extensive modernization has been effected in the Greek railway system, the aims being to improve the existing tracks, to standardize differing metric gauges, to forge links with western Europe, and to coordinate development with that of the roads. The

Greek topography and a late start (railways date only from the 1880s, and Greece was one of the last European countries to develop them) have made this modernization costly and difficult.

Air transport. Air transport is operated by the government-owned Olympic Airways, which took over the financially troubled Greek National Airlines (founded 1951) in 1957. Increasing tourist traffic has resulted in a major expansion of facilities. Athens has a modern air terminal, regional facilities have been improved, and there are new airstrips on some of the islands. **(G.An.)**

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. *The constitution.* The military junta (1967–74) ruled under a constitution approved in 1968 by a referendum held under martial law and preceded by a vigorous campaign in its behalf in the censored press. Political activity was prohibited, and the junta ruled by decree.

The republic

On June 1, 1973, the junta decreed that the monarchy was abolished and replaced by a republic. This was confirmed by popular referendum on July 29, but the validity of this referendum was challenged. After the fall of the junta and the restoration of democratic government, another referendum was held, on December 8, 1974, and 69.2 percent of those voting chose "uncrowned democracy."

On August 1, 1974, the prime minister, Konstantinos Karamanlis, announced that the constitution of 1952 would be reintroduced as a provisional measure until, in his words, "the country acquires a charter fully approved by the people."

The powers of head of state were to be exercised by the president of the republic. The word king in the constitution was replaced by the word president. Under the constitution of 1952 the government had full control over the armed forces, and the judiciary was free.

The draft of a new constitution was published on December 23, 1974, after endorsement by the full Cabinet. It provided for Greece to be a parliamentary republic, with a president as head of state and supreme commander of the armed forces. Legislative power was to be exercised jointly by the president and Parliament, the latter a unicameral body composed of 200 to 300 deputies (the exact number to be determined by Parliament itself) elected for five-year terms. The freedom to form political parties was guaranteed.

The president was to be elected by Parliament for a five-year (later reduced to four-year) term, and he was eligible for a second term.

The president was to appoint as prime minister the leader of the party with an absolute majority in Parliament; if such did not exist, other methods were prescribed. The draft constitution would also create an advisory body, the Council of the Republic, to consist of elder statesmen and the current leaders.

Administration. The Cabinet is composed of the prime minister and ministers of foreign affairs, national defense, coordination and planning, public order, culture and science, justice, national education and religion, employment, social services, interior, finance, agriculture, industry, commerce, public works, transportation and communication, and mercantile marine. There are also a minister for northern Greece and a minister to the prime minister.

Dhiamerísmata and *nomoí*

Officially, Greece is divided administratively into 10 *dhiamerísmata* (regions), although only four (and part of a fifth) have their own governments. A further subdivision is into more than 50 *nomoí* (departments; singular *nomós*); special arrangements are provided for the Greater Athens area, and the peninsula of Áyion Óros (Mt. Áthos) is a self-governing monastic community with a civil governor, appointed by the government, who is responsible for public order outside the monasteries.

Municipal government, with elected mayors and urban and rural councils, was established in Greece in the mid-19th century. Local authorities may levy certain taxation, but, in general, provincial services are supported by the national government.

Political parties. After the downfall of the military junta in 1974 and the restoration of democratic government, two decrees, published on September 23, 1974, authorized the resumption of party political activities in Greece. Four major new parties emerged:

The Centre Union–New Forces Party, a merger of the Centre Union and the Movement of New Political Forces. The Centre Union represented the rump of the Greek liberal movement, the Centre Union Party (founded in 1961) of Georgios Papandreou. The new Movement of New Political Forces was founded to campaign for policies of "democratic socialism."

The New Democracy Party, founded by Konstantinos Karamanlis. It was pledged to work for the establishment of democracy in Greece through political, economic, and social reforms.

The Pan-Hellenic Socialist Movement (Pasok), incorporating two resistance organizations and founded by Andreas Papandreou, son of Georgios Papandreou. Its platform called for a non-aligned foreign policy (opposed to both the North Atlantic Treaty Organization and the EEC) and nationalization of private enterprises.

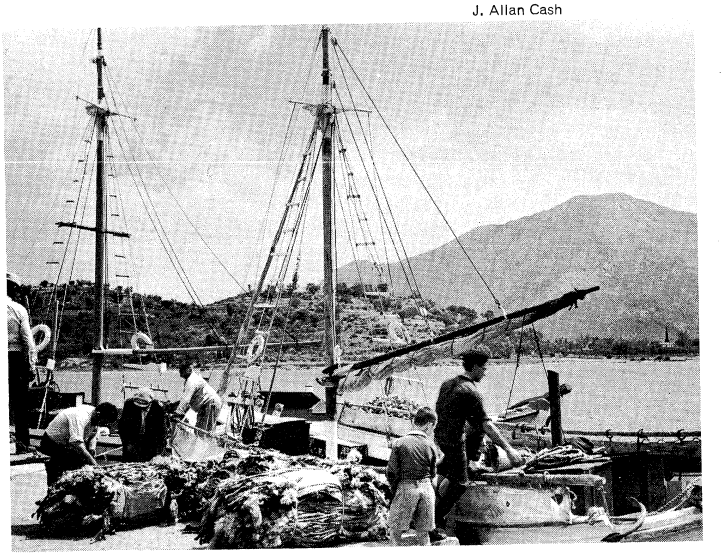
The United Left, a union of the United Democratic Left (EDA) and the two factions of the Communist Party of Greece (KKE). The EDA had been founded originally in 1951.

There were also a number of smaller parties representing various right-wing groups and groups that favoured the restoration of the monarchy.

New Democracy Party in power

Parliamentary elections were held on November 17, 1974, the first since 1964. The New Democracy Party of Karamanlis won 54 percent of the votes cast and a very large majority of the seats in Parliament; the next largest number of seats went to the Centre Union–New Forces Party (20 percent of the vote). On December 9, 1974, the Greek Parliament met for the first time in seven years.

J. Allan Cash



The port of Palaiá Epídhavros (Epidaurus), Greece.

Justice. The Supreme Court consists of four sections, three civil and one criminal. There are 11 courts of appeal, having jurisdiction in cases of criminal and civil law of second degree; in exceptional cases they may also adjudicate in cases of first degree. Below these are the 58 courts of first instance, which function also as criminal courts. They have jurisdiction in cases of first degree and, in exceptional cases, second degree. Criminal and juvenile courts exist in towns where courts of first instance sit. Some towns also have tax courts. There are many courts of justices of the peace throughout Greece, and also magistrates' (police) courts. District attorneys function in all courts except justices of the peace and magistrates'; in magistrates' courts the duties of district attorney are carried out by a public prosecutor.

Procedure in the courts follows the French model; *i.e.*, the investigating magistrate examines the evidence and

interrogates witnesses and, if he decides that there is a prima-facie case, refers it to the public prosecutor, who decides whether or not a charge shall be brought. Judges of the higher courts are appointed for life, and others may be removed only if convicted of a criminal offense.

In addition to the regular courts, there is a State Council, having jurisdiction over administrative disputes, administrative contraventions of law, and revision of disciplinary procedure on permanent civil servants.

Police. There are two main bodies of police, the city police (in Athens, Piraiévs, Pátrai, and Kérkira) and the gendarmerie, the latter a paramilitary force with responsibility for the whole country outside the four cities. Both forces are administered by the minister of the interior. The police forces were reorganized by a British mission after World War II. The gendarmerie and city police provide personnel for the tourist police. In addition, there are small forces of farm police, customs guards, and forest police and a Harbour Corps.

Armed forces. The Greek armed forces consist of an army, navy, and air force recruited on the basis of compulsory military service for all male citizens aged 21, for a period of two years. There is also a National Guard recruited from reservists for local security duties in frontier districts.

In 1967 a clique of army officers, mainly colonels, seized power in Greece by means of a coup d'etat. After the restoration of parliamentary government in 1974, Prime Minister Karamanlis stated (December 11), in a policy announcement, that the aim of the government would be to modernize the armed forces and ensure that they "recovered their concord and discipline." A gradual review of all members of the armed forces was begun, and in March 1975 a reorganization got under way. Clearly, it would be some time before the government could be confident of the loyalty of the military services.

Social conditions. Wages in Greece are low by reason of chronic rural underemployment and because the low agricultural income leads many young peasants to seek employment in the towns and there swell the ranks of unskilled labour. Minimum wages are fixed by the Ministry of Employment.

The Workers' Housing Organization, an agency of the national government, constructs housing and also provides technical assistance and loans for home building. Its projects are largely financed by a national lottery.

Health. After World War II the government took the lead in combatting disease, establishing modern health services and facilities. Malaria, once a scourge, has been virtually eradicated, and standards of hygiene and sanitation have been greatly improved. In the larger cities and towns the water supply is generally safe, though the same claim cannot be made for rural areas. Health measures taken in conjunction with the World Health Organization have been responsible for a great decline in deaths from infectious diseases. The main causes of death are cancer, cerebrovascular disease, and senility.

The Ministry of Social Services is responsible for the maintenance of hygiene, malaria control, establishment and financing of health and medical care centres and pharmacies, and for providing drugs, maternal and child care, and mental care. The large hospitals are concentrated in Athens, Thessaloníki, and Pátrai.

Social insurance. In 1968 the government unified the complex social insurance schemes, which are now controlled and supervised by the Ministry of Social Services through the Social Insurance Institution (IKA), the Unemployment Insurance Organization, the Farm Insurance Organization, and several separate main and auxiliary semi-public funds. The IKA insures workers in government, business, and industry, some agricultural workers, nonfarm self-employed persons, and domestic workers. The main and auxiliary funds offer insurance programs by occupation—*e.g.*, a fund for seamen, a fund for newsboys in Thessaloníki.

Social Insurance Institution (IKA)

Social insurance costs are largely financed by contributions paid into IKA by employees, in the form of payroll deductions, and by matching contributions from employers. Taxes on manufactured products and income from investments made by the insuring agency also help pay for social insurance.

Education. Education is the responsibility of the state, through the Ministry of National Education and Religion, policy being formulated by a Supreme Board of Education. In 1964 a new Education Act introduced radical changes, making school attendance compulsory till the age of 15 and covering both primary school (six years) and part of secondary school (three years). It also made demotic Greek (the colloquial language) the main language of instruction throughout the school system. The military junta in 1967 practically abolished the act of 1964, requiring attendance at school for only six years and reinstating Katharevusa (the literary language) as the principal teaching medium.

Junior secondary schools offer a three-year program, either a general academic course or a more technically oriented course, the former intended for students who plan either to end their education with secondary school or to continue in higher education, the latter for those planning to enter higher technical or vocational schools. Beyond the junior secondary schools are the advanced secondary schools, also offering a three-year program.

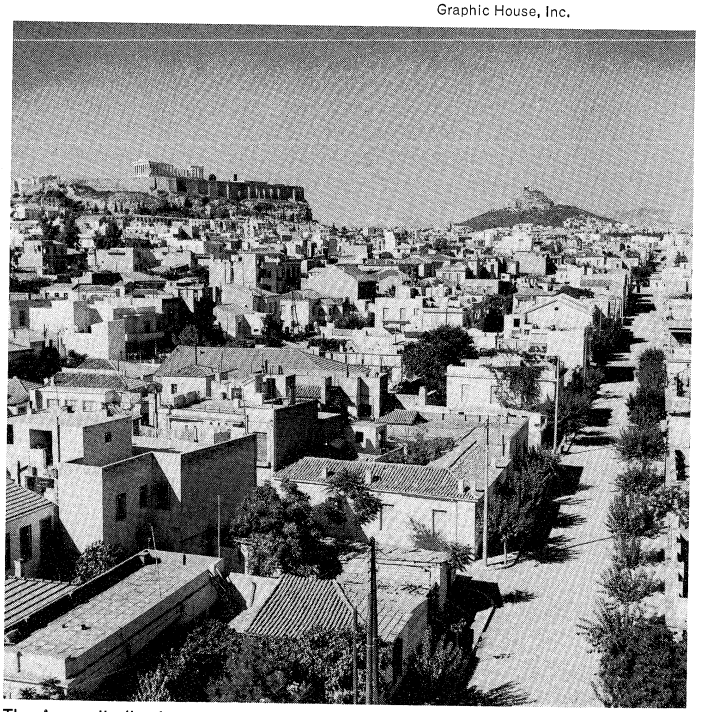
Universities

For higher studies there are universities at Athens (National Capodistrian University of Athens, founded in 1837, and National Technical University of Athens, founded in 1836) and at Thessaloníki (Aristotelian University of Thessaloníki, founded in 1925). The University of Ioánnina (1964) and the University of Pátrai (1966) are newer institutions. Athens also has a number of higher schools with university status: the Higher School of Fine Arts (1836), the Athens Graduate School of Economics and Business Science (1920), the College of Agriculture (1920), the Graduate School of Industrial Studies (1938), the American School of Classical Studies at Athens (1881), the Panteios School of Political Sciences (1930), and Pierce College (1875). The Graduate School of Industrial Studies at Thessaloníki (1958) also has university standing. (Ed.)

CULTURAL LIFE

The physical remains of the culture of ancient Greece, whether preserved *in situ*, in the fine network of museums, or (as a result of past activities) in the museums of other countries, are an ever-present reminder of the country's classical heritage. It continues as an important element in the culture of contemporary Greece and plays

Graphic House, Inc.



The Acropolis (background), in Athens.

a major role, through its attraction to tourists, in the economy. In addition, the deep religious traditions of the country—which found rich expression in the medieval icons and in the mosaics and frescoes that made the 14th century one of the triumphant eras of Byzantine art—continue to provide a fertile cultural source, generating a great variety of contemporary folk art and religious festivals. These are at their most vigorous in the rural areas generally and in the remoter regions in particular. Easter is the major event in the Orthodox calendar, and the sombre processions of Good Friday are followed by the festivities—including roasting of lambs and dancing in traditional costume—celebrating Easter Sunday. Other religious ceremonies occur throughout the year, beginning with the New Year blessing of the sea at Piraiévs (Piraeus) and elsewhere. The summer months are renowned for international music and drama festivals, particularly at Athens and Palaiá Epídhavros (Epidaurus). A revival of Byzantine iconography may be seen in the work of such artists as Fotos Contoglou and Stathis Trahanatzis as well as in the restoration of medieval frescoes and mosaics. The traditional Karankiózis puppet theatre is preserved by such masters as Panayioti Mixopoulos.

Modern Greek writers

Modern Greek poetry is considered by many to be among the best of the 20th century. Poets of international renown include Constantine Cavafy (who spent most of his life in Egypt), George Seferis (who won the Nobel Prize for Literature in 1963), Angelos Sikelianos, Odysseus Elytis, and Yannis Ritsos. Another remarkable writer was Nikos Kazantzakis, who experimented with several forms of expression. Contemporary Greek literature is fortunate in its English translators, especially the Americans Kimon Friar, Edmund Keeley, Themis Vasils, and Theodora Vasils.

The Greek character is reflected in the opening verses of the "Hymn to Liberty" written by the 19th-century poet Dhionísios Solomós. These lines, set to the stirring music of N. Manzaros as the national anthem of Greece, read (in a prose translation by Theodora Vasils),

> I know you by the sword's dread cutting edge, I know you by the look that with vigour measures the earth. Out of the sacred bones of the Hellenes you issue valiant as before, hail o hail, Liberty! You dwelt therein, sorrowful, withdrawn, waiting for a mouth to tell you, "Come again." That day was long in coming, and all were silent, cowering under the terror and the crush of slavery.

The other great "national" song of the Greeks is the fervent Resurrection hymn attributed to an 8th-century monk, St. John of Damascus,

> Christ has risen from the dead, by death trampling upon death, and has upon those in the tombs life bestowed.

The music of these two hymns, a millennium apart in composition, is both joyful and haunting. The listener is reminded of Greek nostalgia, of the deepest fears and hopes of the Greeks and indeed of all mankind.

Greek musicians with international followings include Manos Hadjidakis and Mikis Theodorakis. After the fall of the military regime in 1974 there was a revival in Athens of an irreverent political theatre as well as of the film industry. It remains to be seen what effect a decade of suppression will have on young literary talent. It also remains to be seen what effect the constantly growing tourist industry will have on the cultural integrity of the country.

THE OUTLOOK

The future of Greece remains in serious question, and in the decades ahead the fundamental alternatives facing the Greeks may be the exciting dangers of Balkan and eastern Mediterranean politics or the complacent prosperity of a closer association with the European community. Forces may already be at work among the Greeks that compel them toward further urbanization and industrialization and toward greater exposure to the homogenizing (if not demoralizing) "culture" of the international mass media.

Continuity with the past, a very long past, remains Greece's burden as well as its glory. The past and its significance are often obscure but always present. Their very language constantly reminds Greeks both of what they have been and of what they have aspired to. The ramifications of old influences go deep and are unpredictable; they seem to make permanent solutions impossible, so long as Greece retains its identity. Perhaps it is true for all peoples that there are no "permanent solutions" and that most of what can be done in the present depends intimately on what has happened in centuries past. But these limitations upon self-determination are much more evident in Greece than in most other countries. After all, the light of Greece has long been known to make many things clearer there than they are likely to be elsewhere. (G.An.)

BIBLIOGRAPHY. The most comprehensive single-volume work on Greece in all its aspects is vol. 10 of the *Megali elleniki enkyklopaideia, Ellas*, rev. ed. (1965), with bibliographies that are numerous and extensive but that often do not incorporate works more recent than the early 1930s. A comprehensive and less dated guide to sources is J.E. BAXEVANIS, *Modern Greece: A Bibliography* (1964). The two most extensive works on the geography of Greece are ALFRED PHILLIPSON, *Die griechischen Landschaften*, 4 vol. (1950–59), and the two-volume study *Synchroni geografia* (1965) by P. RHODAKIS and K. TRIANTAPHYLLOS. In English, the BRITISH ADMIRALTY NAVAL INTELLIGENCE DIVISION, *Greece*, 3 vol. (1944), is still the best summary of conditions to that time; later events and trends may be studied in JOHN CAMPBELL and PHILIP SHERRARD, *Modern Greece* (1968). For publications providing the statistical basis for analysis of Greece's population, society, and economy, see works published by the SOCIAL SCIENCES CENTRE (Athens), the CENTRE OF ECONOMIC RESEARCH (Athens), the NATIONAL BANK OF GREECE, and the NATIONAL STATISTICAL SERVICE OF GREECE (chief of which is the annual *Statistical Yearbook of Greece*) and the publications of various European organizations, especially the ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD).

The land. In addition to the general geographical works cited above, reference should be made to *Geological and Physicogeographical Bibliography of Greece* (*Géologiki kai fisikogeografiki vivliografia tis Ellados* (1961), by D. HARALAMBOUS; specific publications dealing with individual aspects of the landscape include E.G. MARIOLOPOULOS, *Klimatografia ton diaforon periochon tis Ellados* (1960), a comprehensive discussion of climate; there is, however, only a very meagre literature on Greece's animal and plant life. Soils are covered in D.S. KATAKOUSINOS, *Les Sols de Grèce* (1963).

The people. The monograph *Greece* (1974) prepared by D. TRICHOPOULOS and G. PAPAEVANGELOU for the World Population Year 1974 gives an authoritative picture of Greece's population both currently and in historical perspective; other useful studies include B. KAYSER, *Géographie humaine de la Grèce* (1964), and D. PENTZOPOULOS, *The Balkan Exchange of Minorities and Its Impact upon Greece* (1962). Publications of the censuses of 1961 and 1971 provide both basic data and analysis.

The national economy. Characteristics of Greece's people and economy are illuminated in B. KAYSER and K. THOMPSON, *Economic and Social Atlas of Greece* (1964), and in H. HOCHHOLZER, *Industrial Atlas of Greece* (1966). Greece's economy is surveyed regularly in the publications of the OECD, COMMERCIAL BANK OF GREECE, NATIONAL BANK OF GREECE, and ATHENS CHAMBER OF COMMERCE AND INDUSTRY as well as in those of the national government and various ministries. The five-year plans provide a general overview and analysis. Other useful surveys include *Problems of Greek Regional Development* (1962), by BENJAMIN WARD, and *Long-Term Prospects for the Greek Economy: A Forecast of Development in the Next 15 Years* (1968), issued by the ROYAL HELLENIC RESEARCH FOUNDATION.

Social conditions. Changing Greek social conditions as well as the relatively unchanging Greek countryside, people, and customs are described in PATRICK L. FERMOR, *Mani: Travels in the Southern Peloponnese* (1958); TIMOTHY WARE, *The Orthodox Church* (1963); ERNESTINE FRIEDL, *Vasilika: A Village in Modern Greece* (1962); JOHN K. CAMPBELL, *Honour, Family and Patronage: A Study of Institutions and Moral Values in a Greek Mountain Community* (1964); BERNARD KAYSER, PIERRE-YVES PECHOUX, and MICHEL SIVIGNON, *Rural Exodus and Urban Attraction in Greece* (1971); GEORGE ANASTAPLO, *The Constitutionalist* (1971), includes citations to 10 of the author's articles on contemporary Greek affairs republished in the *Congressional Record*, and his *Human Being and Citizen* (1975) has discussion of Greek affairs and character. (Ed.)

Greek Civilization, Ancient

The special role of Greece in the history of Western civilization is easier to demonstrate than to explain. The land itself offered no particular advantages such as far earlier had determined the importance of the great valley civilizations of the Nile, Mesopotamia, or the Indus. It plays little part in the history of early man. Its Bronze Age history has something to contribute to the history of civilization but is trivial beside the great contemporary empires of the East and Egypt. Something that may help explain why Greece emerged as a great civilization is the happy combination of geography and temperament: a geographical location that permitted it to receive the benefits of the skills and resources of older civilizations and contemporary civilizations to the north, east, and south, and yet that was not especially vulnerable to attack by them; and a spirit of inquiry, bred, perhaps, in far earlier generations of nomadic life, which was manifested by the Greeks' exploration, travel, and colonization and by a speculation about man and his world, which soon led to a sound basis for philosophical, political, and even scientific theory. Another important factor was a language that developed into the most subtle instrument for the expression of thought, science, and emotion that the West has known.

Achievements of Greek Civilization

The achievements and legacy of Greek civilization can be assessed under many different heads, and yet, the most significant contribution was probably an attitude of mind and an approach to political life rather than works of art and literature. The autocratic dynasties of the East and Egypt with their royal and priestly hierarchies could provide no model for politics in Greece. Instead, the small kingdoms progressed from systems of rule by families or dictators (in the modern sense) to rule by elected assemblies and councils and to recognition of the virtues of democracy, as well as of its practical shortcomings. The practice of government came to be systematically analyzed for the first time, and the systems of different contemporary states compared. The problems of the slaves and the poor could be recognized, if not solved. In the fine arts, the technical skills of the Greeks were mainly learned from foreigners, except perhaps for the art of mural painting. It was in terms of conception and precision that Greek artists outstripped their teachers and set standards of artistic expression and craftsmanship that seem almost superhuman. Their special contribution to narrative art lay in their genius for giving expression to subtleties of mood and feeling while seemingly confining themselves to mere reportage. Their graphic art exhibits similar qualities, expressing a perfect compromise between, on the one hand, a sense of pattern and proportion, and, on the other, shrewd observations of man and the world around him.

Ancient Greece was the most literate of all ancient cultures—witness the profusion of public and private inscriptions, from state decrees to graffiti on walls. Some of the earliest uses of the newly learned alphabetic script were for epic poetry; for law codes; and for lyric poetry, private and public, on themes of state, war, humour, and love. Eventually, there emerged choral and dramatic treatments of the traditional myths, done in a spirit of moral speculation, religious affirmation, or sheer irreverence. The great prose histories are models of disinterested inquiry and of a critical concern for recording man's motives as well as his behaviour.

The attempt of Greek philosophy to provide a theory of the universe to replace the cosmologies of myth eventually led to practical scientific discoveries. And, by exploring man's processes of thought and argument, principles of description and logic were defined that were both an expression of and an influence upon the practice of political and personal morality. The search for principles was in fact carried so far that, too often, theory was preferred to the more mundane process of observation, so that progress in the natural sciences and medicine was slow compared to achievements in other fields.

That man is the measure of all things and mankind his special study may appear today the most remarkable tenets of ancient Greece, especially in light of the obsessional religions, oppressive governments, and conventionalized arts of other early societies. This is, on the whole, a just appreciation of the Greek achievement, but in the past it has led to an unreal, not to say sentimental, idealizing of Greek standards of virtue, wisdom, and beauty. More recent studies have dwelt more objectively on the Greeks as one society out of many in antiquity, whose debts to others were at times considerable and whose success must be measured against the not inconsiderable achievements of other peoples. But it remains true that no other ancient people exhibited such a wide range of genius or left such a vigorous legacy. The heroic quality of Greek history—the stand against Persia, the exploits of Alexander, expressed so proudly by their own historians and poets—is, if anything, enhanced by this new objectivity, by the understanding that the Greeks were subject to the same fears, hopes, motives of power, self-interest, love, duty, and cruelty that have beset men throughout history.

The geography of the Greek homeland. The Greek landscape is that of the miniaturist—with rapid but never drastic transitions from valleys to hills and, for such a small country, with so many geographically distinct enclaves, including its many islands, that it is totally unlike the broader tracts of plains and mountains that are characteristic of most other Mediterranean lands. This simple geographic factor explains the separate development of the Greek city-states, which were bound by a common language and interest rather than by any sense of belonging to a distinct geographical entity—until, that is, a sense of national unity was imposed on them from outside. Greece's long coastline and the difficulties of land communication meant that seafaring always played an important part in Greek life and politics and probably contributed to the Greeks' readiness to explore the farther Mediterranean shores. Unlike Asia Minor, or even Italy, Greece is a country of sea routes, not roads.

Influence of geography on Greek life

The provinces of Greece represent geographic, racial, or dialect unities rather than political unities. In the north, Macedonia, a peripheral Greek area until the age of Philip and Alexander, marked the end of the main routes from Europe. To the south, the country is split lengthwise by the Pindus Range, with Epirus and Aetolia to the west, facing the Adriatic, and the broad Thessalian plain sloping to the Aegean. The mountain range is difficult to cross at any time and almost impossible in winter. East central Greece, east of Mt. Parnassus and Delphi, forms a separate enclave with the plain of Boeotia, the hillier country of Attica with its thinner soil, and the long and rather bare island of Euboea flanking them. The Corinthian Gulf nips the peninsula, making the isthmus to the Peloponnese a natural focus of routes and power. The isthmus was dominated by Corinth, which built a channel (*diolkos*) for dragging ships overland. From the east, the seaward approaches were controlled by Athens and Aegina, while Megara, north of the isthmus, must have owed much of its wealth to its position beside the only north–south route overland. The mountains of south Greece, the Peloponnese, were equally effective as dividers of the land. There are only two plains of any importance, serving Argos and Sparta, and the coastline is generally inhospitable. The jutting masses of Euboea, Attica, and Argolis are continued southeast in a broad scatter of islands reaching across to the southwest corner of Asia Minor. These are the Cyclades, with Dorian Rhodes and the Dodecanese at the eastern end. South lies the great island of Crete, halfway from Athens to Africa, geographically the southern border of the Aegean world.

The west coast of Asia Minor is quite unlike Greece, with broad river valleys, massive mountain ranges running inland to the east, deep gulfs, and large offshore islands. The Greeks, whose communications ran more readily by sea than by land, naturally came to regard the coastal areas and islands of the east Aegean as theirs to exploit, which they did through much of the period, coming regularly into conflict with the mainland powers

(Lydia, Persia), which could not brook foreigners on their flanks.

The Greek climate

The climate of Greece is temperate, with long dry summers favouring the cultivation of vine and olive, and wet winters, especially in the western areas, but not generally very harsh. Barely one-fifth of the land area is arable, and, in parts of the country and islands, it can be seen that bad farming and erosion have reduced the areas once cultivated. Wheat grew well enough, but, even as early as the 8th century BC, not in sufficient quantity to satisfy the demands of growing urban populations. Specialized crops producing oil or wine, which could be traded for wheat, proved a more economic use of land for some states (Athens and Chios, for example). The hillslopes provided pasture for cattle, sheep, and goats, which played an important part in the provision of food and clothing. Greece was well wooded, but the demands of housebuilding and shipbuilding, as well as the attentions of the goats, considerably diminished the forested area even in antiquity. There is good fishing in the Aegean, especially of tunny, and the Greeks ate much fish, fresh and dried.

Climatic changes since antiquity have probably been slight, but there may have been a period of drought in the Early Iron Age which contributed to the severe depopulation of the country in about the 12th and 11th centuries BC. There has been a general rise in sea level since antiquity of at least two metres (6½ feet). Some parts of the coastline, notably in Crete and parts of the Peloponnese, have suffered local disturbance through earthquake. Most of Greece, especially southern Greece, is subject to these disturbances, which have at times proved disastrous. Mineral resources exploited in antiquity were few—some gold and silver and slight deposits of copper and iron. But there were rich sources of fine-quality stone for building and statuary and plentiful deposits of sedimentary clays for potting.

Sources for the study of Greek civilization. The prime source for all except the earliest periods is the record of ancient historians, describing events recent and remote, local and national. Early accounts of the foundation of cities and of political or military conflicts were in verse, in a literary form influenced by epic and barely recognizable as history in the modern sense. They were, nevertheless, of quite a different character from the bald chronicles of the East and Egypt, which were narrow in theme and almost never explored or explained motives. In the 6th century BC, the study of geography (Anaximander) and the writing of travel books (Hecataeus) recorded much information of historical value, but of this work only scraps have been preserved in later authors. This is true also of many works by important 5th- and 4th-century-BC writers. The handling of this patchwork of information requires great care and observation of the context and of the intention of any later author who recorded a quotation or excerpt from an earlier work.

The work of Herodotus

The first major work to have survived in nearly complete form is Herodotus' *History*, composed in the third quarter of the 5th century BC and devoted to a study of the background to and course of the wars against Persia. Herodotus travelled widely, collecting written and oral evidence. He displays a spirit of critical inquiry that has earned him the title of the "Father of History." He is ready with digressions that are often informative about earlier periods of Greek history; his interest ranges over problems of religion and culture; and his account of the customs and history of the neighbouring barbarian empires presents a more vivid picture of them than do their own records. His history can be seen as a logical development of the work of the earlier geographers and ethnographers.

Thucydides, writing at the end of the 5th century BC, told the story of the Peloponnesian War and incidentally of the years of the Athenian Empire after the Persians were repulsed from the Greek homeland. He succeeds in being a dispassionate observer of the greatest internal struggle that Greece had seen. It was a struggle that tested the quality and ethics of the political life of the city-states involved. Fortunately, Thucydides, too, was interested in the motives and interests that led to decisions by states or individuals, and he discusses these in passages of personal speculation or through the more dramatic device of speeches placed in the mouths of generals and statesmen.

Of the 4th century BC, there is Xenophon's *Hellenica*, which continues Thucydides' history. Xenophon did not have the flair of his predecessors, but he was a man of action, and his account of military operations carries the authority of experience. Ephorus' universal history is lost, but it was much used by the Sicilian Greek Diodorus Siculus, a writer of the 1st century BC, on whom it is necessary to rely for a consecutive account of events in the 5th and 4th centuries BC. Other historians of the Roman period draw on various earlier sources for Greek history. Arrian is a notable example; he gives an account of the reign of Alexander. And Plutarch's *Lives* and *Moralia* are rich in quotations of earlier sources, as well as being interesting historical commentaries in themselves. No less valuable than the historians as contemporary sources are the 4th-century-BC orators Isocrates, Demosthenes, and Aeschines, whose speeches have survived more for their literary than their historical content. Works on political theory by Plato, Aristotle, and others are also of use as sources. The remoter the writer from the events he describes, the greater the likelihood that the account has been distorted by invention or systematization either for political ends or for the sake of continuity. This most affects genealogies, and to these many events of early Greek history are tied.

A different type of literary source is offered by inscriptions on stone or clay. Some are private, recording epitaphs or dedications, but even these may allude to historical events of wider significance. Many are public documents: state decrees, treaties between states, honorific decrees, and tribute lists. They are valuable contemporary records and can usually be dated either by the context in which they were found or by the letter forms used. Most of importance are later than the 6th century BC in date.

Archaeological sources

Strictly archaeological evidence is being used increasingly by historians, especially for the early period of Greek history. Study of style and scientific aids have helped to determine a fairly reliable chronology for even the commonest objects recovered from an excavation, and this often makes possible a historical account of an area or town that written sources ignored. Archaeology can provide direct evidence for trade, but only for trade in commodities or containers that have survived. More important, it may indicate, in early periods of colonization or trading, which states of homeland Greece were operating overseas. While the ancient historian was inevitably attracted to the great names and events, archaeology can complete the picture of life in parts of Greece that escaped the limelight and can explore problems of society and technology that ancient writers generally ignored. It is a major source for the history of myth and religion and the only source for the history of art.

This article is divided into the following sections:

I. The Early Archaic and Archaic periods
  The Early Archaic period
  The Greek colonizing movement, *c.* 750–*c.* 500 BC
  Archaic Greek culture
II. The Greek city-state
  The growth of the *polis*
  Sparta
  Athens
  Other important *poleis*
  Relations among the Greek *poleis* to the end of the 6th century BC
III. The 5th century BC
  The period of the Greco-Persian Wars
  The Athenian Empire
  The Great Peloponnesian War (431–404 BC)
  The cultural life of Athens during the war
IV. The Greek world from 404 to 323 BC
  The Spartan hegemony and decline
  The rise of Macedonia
  The western Greeks
  Greek culture and society in the 4th century BC
  The conquests of Alexander the Great

V. The Hellenistic Age (323–30 BC): Greece in the Roman period
The establishment of the Hellenistic kingdoms (323–276 BC)
The Hellenistic monarchies in the 3rd century BC
Relations among the Hellenistic states, 275 to 27 BC
Hellenistic political, social, economic, and cultural institutions
Macedonia and Greece under Roman rule (to *c.* AD 395)

I. The Early Archaic and Archaic periods

In this article, the story of ancient Greek civilization begins with the Early Archaic period; that is, the period from the disintegration (*c.* 1200 BC) of the Bronze Age civilization in Greece to the beginning of the literary and cultural revival of the Archaic period proper (*c.* 750–*c.* 500 BC). For the history of the earlier Bronze Age cultures of Greece see AEGEAN CIVILIZATIONS.

THE EARLY ARCHAIC PERIOD

The Dorian Invasion and Proto-Geometric period, c. 1100–900 BC. The collapse and gradual destruction of the Bronze Age civilizations in Greece are subjects for archaeological investigation only, and it is natural to look first to archaeology for any account of the period that immediately followed. During this following period, the few remaining centres of the earlier Mycenaean culture were abandoned or destroyed. These included not only settlements at the older capitals, like Mycenae, but shorter lived refugee settlements in west Greece (Achaea, Cephallenia), east Attica (Perati); settlements on eastern Aegean islands (Emporio on Chios); and settlements in the islands that had hitherto escaped the worst effects of the invasions.

The "invasions" may be thought of as having occurred in waves for over a century. It is possible that many of the effects attributed to them were in fact the result of internal dissensions. Observing the areas that seem to have been least disturbed by the earliest attacks, and the directions in which refugees moved, it may be judged that the main course of invasion passed through Greece from the northwest. The identity of the earlier invaders is not abundantly evident in the archaeological record, which probably means that the attacks were carried out by raiders who may have quickly passed on or turned back with their prizes. Some objects of new types, or of types only familiar on the borders of the Mycenaean world, are introduced: an efficient bronze sword, violin-bow fibulae (clasps), some distinctive armour, knives, and handmade pottery. Prototypes for these are readily found in northwest Greece and the Balkans beyond.

The overall effect of the invasions was a depression of all cultural standards and considerable depopulation, except in Athens, parts of Thessaly and the Argolid, and in the refugee areas. It has been suggested that a contributory or even dominant factor was a climatic change that brought about conditions of drought and famine. The evidence for or against this hypothesis is difficult to judge, but the sudden upheaval, after so many years of prosperity and heavy settlement, is hard to attribute solely to the hand of man.

It is concerning the final wave of invasions or destruc-

From W. Shepherd, *Historical Atlas;* Barnes and Noble, Inc.



Ancient Greece.

tions in the early and middle 11th century that the archaeological record is somewhat more informative. It reveals some continuity of settlement with a new population or, at least, a somewhat changed way of life. This is most easily observed in Attica and the Argolid.

The new culture is distinguished from what had gone before in a number of ways. The Mycenaeans buried their dead in family graves. The new cemeteries are of individual burials in slab-lined cists. These are not altogether unknown in earlier Greece; they were normal in the Middle Bronze Age, and, toward the end of the Bronze Age, cist cemeteries were made in north Greece (Epirus and Thessaly). As an alternative burial practice, cremation was introduced and became dominant in the following period in Attica and Crete but remained rare in the Peloponnese and is only occasionally found elsewhere. Cremation was very rare in Mycenaean Greece but was practiced during the Mycenaean period in the central Balkans and was regular farther north. Straight dress pins make their appearance, and iron for weapons. These were to be characteristic of the new Proto-Geometric period, whose other characteristics are examined below. It is impossible not to associate the invasions of this period and the resettlement or continued settlement of the surviving Mycenaean sites with later Greek tradition about the Dorians. The Dorians were Greeks from northwest and north central Greece, and there is nothing in the archaeological record to upset seriously the theory that they were the last invaders and settlers of the Mycenaean Greek world. They effected a shift of emphasis rather than a revolution in the Greek way of life. Tradition told of the "Return of the Heraclidae," suggesting a number of earlier or repeated raids from the north. There were moves from Thessaly into central Greece and Boeotia and moves from Epirus and the northwest into the Peloponnese. Not all the immigrants were Dorian, but it was the Dorian invasion of the Peloponnese that overthrew the remaining Mycenaean centres and led to the strongest new settlements in Laconia, the Argolid, and Corinthia. Athens was marginally affected, perhaps mainly by refugee populations. A Mycenaean "pocket" remained in Arcadia, using a dialect related to that of the Mycenaean refugees in Cyprus. Dorians settled also in the southern Aegean islands, in Rhodes, and especially in Crete, where again there were pockets of Bronze Age survival—some even non-Greek Mycenaean (the Eteocretans). It is possible that less of Greece was totally depopulated during the invasions than has been thought. It may well be that the interaction between the new and old Greeks in what had been centres of Mycenaean power proved responsible for developments in religion and myth about which it is now only possible to speculate, since no contemporary written records survive. In Greece itself, the art of writing had been lost with the palace bureaucracies that disappeared as a consequence of the invasion. One of the prime achievements of the new Proto-Geometric period of which there is material evidence was in pottery and pottery decoration—from this the period gets its name. The notable refinement in shapes and decoration exercised by the Proto-Geometric potters represents not a break with earlier tradition but a reworking of old forms and patterns with a new precision and eye for proportion and design that the Mycenaean Greeks had failed to achieve in their adaptations of Minoan art. This is the beginning of Greek Classical art. It seems most likely that Athens saw the inception of the new style, and it was certainly in Athens that it was most successfully developed. And from Athens other areas of Greece drew inspiration. Athens was never Dorian, but it shared with the Argolid the first manifestations of the new style. Skeletal material from Athenian graves reveals a marked intrusion of northern types, and so the responsibility of the new Greeks for the changes observed in Athens may well be real, and possibly not wholly indirect.

Identification of the invaders

Cemeteries are the main source of information for the Proto-Geometric period, and it can be seen from them that as the style develops, through the 10th century, there is a growth of population accompanied by growth of prosperity. This prosperity is indicated by the increasing number of metal objects deposited in the graves. Apart from east central Greece and Argolis there was a strong unity of development in this period in Thessaly, Euboea, and into the Cyclades. In these areas the Proto-Geometric style died hard. The Euboeans and islanders, in particular, seem already at this time to have been busy seafarers; and, from the site of Xeropolis (Levkandí) on Euboea, between Eretria and Chalcis, there is evidence for a variety of imported goods from Cyprus and the Near East. By comparison, Laconia and western Greece were backward, slow to repopulate and to trade. Crete is a special case since the island suffered less than most of Greece from the disasters attending the end of the Bronze Age. Some centres in the island seem to have developed with a certain vigour and degree of independence. There are clear signs in Crete, also, of contact with Cyprus and the Near East.

Apart from archaeological sites, pottery, a restricted range of metalwork, and cemeteries, little can be assessed of Proto-Geometric culture. Virtually nothing is known about temples or even houses. Yet it was during this period that the tenets of myth-history and religion found in an already developed form in the Homeric poems at the end of the following period were being established. Much of the ordinary way of life described in those poems was probably typical of the Proto-Geometric period. And, despite an overall unity of culture in Greek-speaking lands, it is possible to detect that differentiation of customs (mainly burial) and style that came to be most fully expressed in the stout independence of the city-states.

The Ionian migration

There is one other event, rather loosely described and dated by ancient writers, which archaeology suggests may be placed in this period—the "migrations" to the east coast of the Aegean. The area had already been much visited and partly settled in the Bronze Age, and offshore islands had received refugees from the Mycenaean world. But none of these communities survived into the Proto-Geometric period, and for most of the area there is reason to believe that there was a break in Greek occupation, although not so obliterating that later Greeks were unable to find their way again to the same sites. The migration was described by ancient sources as a single event involving the marshalling in Attica of refugees from Pylos, together with other Greeks from Boeotia and elsewhere, and their passage east to found Miletus and Ephesus, and thence the other major cities of Ionia. There may be archaeological evidence for such refugees in Late Helladic Attica, and there is certainly archaeological evidence for the settlements in Ionia in the 11th and 10th centuries. The style of pottery found in them indicates some relationship with Attica. Miletus was founded by the start of the Proto-Geometric period, and this seems to support the literary tradition. Finds in most of the other major cities indicate that all were probably established before the end of the 10th century, but the diversity of the records about their composition and founders and the inadequate evidence of excavation make it difficult to decide how much was effected by fresh immigration from the homeland and how much by local expansion.

By about 900 BC Athens had emerged as the cultural centre of the Greek world; the Euboeans and islanders had begun their eastern voyages, which were to lead to new trading ventures and transform the prosperity of the Greek world; and the Aegean had become for the first time a truly Greek sea.

The world of Homer: the Geometric period, c. 900–c. 700 BC. The Geometric period is named after the style of pottery decoration that was developed in Athens out of the more austere Proto-Geometric idiom. It was wholly abstract at first but eventually admitted figure decoration. The closing phases of the period saw the introduction of elements derived from the arts of the Near East, a phenomenon that resulted from new Greek trading enterprise in the east Mediterranean. The desire and need for this trade reflect the more stable and prosperous conditions of life in Greece, which gave rise to a

notable increase in population and the need for colonizing ventures. The same period sees the introduction of writing and the possibility of written records. Greece moves into the light of history.

The material culture of the new period is best judged in Athens, where the cemeteries encircling the city have proved most productive, and in the countryside of Attica. The method of burial becomes more varied, ordinary inhumation becoming as common as the cremation that had been dominant in the preceding period. Grave offerings, with weapons and metalwork, give some indication of the new standards of life, while some burials in Athens (the so-called Dipylon graves) are marked by monumental clay vases decorated with scenes of the laying out of the body and its removal to burial, implying funeral rites and celebrations of some complexity and expense. Iron appears more often and is used for swords and spearheads; bronze is used for dress pins, ornaments, and implements. There is a considerable increase in the production of minor works in bronze and clay for dedication in sanctuaries, mainly in the form of men and animals. The most impressive bronzes are the great tripod caldrons, best known at Olympia but also in Athens and elsewhere in Greece. Precious metal was still comparatively rare, but immigrant eastern goldsmiths were at work in Attica in the early 8th century, and there was some local production of goldwork in the Geometric style. Some eastern metalwork and jewelry had arrived by the 9th century and had been deposited in graves. Another luxury craft, forgotten since the Bronze Age, was the cutting of seals in ivory and stone, indicating some sophistication in the safeguarding of property and identity of ownership.

This Geometric culture of Athens is repeated all over Greece, although nowhere else with comparable complexity and wealth, except possibly in Crete. Athens still seemed to give the lead in pottery decoration, but regional schools developed, especially in the 8th century. The new wealth and growing population of the Greek states were soon to lead them to test their strength and independence against each other in fields other than that of craftsmanship.

The Proto-Geometric unit formed by Thessaly, Euboea, and the Cyclades remained a strong entity, with Euboea the dominant partner. At Levkandí, the site between Eretria and Chalcis in the Lelantine Plain, there is clear evidence for metalworking and signs of trade with the Near East. The town was destroyed before 700 BC, while the site of classical Eretria was first settled only in the early 8th century. These events are in some way to be related to the Lelantine War, a conflict between Eretria and Chalcis, which, according to ancient authors, split Greece into two camps. The Thessalians were, not unexpectedly, directly involved. The affiliation of Levkandí and the outcome of the conflict, which may have been protracted, are not clear, but the immediate result was considerable restriction in the activity of the Euboean states at home and overseas. In the islands there were several important Geometric towns, notably on the sacred island of Delos.

The Lelantine War

In Boeotia, Thebes seems the strongest site and most active culturally, although dependent on Athenian styles. Delphi was prosperous by the 8th century and received many Corinthian goods.

In the Peloponnese, archaeological evidence indicates the rapid growth in wealth and size of Corinth and a considerable concentration of prosperous sites in Argolis —at Argos, Mycenae, Tiryns, and Nauplia especially. While there is not the volume of imports and precious metals found in Attica, it is in Argos that is found the earliest suit of all-metal armour of the type later worn by hoplite soldiers, as well as offerings of spits (*obeloi*) in sets of six. Corinth's prosperity began in the 8th century, especially after the middle years, when its pottery began to appear in quantity on a great many Greek sites at home and overseas.

Crete presents a checkered picture, with several important sites in the centre (notably Knossos) and the east. Relations with Cyprus remained strong, and by the end of the 9th century a group of eastern goldsmiths had settled in Knossos. They worked there, in gold, bronze, and stone, for some three generations. Crete, like Thera —another Dorian island that prospered at this time— generally preferred cremation for its dead. Some tombs, though used for the new rite, retain the form of the old Bronze Age *tholoi*: the Minoan heritage is imprinted in a number of ways on Archaic Cretan culture.

In east Greece (Asia Minor and the eastern Aegean islands), several towns on Dorian Rhodes were flourishing and developing a regular trade with Cyprus by the 9th century. It is not until the 8th century, however, that the major sites in Ionia show marked signs of growth, notably at Smyrna and on Samos, where the first temples of the Heraeum were built. The archaeological record from other cities and islands is sparse. Farther north, the coastline of Aeolis as far as Troy was settled. Finds on sites as far south as Chios and Samos show evidence for increasing awareness of the hinterland powers of Phrygia and Lydia, while farther south Greek fashions were penetrating Caria.

The material pattern of life in Geometric Greece is characterized by growth in population and wealth, but a special quality is also increasingly apparent, and this must be summarized and explained. Objects from Cyprus, Syria, and Phoenicia (and to a limited degree Egypt) arrived in the Greek world in increasing numbers through these two centuries. The presence of immigrant craftsmen can be detected in Attica and Crete; there may have been ivory workshops in the Peloponnese, Attica, and east Greece; and, by the second half of the 8th century, most Greek crafts were beginning to betray the imminence of the influence from the East that became more fully apparent in the 7th century. This appears most clearly in Corinth, Crete, Rhodes, and Attica. The goods and craftsmen followed routes that may not at first have carried much else by way of serious trade, but the expanding Greek states looked overseas for new resources, especially for metals—copper, tin, iron—that were in short supply at home and that were especially required for weapons. Tradition has it that the Phoenicians conducted much of the early trade with Greece, but physical evidence for this is hard to find, and the first clear sign of organized trade is a Greek enterprise. This involved the establishment of a trading post at the mouth of the River Orontes in north Syria at a site known as al-Mīnā. There is no literary record of the venture, but it, or a similar operation, provided Greece with its first opportunity to keep written records, since it was through intercourse with the Aramaic-writing peoples on this coast that the Greeks learned and adapted for their own use alphabetic writing. This must have happened by the mid-8th century, and if Olympia had accurate written records going back to 776 BC, then it may be surmised that writing reached the Peloponnese about that time.

Eastern influences on the Greeks

The physical world of the Homeric poems is dominantly that of Geometric Greece, overlaid as it is with the myth-history and geography of an earlier age and with stock epithets and descriptions perhaps out-of-date by the 8th century. The picture is a convincing one of independent kingdoms centred on single cities with dependent villages and linked only by bonds of kinship or mutual convenience. Rule is generally autocratic and fighting an affair of the nobility, the citizen mass not yet being a decisive force. Hesiod, writing around 700 BC, paints a more sombre but moralizing picture of the hardships of a farmer's life in Boeotia.

Homer and Hesiod are poor guides to religious life in Geometric Greece. Homer's gods are already acquiring set attributes and an Olympian family life, imposed in the interest of the story. The process may have been going on for some time, but this is bards' work and for long afterwards bore little relationship to the Greeks' view of the town's patron deity and that deity's functions and dues. Hesiod's cosmogony is, if anything, even more artificial and owes much to Eastern cosmic myths. His systems and tales also find little echo in later Greek religious practice, but in parts of his work natural justice and the will of the Father of the Gods correspond

Religion in the Geometric period

well enough to permit recognition of an attitude to the gods based on humane respect rather than fear. Archaeology can offer evidence of only the trappings of religious life. The earliest cult places are altars for sacrifice, which also serve as the focus for offerings. The provision of a house (*oikos*) to accommodate the symbol or cult image of the deity comes late, and the few Geometric examples of such primitive temples all resemble contemporary domestic architecture. The images themselves may well have been aniconic (symbolic rather than representational). The offerings are bronze or clay substitutes for the offerer himself, representations of the deity, or intrinsically valuable objects such as the metal caldrons that in the Homeric poems are offered as prizes. In the grave a warrior may be buried with his weapons, sometimes rendered useless (swords bent or folded), but most offerings are of pottery and seem to indicate provision for a meal, no doubt for the journey to the other world. There is no clear evidence for continued rites at the graves of the recent dead, but in the 8th century there is evidence of attention being paid to some Mycenaean tombs and of the introduction of some hero cults which may have been inspired by the Homeric poems, as at the Agamemnoneion of Mycenae.

Certain other aspects of Greece's remarkable Bronze Age past must have affected its new and rapidly developing Iron Age civilization. The walls of Mycenaean citadels were viewed as the work of giants, the Cyclopes. These physical monuments, together with the poems, were clear evidence for a golden age, long past and replaced by one of iron. Geometric and later artists were very rarely moved to imitate these works. The degree of continuity from the Bronze Age is more difficult to assess but has been enthusiastically championed. The palatial arts were lost and had to be relearned from the East. Old sites reoccupied had led scholars to believe in the continuous use of cult places and sanctuaries, but they are difficult to identify in Bronze Age terms.

Although their new literacy may have awakened an intelligent interest in their past, the strongest motivation for progress in the Geometric period came from the Greeks' own efforts to improve their lot at home and from their enterprise overseas, which brought them both valuable new resources and invigorating new ideas.

THE GREEK COLONIZING MOVEMENT, C. 750–C. 500 BC

The Mediterranean world and Greek needs. The geography of Greece determined the Greeks' interest in seafaring, and, when their homeland could no longer support their growing population, it was their seafarers who offered the solution. Early sailing was a hazardous business, generally restricted to a few months in the year and to routes which, so far as possible, never took the mariner too far from sight of land. Not until the 6th century were maps being drawn, and, even then, they were of little practical value, compared to experience and the reports of others. The need now was for land and new resources. What could the Mediterranean seaboard offer? First, a climate that was at no point so markedly unlike that of Greece that Greeks could not readily adapt themselves to the new conditions of life and agriculture, and, second, sea and river routes to remote peoples and resources. The sources of raw materials were discovered early and exploited fairly easily by trade, even if in the teeth of foreign, usually Phoenician, competition. The western Mediterranean gave access to the tin and silver of Spain, France, and even Britain; central and north Italy could offer copper and the iron of Elba. The richer mineral resources of Europe were reached only with more difficulty. On the shores of the Black Sea there was gold (in Colchis), and there was iron in north Asia Minor. The copper island, Cyprus, pointed the way to areas in which metallurgy had long flourished and from which many even more distant sources had been exploited. But food was even more important than metal, and this involved seeking new land for settlers. This, of course, depended upon finding unoccupied but cultivable land or upon the goodwill of new neighbours.

Expansion and trade

The Mediterranean world offered several such opportunities for expansion and trade. The northern shore of the Aegean flanked the territories of the minor kingdoms within Macedonia and Thrace. The Macedonians were Greek in culture; but neither they nor the Thracians were seafarers, and the coast of Macedonia, especially the peninsula of Chalcidice, proved an important colonial area for the earliest of the active colonizing cities in Euboea. Farther east, new towns were founded from the islands and east Greece, and the gold of Mt. Pangaeum was soon exploited. On the east shores of the Aegean the hinterland powers also had little interest in the coast, and the already established Aeolic and Ionian cities soon expanded along the coastal strip up to the Dardanelles. The passage up into the Black Sea, against wind and current, was difficult but not impossible. The Thracians to the west seem to have welcomed the Greeks. In the north the newly arrived nomadic Royal Scythians provided a ready market rather than competition, and again the Greeks were able to benefit from the fact that they were more interested in the sea and coastal areas than the natives were. They soon learned to exploit the riches of the south Russian cornfields and fisheries. The less hospitable eastern and southern shores of the Black Sea offered fewer but important places for foundations that could tap important mineral resources.

In Cyprus, the Near East, and Egypt the situation was very different. All the local states were well established, and all had an interest in seagoing. The Egyptians also had a deep-seated suspicion of all foreigners. Their trading routes inland were well trodden, and there was no foothold there for a colony. Libya was a different matter, with a broad green plateau and disorganized native tribes which appeared cooperative. Farther west, the Phoenicians had forestalled the Greeks, and Carthage was founded while the Greeks were still exploring Italy and Sicily.

The Greeks were slow to explore routes up the Adriatic and crossed straight to south Italy and up to the promising markets of Etruria. There, as in Egypt, the opportunities were for trade. The natives of south Italy and Sicily were landlubbers, their natural ports and coastal plains mainly empty, it seems. There the Euboeans, followed by Corinth and other Peloponnesian states, were able to establish important agricultural communities. East Greeks later prospected farther afield, to France and Spain.

There were solutions to the land problem other than the sending out of colonies; but they were rarely resorted to. Sparta annexed the land of its neighbour, Messenia, acquiring a large slave population and its attendant problems. A dissident group in Sparta threatened state unity and was dispatched to found Tarentum in south Italy. Athens had been an active seagoer earlier in the 8th century, but it did not join the first colonizing movement and later avoided the necessity by specializing its economy in export materials—oil, wine, and silver. The Euboean cities, Eretria and Chalcis, were the earliest and most energetic traders and also the earliest colonizers, since their home island offered little scope for their new wealth.

Relations between colony and mother city

The first colonizing parties were probably very small, and families from neighbouring states seem often to have been free to join. This contributed to mixed feelings of loyalty, and, given the forced circumstances under which some colonies were dispatched, it is not surprising that relations with mother cities were sometimes less than cordial. When the colonizing states were seagoers, like the Euboeans and Milesians, the routes and possible destinations were already known to their merchants. The Therans turned to a Cretan who knew it was safe to bear south for the Libyan coast. Other states may have sought information from the Delphic oracle: the latter's newfound wealth and reputation may have been in some part due to the role it played or claimed as promoter of the siting, cults, and laws of new foundations.

The immediate needs of any new city may have been quite basic, but its identity as a full Greek *polis*, or city-state, on the model of its mother city, was immediately

affirmed by the establishment of the appropriate cult places and the demarcation of an agora area for political and commercial assembly. The new cities had not the architecture of an old city to impede or limit their planning, and their political and legal institutions were perhaps less trammelled by local, family, or territorial traditions. But it did not take long, with the growth of wealth and strength, for their political and social life to acquire all the complexity of the homeland.

The trading posts. There had been refugee Mycenaean Greek settlements in the island of Cyprus, and some surviving Greek element in the culture and population of the island may account for the fact that it was only with Cyprus that some sort of trade relations, even if casual, were maintained through the "dark ages" of Proto-Geometric and Geometric Greece. Archaeology shows that this was true especially of Crete, Euboea, and eventually Athens. Cyprus itself, from its geographical position, was a cultural extension of the Levant, and, by the 10th century at least, new Phoenician cities had been founded in the island. Before the end of the 9th century it seems that Cypriots had led Euboean traders to the eastern mainland and, with them, established a trading port at the mouth of the River Orontes, at a site known as al-Mīnā, excavated before World War II but not recorded by any ancient author. Euboean and island interest was dominant there until about 700 BC, and this site and port, with immediate access to the Syrian hinterland and Mesopotamia, may have been the major source for the flow of Eastern materials into the Greek world of the 8th century. After 700 the Cypriot and Euboean interest gave place to one that was mainly east Greek but shared by a mainland Greek state, perhaps Corinth or Aegina—at least a state using Corinthian pottery. By this time there may have been Greek trader elements in other eastern cities, such as Tarsus in Cilicia and farther south on the Syro-Phoenician coast (at Tall Sūkās, for example). There would have been no serious rival to this route, which brought prosperity to another Greek staging point for Eastern objects destined for the Greek world—Rhodes. In Asia Minor, the roles of Phrygia and Lydia as purveyors of Eastern goods and ideas seem to have been strictly limited to the east Greek world, and the direction of the flow was soon turned.

In Cyprus and north Syria there was no question of establishing colonies, although in time Cyprus was to become thoroughly Hellenized and isolated from its Eastern associations. Greek trading in Egypt gave rise to a still different type of settlement. In the mid-7th century the Egyptian king Psamtik I had employed east Greek mercenary soldiers, some of whom were allowed to settle in the country, and it is evident from excavations that by about 620 BC a Greek trading town had been established at Naukratis on the western branch of the Nile Delta. Herodotus describes a reorganization of the town under King Ahmose II in the mid-6th century that confirms what the finds indicate: that it was in the hands of various east Greek states and of Aegina. This was virtually the same consortium that had been operating at al-Mīnā since about 700 BC. In Egypt their operations were limited to trade through Naukratis, carrying in wine, oil, and silver, and carrying off papyrus, manufactured goods, and, eventually, corn. The community was better established than at al-Mīnā, since the Egyptians allowed a degree of local government by overseers appointed by each trading state and a number of temples, including a joint Hellenium. The latter may have been involved in collecting the taxes that had to be paid over to Egyptian temple treasuries.

The Greek trading settlement in Egypt

Precolonial trade in other areas of the Mediterranean world was far less firmly established and was conducted, it seems, by regular visit rather than through trade settlements. Evidence for it, therefore, is in the form of Greek objects in native towns. These appear sporadically in south Russia, up the Dnepr and the Don, and more often in Etruria. Some Greek place-names in the Carthage area suggest that Euboeans may have explored this coast, too, but Phoenician settlement there by about 750 BC forestalled possible development of trade or settlement.

Greek colonies in the west. *South Italy and Sicily.* Finds in Etruria show that Euboean merchants had been visiting its shores early in the 8th century BC. Before the middle of the century, a settlement had been established on the island of Ischia, offshore of Naples, at a site called Pithecussae. It was easily defensible and offered a good harbour at a convenient point on the approach to the Etruscan cities in the north. Its early years were its most prosperous, it seems, and the Greeks on this agriculturally unpromising volcanic island must have relied heavily on trade. There are early traces of ironworking, with material no doubt acquired from Elba. By the middle of the century, a second colony was established, on the mainland at Cumae, mainly by Chalcidians from Euboea. This provided better opportunities for livelihood off the fertile coastal strip, and it naturally came to eclipse Pithecussae in importance. These, the first of the Greek colonies, were also the most remote on this coast and the nearest to Etruria. The approach to them from Greece lay through the Strait of Messina, and the Chalcidians soon took note of the Sicilian coast—both for staging points and for its resources. The Chalcidians founded Naxos in Sicily in 734 BC. It seems that the Greeks had to displace a native settlement, and the same was true of the Chalcidians' next foundation in Sicily, at Leontini, which lay to the south of Naxos and a little inland, commanding a rich plain. Lower down the valley, on the sea, they founded Catana in 729 BC. This had a good harbour and plain. Thus it can be seen that, as in Italy, the first settlements were determined as much as anything by marine or mercantile needs and were soon supported by agricultural foundations.

The passage through the Strait of Messina had also to be safeguarded. Chalcidians, with a party from Cumae, founded Zancle (later called Messana, modern Messina) on the Sicilian side, in a position that was more strategic than comfortable and that was supported in 716 BC by the foundation of Mylae, at the other side of the northeast spur of Sicily, where there was a plain. On the Italian side, Rhegium was founded at about the same date by Chalcidians and a party invited from Messenia in the Peloponnese (noble refugees from Spartan expansion). The Euboeans had sought to safeguard their approach to the west by a colony on the western Greek island of Corcyra. In 733 BC the Eretrians were expelled from the island by Corinth, whose own settlers were soon at loggerheads with their mother city. They did, however, eventually press on up the Adriatic, founding Epidamnus in Illyria in 627 BC. Corinth's eyes were not on the north, however, but on the west and the trail blazed by the Euboeans, and its further colonizing along the western coast of Greece was mainly in the interest of protecting the approaches to the Corinthian Gulf.

The founding of Syracuse

According to ancient sources, it was only a year after the first Euboean foundation in Sicily that Corinthian settlers occupied the best port site on the east coast, and one which was to become one of the strongest cities in the whole Greek world—Syracuse (*c.* 733 BC). This was, in fact, an island site, soon linked to the mainland by a causeway; it commanded a broad coastal plain that seems to have been virtually unoccupied. Settlers from Megara had a harder time, spending a while in Chalcidian Leontini, then moving to a bare promontory at Thapsus before being invited by the native Siculi (Sicels) to what was to prove a valuable coastal site—Megara Hyblaea—in 728 BC. All these foundations were in the east of the island, and it was another generation before some Dorians, from Rhodes and Crete, ventured a little way along the south coast, to Gela, in 688 BC. Farther west lay the Elymi (Elymians), but active Phoenician interest in the west of the island and the establishment of a Phoenician town at Motya some time before 700 BC discouraged the Greeks from pressing far in this direction, although there was at this time no direct confrontation between the two great colonizing and trading nations.

The Euboeans established colonies in Italy that were as much strategic as agricultural. They were soon followed by Peloponnesians who had an eye rather to the rich and unoccupied coastal plains that the Euboeans sailed past.

Most came from Achaean cities. Sybaris, whose prosperity was short-lived but became legendary, was founded in about 720 BC, followed by colonies at Metapontum, Croton, and eventually Caulonia. These enjoyed the best of the land on the southeast coast, but there were valley routes northwest to the Tyrrhenian Sea, bypassing the Strait of Messina, and there was an early flow of Greek goods through the native towns of the hinterland. The other important Achaean colony, Poseidonia (Paestum), lay at the end of one of these routes and was established by about 700 BC.

The Achaeans were not alone in south Italy. In about 706 BC the Spartans settled the fine port at Tarentum, beside a native town that had been in touch with Greeks long before. Locri was settled from central Greece in about 673 BC. The first Ionian colony was at Siris, settled by Colophonians who had been displaced from their homes by Lydians—a circumstance that was to set other Ionians on the move west in the next century. Within less than a century, and mainly in its first 50 years, Greek states had found for themselves new and prosperous homes on the coasts of south Italy and Sicily. They had established generally good relations with their neighbours, were well placed to dominate trade with Etruria, and were already courting the jealousy of the Phoenicians and Carthage.

South Italy and Sicily: expansion and consolidation. The first colonies had all been on or very close to the sea. The period of expansion and consolidation, after the mid-7th century, saw more inland foundations, bringing the Greeks into closer contact with the native peoples. New cities toward western Sicily began to challenge the Phoenician hold on that part of the island. Although some of the new cities were manned by newcomers from Greece, much was undertaken by the colonies themselves, which sought to confirm and safeguard their positions. Syracuse established its hold on southeastern Sicily by founding Acrae (663 BC); Casmenae, yet farther inland (643 BC); and Camarina (598 BC). Smaller settlements, as at Helorum on the coast south of Syracuse, confirmed its domination of the corn lands and, by enslaving the Siculi, Syracuse had, by the early 6th century, progressed from being an outpost on a foreign shore to command of an area and cities rivalling any controlled by a homeland city. Camarina alone claimed a measure of independence: it quarrelled with Syracuse and was defeated and laid waste in 552/551 BC.

Expansion of Syracuse

Megara Hyblaea, hemmed in by the Chalcidian cities and Syracuse, looked to the opposite end of the island and founded Selinus soon after the mid-7th century in territory close to the Phoenician town of Motya. This was to prove another of the most prosperous western cities, long to outlive its mother city. At about the same time (648 BC) Chalcidians, joined by a group from Syracuse, also struck westward to found Himera, lying about as close to Phoenician Panormus (Pallermo) as Selinus did to Motya.

The last of the important Sicilian foundations was made by Gela at Acragas (Roman Agrigentum) along the coast to the west, in 580 BC. Eastward expansion was contained by Syracuse, and the Geloan cities were to be a serious challenge to Syracusan supremacy in Sicily.

In south Italy, the new foundations were mainly of the 6th century and were sponsored by the existing colonies establishing routes across the peninsula or "filling in" the coastline. The only newcomers were from east Greece. They had considerable trouble establishing themselves and provoked the first serious troubles with the Phoenicians. A party from Rhodes and Cnidus tried to settle too close to the Phoenicians at Lilybaeum in west Sicily and were expelled soon after 580 BC following a dispute involving the Phoenicians, Selinus, and Segesta—a native Elymian city that was strongly Hellenized and constantly at odds with Selinus. The unlucky colonists moved on to the Lipari Islands, where they established some form of communal ownership of land and annoyed the Etruscans by attacks on their shipping.

In the face of the Persian invasion of western Asia Minor in the mid-6th century, several east Greek states considered the possibility of mass emigration to the west, one suggestion being that the Ionians should move to Sardinia. The Phocaeans lost their city to Persia and sailed west to join the colony they had founded at Alalia on Corsica in 560 BC. Their piratic behaviour produced an alliance against them between the Etruscans and the Carthaginians, and, although the Phocaeans defeated them at sea, they had to leave Alalia and seek a new home. They stayed a while at Rhegium and then founded Elea (Velia) on the coast just south of Poseidonia in about 535 BC.

Political life of the colonies

Internally, the Greek colonies developed much as their mother cities at home. There were rare examples of political innovations, as at Lipara. The progress was generally from aristocracy to rule by tyrants, but the landowners of Syracuse were long influential. Disputes between the colonies did not always follow "national" lines, and the division between Dorian and Ionian had little weight in the colonial period.

Relationships with the native population rapidly became a matter of master and slave in Sicily, where only the Elymians enjoyed some security under the protection of the Phoenicians in the west. The other native towns of Sicily became progressively more Hellenized. Many were taken over, so that by the end of the 6th century the large native population was no longer a significant force in the island. Much the same was true in the southernmost part of the mainland, but the native tribes of Apulia, Lucania, and Campania were made of stouter stuff, retaining an individuality of culture well into the Classical period.

Conservatism is perhaps a characteristic feature of colonial areas. It was certainly a feature of the culture of the western Greeks. Their new cities were rich and powerful, and they sought to demonstrate this wealth both by the splendour of their offerings and by athletic missions to the national sanctuaries and games of Greece itself, and by their building at home. At Olympia at least half the pavilion treasuries dedicated by individual states were paid for by westerners—Gela, Metapontum, Sybaris, Selinus, and Syracuse. A majority of the surviving temples of the Greek world constructed in this period stand in the west, and in Syracuse a start was made to emulate even the colossal temples of the Ionian cities. This was an isolated venture in the Ionic order: most western work is Doric, massive and conservative in proportions, but often admitting uncanonical decorative detail, "vulgar" by metropolitan standards. Lacking fine white marble, the statuary is in stuccoed limestone or clay and is decidedly provincial in appearance. Only in the minor arts of clay and bronze did some western studios make original contributions to the development of Greek art.

Spain, France, and north Italy. The Phoenicians had been trading with the south of Spain since the later 8th century, and by 700 BC they had established at least one trading post there. The Greeks also took an early interest in these western waters and for a while held their own against the Phoenicians, whose foundations in Africa, Spain, Sardinia, and western Sicily eventually led to their domination of this area. The attraction was the tin and silver obtainable in Spain and, via Spain, from more northerly sources. Greek goods continued to arrive in Spain through the 6th century, but many, like the earliest of them found in the Phoenician trading post at Almunecar, may have been carried by Phoenician ships.

An alternative route for the tin was overland, through France, down the valley of the Rhône to Marseille. Here the Phocaeans, who seem to have been the most active of the east Greek merchants in these seas, founded Massilia in about 600 BC. A vigorous trade into France and north Spain developed and was later reinforced by other colonial foundations from Massilia. By this route Greek works of art travelled deep into France and Germany, and a Celtic chieftain at Heuneburg near Munich seems to have employed a Greek to plan his fortification for him. But, by the end of the 6th century, the balance of power had shifted, and, as a result of the diminished trade along the Rhône, Massilia's days of high prosperity were over. Its fortunes are an example of how much and

how long some colonies relied on trade rather than on local resources for their riches and even livelihood.

Greek colonies in the north. *North Greece.* The Euboeans, who were the first in trade and colonizing in east and west, were also the first to exploit the northern shores of the Aegean, where their neighbours were the part-Greek Macedonians. The centres of activity were the lower valleys of the Axius and Strymon and the three-pronged peninsula between them that took its name, Chalcidice, from one of the colonizing cities. It seems that both Eretria and Chalcis had been active in these regions in the 8th century: the Chalcidians at Torone and Dicaea; the Eretrians at Mende and then at Methone. After the Lelantine War Chalcidian interest was dominant in this area, and in the mid-7th century Chalcis sponsored colonies sent out from the island of Andros to sites on the east coast of Chalcidice—Acanthus, Argilus, Stagirus, and Sane. Eretrian control seems to have been confined to the westernmost promontory, Pallene. It is likely, however, that it was by agreement with Chalcis rather than Eretria that Corinth established Potidaea at the neck of the Pallene peninsula in about 600 BC.

Colonization of Chalcidice

The important Euboean foundations had probably been made by about 700 BC. In the 7th century, when the coast to the east was being colonized, it was other Greek states in the islands and east Greece that were active. An early foundation, *c.* 680 BC, was by Paros on the island of Thasos. Thasos soon set colonies on the mainland opposite, at Neapolis (Kaválla) and Oisyme. From Ionia the Chians came to colonize Maroneia and the Clazomenians to colonize Abdera, on the coast east of Thasos. There is indication of serious trouble with the natives—Thracians in this area, belligerent and reluctant to Hellenize. They expelled the Greeks from Abdera for a while, but, after the mid-6th century, it was settled again from Teos in Ionia, by Greeks escaping from the Persian threat. These settlements had a trade interest in the minerals and timber of the mainland, but they were also strong agricultural communities where the vine was grown with considerable success.

The Black Sea. Dates as early as the 8th century BC, which have been proposed for some of the Black Sea colonies, are supported neither by finds nor by considerations of geography, which show Greek colonial interest in the approaches to the Black Sea only at the end of the 7th century. But references and names in Greek poems show knowledge of the area by about 700, as in the Argonaut story. Miletus was the leading colonizer in this area, and most of its foundations were made in the later years of the 7th century, sometimes with the collaboration of other Ionian cities. The progress was not one of gradual exploration along the shores but of settlement, within a single generation, of key sites around the whole of the inland sea, especially at river mouths and points propitious for trade with the natives. The greatest Milesian city was at Olbia, at the mouth of the River Bug, but settlement there seems to have been preceded by a post on the island of Berezan, farther out in the Bug-Dnepr estuary. Olbia was the main point of contact with the Scythians of the south Russian steppes. Herodotus tells of the Scythian prince who was so corrupted by the Greek way of life that he was murdered by his followers.

South of the Danube delta the Milesians founded Istrus, and later Tomis (modern Constanţa, Romania) to the south, on what was to be a frontier of the Roman Empire. The other important area of Milesian colonization lay around the Kerch Straits, leading into the Sea of Azov east of the Crimean Peninsula. The oldest and strongest city in this area was at Panticapeaum (Kerch).

On the east coast, the wealth of the Caucasus was known to Greeks, and there was settlement there by Milesians at Phasis and elsewhere, about which there is still much to learn. The coastal route back to the Aegean lay along the north shore of Asia Minor, which is notoriously inhospitable to shipping. Three important staging points were settled as colonies along this coast, at Sinope, Amisus, and Trapezus (Trebizond). Apart from their value to transit shipping, Amisus, in particular, also gave access to the mines of Anatolia.

The Megarians, from mainland Greece, were other notable colonizers in the Black Sea region. They placed Mesembria to the north, and Heraclea just east of the straits, and in the 5th century crossed the sea to Chersonesus in the Crimea.

Economy of the Black Sea colonies

The prosperity of the Black Sea colonies depended on various factors. They generally had no lack of land for local needs and were able to introduce Mediterranean plants without too great difficulty—the climate of the Crimea can be remarkably like that of the Aegean for much of the year. They relied, too, on their trade in the products or raw materials of the hinterland. Corn from the Black Sea was already on its way to Greece before the Greco-Persian Wars, and the local production of works of art in precious metals in some of the Greek cities is a fair indication of how much of the same material might have been sent home. The waters of the Black Sea also yielded their riches. Kerch herring were prized, and quantities of pickled fish were sent back to Greece. In the Scythians they had to deal with a people whose culture was in many respects barely inferior to their own. Although passage of goods was free between them, the Greeks never really succeeded in assimilating the Scythian, in either their arts or manner of life, in the way in which they could the Eastern or the Egyptian. The frozen north remained for the Greeks the home of ogres and gifted savages.

Greek colonies in Africa. The easy route south from Crete to the Libyan coast may well have been travelled by Greeks in the Bronze Age. When, in the mid-7th century, the island of Thera was obliged to send away part of its population, it was from its Dorian neighbour Crete that a pilot was found to lead them south. They settled for two years on an offshore island and were visited there by Colaeus the Samian on his famous voyage west. They then moved for six years onto the mainland, to Aziris, before being led by the apparently friendly natives up on to the Libyan plateau where they found a site whose climate and setting closely resembled Greece, at Cyrene (about 632 BC). The attraction was agricultural, since the plateau was fertile and well watered. It was soon to become famous as a breeding ground for horses and as a nursery of the valuable silphium plant (now extinct), and in time it became an important source of corn for the Greek and Roman world. The foundation at Cyrene was soon followed by exploration along the coast. Apollonia was founded as the port of Cyrene, and to the west, the site later called Ptolemais, Taucheira, and Euesperides (modern Banghāzī). All of these, with the possible exception of the last, were founded before the end of the 7th century. The royal family at Cyrene was not always popular, and in the mid-6th century Barce was founded by an oligarchic faction, together with brothers of the king, farther west on the plateau.

This was the one colonial area that early suffered at the hands of the Persians. The cities submitted to Persian rule once Egypt had fallen to the new empire in 525 BC and had to suffer one armed Persian raid as far as Euesperides. But, even in this period, it is clear that Cyrene prospered, as its temples and tombs matched those of some of the richer western colonies. Dorian interest in the area was maintained not only by the founder islands, Thera, Crete, and Rhodes, but also from the Peloponnese, notably by Sparta.

In Egypt itself, the fortunes of the Greek trading town at Naukratis flourished through the 6th century and were barely dimmed by the arrival of the Persians. There were other small Greek communities in Egyptian towns that may have admitted "native quarters" or mercenary garrisons, but never anything like a colony. Out in the Libyan desert, at Siwa Oasis, the cult of Amon attracted Greek attention and soon became a recognized and respected oracle for all Greeks. The god was recognized as Zeus Ammon and figures on coins of Cyrene.

To the west, any Greek exploration had been cut short or wholly forestalled by the Phoenician settlement at Carthage. The only Greek attempt in this area came soon after 515 BC, when the Spartan Dorieus was led by Therans to make a settlement at Cinyps on the bare coast of

Tripolitania. Here he was pressed by both Phoenicians and natives and had to retire.

ARCHAIC GREEK CULTURE

Religion in Archaic Greece. The Archaic period offers a first opportunity to study more closely the religious life of the Greeks without paying undue attention to the rather artificial and literary pantheon of the Olympian family and dynasties of heroes. These, however, still provided the material for song and epic and were increasingly invoked or manipulated to serve political or personal ends. In the city-state can be distinguished two elements: the agora, which served as assembly and market place, and the temple area, where its god or gods were honoured. All the monumental architecture of Greece in this period was devoted to temples, altars, and other sanctuary buildings, and only toward the end, in Athens, were public administrative buildings treated in a comparable manner. The major deity of a city may have been determined by its Bronze Age history and the identification of a warrior god or fertility goddess as one of the Olympian pantheon; or by the promotion of a local deity or hero, often associated with a natural feature, to the same rank; or, in a colony, by the precept of the mother city or the Delphic oracle. The temple was at once a statement of confidence in the guardian power and a demonstration of wealth. Loyalty to the city god could be a stronger incentive to concerted action than loyalty to a family or tyrant. And, as the national sanctuaries won greater importance, either through their oracles or the staging of games, so loyalty to their deities—the Zeus of Olympia, the Apollo of Delphi—strengthened, together with the bond of language and their attitude to the "barbarian."

Dominance of religious architecture

Festivals were of different types and could serve different purposes. The state festivals, notably the Panathenaea in Athens, were occasions for the display of citizen solidarity and were readily exploited by tyrants to reinforce their own position. The many other festivals generally had a connection to agriculture—spring or harvest festivals, or celebrations associated with particular crops or activities (Dionysus for wine, Artemis for hunting)—or to the hero of a locality. The agricultural character of these is apparent in accounts and representations of the Archaic period, but their significance gradually became overlaid by other considerations when they were performed by city dwellers, whose life and interests became more and more remote from the countryside.

The conduct of festivals and of sacrifices by private individuals for favours received or hoped for was in the hands of priests whose office went by lot or family. Burial was a private affair but was regulated by the state. The Homeric magnificence of burials of the Geometric period was restricted by legislation (as by Solon in Athens) generally directed against excessive expense and public display. In the Archaic period, the dead were better served above ground, by monuments which reflected as much credit on the living, than below, and the few offerings laid with the body were but a token of a presumed brief need for sustenance on the way to the other world. But some religions or sects provided for man's fears for his afterlife. This was achieved through ritual purity in way of life or through initiation in a "mystery" religion such as that at Eleusis. The former method involved a complicated theory of evolution and matter and a mythology of the underworld with precepts for the journey there. Such was Orphism. The philosophy of Pythagoras, who left Samos to found a school in Croton in the later 6th century, was similar, teaching the migration of the soul into another body, which might be human or animal. Pythagoreanism codified many precepts that elsewhere would be classified as magic. Orphism, together with aspects of the Dionysiac worship (which was late to win importance in Greece), was also rooted in a religion of ecstatic "possession" by the god working through a priest or in the bodies of dancing worshippers. In the Archaic period there was, thus, little in the religious practices of the Greeks to mark them from other early societies. The difference lay in their literary treatment of stories about gods, from Homer on. Only with the Classical period did their interest in man and man's behaviour begin seriously to condition their views about their gods.

Orphism and Pythagoreanism

The early philosophers. The cosmogonies of Hesiod and other early poets described the beginning of the world in terms of the supernatural, but in poems that, at the same time, showed the artists' shrewd observation and appreciation of man as an object deserving study. The sometimes ribald treatment of divine behaviour probably accelerated the process of critical inquiry and scepticism about the validity of myth and the attempt to replace it with a view of the world and man's place in it that did not depend upon the antics of anthropomorphic deities. In the 6th century the centre of this new criticism lay in Ionia, and especially in Miletus, the city of explorers. The exercise of reason and the seeking of truth through argument were the two methods of inquiry, and they owed virtually nothing to Eastern science and very little to scientific observation. This is why the systems devised by the early philosophers seem so naïve and impractical to the modern mind, but what is most interesting is why they saw the need to devise them at all. They looked for the basic principles of matter and the world around them—in water (Thales); in the "boundless" separating into opposites (Anaximander); in air (Anaximenes); and fire (Heracleitus). Yet they were practical men, too, involved in the commerce and politics of their day, and it was probably due to their understanding of some of the mathematical and geometrical discoveries of the East and Egypt that certain of the more ambitious technical achievements of late Archaic Greece were made possible.

The Archaic style in the arts. The Archaic period saw the absorption by Greek artists of the artistic styles and techniques of the Near East and Egypt and their molding into a distinctive and original idiom in which lay the seeds of everything that was to be achieved in the Classical period.

The 8th century had seen foreign craftsmen and goods in Greece, but only toward the end of the century did these new arrivals begin to exert a profound effect on the work of Greek artists. Their metallurgy was always of a high order, and they refined techniques of casting and hammering bronze long before this was practiced in the East. New materials required new techniques, as in the case of ivory, and this, together with the filigree and granulation of the goldsmith, was a craft that had to be learned afresh, although it had flourished in Bronze Age Greece. The formal changes were more important. Geometric art could be subtle, but the more realistic conventions of Eastern art, involving the rendering of detail, offered far greater possibilities in the depiction of action. This was not, however, immediately applied to the needs of narrative or myth, and the hallmark of Eastern-influenced Greek work is the appearance of animal-frieze decoration. With the animals came a selection of monster motifs like the sphinx and griffin, to which the Greeks could add their own subhuman creations like the centaur and the satyr. Another new element was floral decoration, which gradually replaced the geometric and offered the possibility of considerable compositional invention—of which the Greeks were not slow to take advantage.

In the mid-7th century the Greek feeling for the monumental was stimulated by knowledge of Egyptian art. This gave rise to the first sculpture in hard stone (white marble) at life size or more, which successfully overcame the sterile conventions of work done in more easily managed materials and at a smaller, purely decorative scale. It also led to the creation of monumental stone architecture, with "orders" of moldings and columns based on timber prototypes (Doric) or on Oriental patterns (Ionic). This successful expression of the monumental probably helped the vase painter shake off the more restrictive conventions of decoration and indulge wholeheartedly in scenes of everyday life or myth, as on Athenian black-figure vases of the 6th century. A set iconography of scenes and figures was soon established, but never so rigidly as not to allow for the invention of a gifted artist or for originality in the execution of detail.

Developments in sculpture

Progress in the physical appearance of figures in Greek

art through this period was not rapid, despite the ease with which technical problems had been overcome. The figures were rendered with greater detail, and this was inevitably more realistic. Overriding considerations were proportion and pattern—the pattern of anatomical detail or the pattern of drapery pulled tight across limbs or hanging free from the body in a cascade of folds. In representing poses, the artist only very slowly broke with the uncompromising frontal or profile views and the abrupt juxtaposition of these views. But the change was not toward greater realism for its own sake. It was rather a refinement of the expression, by means of symbols for heroes, men, or monsters, of more and more complicated action and compositions. It was only just before the Greco-Persian Wars that the Greek artists began to see the merit of observing life and living forms and of adding a whole new dimension of expression to their storytelling.

Commerce. In Homeric Greece commerce was not an honorable profession, but, by the end of the Archaic period, the survival and wealth of most Greek city-states depended on trade, and commercial advantage could be a cause of war. The demand for raw materials, manufactured goods, and food in the expanding Greek economy was generally in this period met by private enterprise. A merchant embarked with his goods, paid his passage, and sold where he would; or, like Colaeus the Samian, it might be the captain of a ship who took what opportunities and profits were presented. Although it was to the state's advantage that there should be enough food or bronze for a hoplite army, it was some time before the state, as such, took a leading part in fostering or directing trade, although some tyrants did, it seems, as at Corinth. And, in early 6th-century Athens, Solon's reforms were in part directed at the Attic economy by encouraging an export market in oil. There was also an indirect interest in trade wherever taxes were levied on goods entering or leaving ports or passing through controlled territory. Slowly an ethic of commerce was evolved as trade became of more pressing importance to states. Piracy probably became less menacing, which is remarkable since there was no central government, let alone any policing force; it speaks well for the common sense of the Greeks and their readiness to submit dispute to the arbitration of fellow Greeks. It seems to have been remarkably easy for goods to pass between Greece and non-Greek enemy territories. There was a naïve simplicity about Archaic trade, but the days of tariff barriers and economic warfare were not far off.

Local or retail trade was, for most of the period, conducted by barter, the countryman exchanging his foodstuffs for manufactured goods or recognized precious objects, such as iron spits, which served as a form of primitive currency. But there was no question of the weighing of precious metals for retail payment, and for a long time Greeks were incapable of precision weighing. When dumps of precious metal were first stamped with a state or personal blazon (by about 600 BC, in Lydia, it seems) this guarantee was of the issuing authority rather than of weight. The use of coinage spread quickly through Greece. It was for local exchange, however, perhaps for payment of mercenaries by the state, and certainly for payment of fines or taxes. But there was seldom any provision of small change to serve a retail market and no passage of money from state to state to indicate its use for general trade. Only toward the end of the Archaic period and in places in which the precious metal was in good supply, as in Athens, was the production of coinage by the state apparently intended to serve "internationally" as recognized bullion. This is the beginning of a money economy, but coinage was not invented to serve this end, nor did it serve it for a century or more.

(Jo.Bo.)

II. The Greek city-state

THE GROWTH OF THE POLIS

The period of nearly three centuries in Greece after the great migrations of *c.* 1100–*c.*1000 BC is often called the Greek "dark ages" because little is known about it. Habitation was in villages and towns, which have left few traces except for pottery of Proto-Geometric and Geometric types. Rather before 700 BC a revival began, largely because of Oriental influence, and the term Archaic is given to the age *c.* 750–*c.* 500 BC to distinguish it from the Classical period that followed. Among the outstanding developments of the Archaic age was the growth of urban life and political institutions.

Character and early condition. The type of settlement emerging in many parts of Archaic Greece was the *polis* (plural *poleis*), or city-state. The *polis* was a sovereign political unit. Usually it had a fortified centre on a hill, where the inhabitants took refuge from attack; a residential and commercial town arose at the foot of the hill, and the town was surrounded by a dependent territory, which served for agriculture and pasture and, in the case of large *poleis*, included villages subordinate to the main city. Thus at Corinth the fortification was on the steep hill called Acrocorinth; the town of Corinth arose at the foot of the hill and commanded an extensive territory, which included subsidiary settlements such as Lechaeum, Sidus, and Crommyon. (It should be noted that this was not the only characteristic type of settlement; in much of European Greece, especially in the northwest, the unit was not the *polis* but the *ethnos* [plural *ethnē*], or nation. An *ethnos*, such as Phocis or Aetolia, held extensive land, which was divided among unfortified villages; the villages were linked in a loose federation, and no single centre predominated. But until the 3rd century BC, the political development of the *ethnē* was slow.)

In the early Archaic *polis*, public authority was weak. Society was agrarian, and each community had people of quite varied status—wealthy landowners, small farmers, landless labourers, artisans, and in some places serfs. The strong unit was the family; a large and powerful family attracted numerous dependents, giving them protection and economic help in return for their labour.

The divisions in society, consequently, were not horizontal but vertical: conflicts were not between classes but between groups. Each group was led by one or a few strong households, which had many dependents. (These groups have been aptly compared to pyramids because each had a broad base of humble men.) Eventually these relationships became formalized and hereditary; the consequent terms are known in the case of Athens. There *genos* (plural *genē*), or clan, was the word for a group of powerful families that led a body of dependents; each *genos* claimed descent from a common ancestor, who was usually mythical. The *genos*, together with the body of dependents attached to it, constituted a phratry, or *phratria*.

The citizens of each *polis* were also divided into hereditary units called *phylai* (singular *phylē*), a word often translated by the Roman term tribes. In Dorian *poleis* there were usually three tribes—the Hylleis, the Pamphyli, and the Dymanatae; among the Ionians the number of tribes was less uniform, but some of the same tribal names occur in several Ionian *poleis*. Evidently the tribes date back to the age of migration, before their members had settled in cities.

Political institutions

Three political institutions were commonly found in each early *polis*—a king, a council, and an assembly of adult male citizens. These institutions arose understandably from the migrations. Each migrating host required a war leader, whose office developed into hereditary kingship when the host found land and settled down. The war leader consulted the leaders of distinct contingents (his council) among his followers before taking action; and he announced his decisions to a gathering of all his able-bodied men, assembled to receive marching orders. (Such practices are reflected in the *Iliad* and the *Odyssey*.) But in sedentary conditions changes came about. The kingship was more vulnerable than the council: the king might be a minor, or there might be a disputed succession. In many cities additional offices, often elective and annual, were created alongside the kingship; even the kingship became annual and elective in some places, including Athens.

The council, though at first merely advisory, gained in

power from the decline of the kingship and became the primary scene of political rivalries. The assembly changed from a gathering that heard decisions to a gathering that reached decisions. This does not necessarily mean, as has often been supposed, that the common people, working through an assembly, asserted themselves against an aristocracy, working through a council. More likely, it means that when leaders of prominent families could not get their way in council, they took the issue to the assembly, calling out their followers to fight for and later to vote for them; and as this practice became habitual, the assembly became a second place for settling political rivalries.

Oriental influence. During the 8th century BC, Oriental influences began to reach the previously static societies of the Greek communities. A Semitic and probably Phoenician alphabet was borrowed and adapted for writing Greek. Perhaps in the same period Greeks learned the Oriental stories that appear in the *Theogony*, an account of the beginning of creation and the birth of the gods by the poet Hesiod (fl. *c.* 800 BC). Changes in pottery provide the clearest indication of foreign influences in Greek societies. Geometric pottery was superseded by new naturalistic styles in different cities at different times during the 8th and 7th centuries. Collectively the new styles are called "Orientalizing" styles, but they varied according to place of manufacture. The pottery made at Corinth was more widely exported than any other during the 7th century; the subdivisions of the proto-Corinthian and Corinthian styles span the period *c.* 725–*c.* 550. The adoption of the Orientalizing styles and the fact that these new styles travelled farther than had Geometric pottery indicate a growth of overseas trade, and this was likely to dislocate the traditional societies of the rudimentary Greek *poleis*. The colonization movement (*c.* 750–*c.* 500 BC) is evidence of profound social changes.

Hoplites. A change in military practice—the introduction of hoplite equipment and hoplite tactics—had effects that are difficult to estimate. In the "dark ages," fighting was mainly conducted by comparatively few warriors: each had a shield, a sword, and spears, usually for throwing, but lacked armour for the head and body; warriors fought as individual champions, not as an organized group. In the 6th and 5th centuries the characteristic force of the developed *poleis* was the phalanx of hoplites, or heavily armed infantrymen. The hoplite had a sword and a thrusting spear; and above all, he had plentiful protective armour, including a closed helmet, a bronze plate corslet, greaves, and a large round shield with armband and handgrip. In traditional practice the front row of hoplites tried to push the enemy off the field, and the succeeding rows of the phalanx added weight to the impact.

Until recently it was believed that many Greek states had adopted the hoplite phalanx within a few years of 700. But fresh study of vases that depict armed men has revealed a more complex development of the hoplite phalanx. The various items of hoplite equipment were adopted separately in the period 750–700, but at first the old manner of fighting by uncoordinated champions was not changed. The complete hoplite panoply is first attested on a vase *c.* 675, and the massed hoplite phalanx first appears *c.* 650. Moreover, the social effect of the change in tactics is disputable. A widely accepted theory says that in consequence of the change, the burden of fighting passed from relatively few aristocrats to a wider class of substantial peasant farmers; that the latter accordingly demanded a share in political power; and that revolutionary rulers, the tyrants of the 7th century, arose as their leaders. Against this view it has been argued that peasant farmers were more likely to be conservative than revolutionary, and they would dislike war and disturbance as threatening their land and taking them away from tillage; in Etruria and Rome hoplite tactics were adopted during a period of aristocratic ascendancy. Perhaps the social significance of the hoplite phalanx was something different: the new type of fighting was much more highly organized than was the old, and so hoplite discipline contributed to the growth in the authority of state.

Tyrannies in Archaic Greece. Some increase in public authority was brought about by the tyrants, who arose in many cities—especially of the Peloponnese—during the 7th and 6th centuries BC. The words tyrant and tyranny do not necessarily imply oppressive rule. Essentially a tyranny began when a man without any hereditary or

Adapted from W. Shepherd, *Historical Atlas;* Barnes and Noble, Inc.



Greek expansion *c.* 550 BC.

official claim to rule seized control of his city. Usually a tyrant held no office and did not alter the city's laws or institutions. He exercised an informal and extralegal preeminence; the nearest modern equivalent is the "boss" of a small town.

The rise of tyrannies

Because several tyrannies began in the Peloponnese within a limited period, scholars have often sought a common cause for them all. Some scholars have suggested commercial factors; others have argued that tyrants arose as champions of the pre-Dorian element in the population against Dorian aristocracies; others have attributed their rise to class conflicts consequent upon the introduction of hoplite tactics. But no one factor can be adequately attested for all the tyrannies of the Archaic Peloponnese. Local conditions varied, and the only note common to all tyrants was one of magnificence; they made a display of their power and wealth.

The word tyrant was not Greek in origin; it may have been a Lydian word for "king," and the Lydian kings of the Mermnad dynasty (*c.* 680–*c.* 544 BC) impressed the Greek imagination with their power and display. In Greek the word tyranny first occurs (7th century) in a quatrain of the poet Archilochus, who there associates it with his contemporary, Gyges, the founder of the Mermnad dynasty, and with ostentatious wealth.

Pheidon. Pheidon, a hereditary king of Argos, was a forerunner of the tyrants. Aristotle says, enigmatically, that he began as king but became a tyrant. He brought the whole of the Argolid (the area of Argolis) under the control of Argos; this was the basis for further expansion. Pheidon proceeded southward into the plain of Thyrea, which the Spartans too wished to seize; the Argives defeated them at the Battle of Hysiae in 669, and this ensured control of Thyrea to Argos for more than a century.

A tradition known to Herodotus said that Argos once ruled the whole southeastern coast of the Peloponnese as far as Cape Malea and even the island of Cythera. This tradition, if sound, must refer to the time of Pheidon, but it may be exaggerated. Probably in 668, although the date is disputed, Pheidon intervened at the Olympic festival; he excluded the regular Elean presidents of the competition and took the presidency himself. In this action he took advantage of tension within the territory of Elis. There the Eleans, who had arrived during the migrations, had ascendancy; but there was a rural and underprivileged class, the Pisatans, who were descended from those who had inhabited the land before the migrations. Pheidon promoted the aspirations of the Pisatans, and from his time intermittently until 580 they succeeded in seizing the presidency at the Olympic competition. Pheidon evidently achieved military ascendancy in much of the Peloponnese. He struck standard weights and measures, dedicating the models at the Argive Temple of Hera, and these standards were accepted permanently by many Peloponnesian states.

Cleisthenes of Sicyon. The fullest information on early tyranny concerns Corinth and its western neighbour Sicyon. Under Cleisthenes (reigned *c.* 600–*c.* 570 BC), Sicyon achieved greater power than at any other time. Herodotus, the best source, tells two stories about Cleisthenes of Sicyon. In one he relates how Cleisthenes, being at war with Argos, took several anti-Argive steps in domestic affairs; for example, he tried to suppress the cult of the Argive hero Adrastus. Further, according to the story, Cleisthenes altered the names of the tribes at Sicyon. Previously there were the three Dorian tribes of Hylleis, Pamphyli, and Dymanatae, and a fourth tribe called Aegialeis; Cleisthenes gave his own tribe the name Archelai (Rulers of the People), but he renamed the Dorian tribes, calling them Hyatae (Pig-ites), Oneatae (Ass-ites), and Choereatae (Swine-ites). Thus Herodotus represents the tribal change as merely a change of names carried out because of the war against Argos; Cleisthenes did not want the Sicyonians to have the Dorian tribes in common with the Argives. But something more serious must have been intended. By Herodotus' time the tradition was already becoming anecdotal; the story of the change in names conceals a real reform. Previously, the non-Dorian element in the population—that is, the descendants of those inhabitants who held Sicyon before the Dorian settlers arrived—were underprivileged, at least to the extent of constituting only one tribe in contrast to the Dorian three. Cleisthenes altered this condition drastically; the change was in the nature of a social revolution. Doubtless it was carried out not because of the war with Argos but for internal reasons. Moreover, the conjecture can hardly be avoided that this social reform was the source of Cleisthenes' power: he ruled as the champion of the pre-Dorian element in the population.

Reforms of Cleisthenes

The other Herodotean story tells of how Cleisthenes sought a bridegroom for his daughter, Agariste. On winning a chariot victory at Olympia, he invited prospective suitors to come to Sicyon within 60 days. Men came not only from the Peloponnese but from places as far away as Crannon in Thessaly and Sybaris and Siris in southern Italy. Cleisthenes inquired into their descent and kept them at his court for a year, testing them in all manner of gentlemanly exercises. He was inclined to prefer the two Athenian suitors, Hippocleides and Megacles, and especially Hippocleides. But on the last evening, after dinner, Hippocleides climbed on a table, stood on his head, and danced with his legs in the air. When Cleisthenes told him that he had danced away his marriage, Hippocleides retorted that he did not care. So Cleisthenes chose Megacles, but he announced his choice to the suitors in a banal speech; he said that he honoured them all and would gladly gratify them all, but with only one daughter to give away he could not please everyone.

This story may have gained in the telling, but it reveals much. Hippocleides of the Philaid clan traced his descent to the legendary hero Ajax. Later writers said that Cleisthenes was the great-grandson of a butcher or cook; Herodotus made the point more effectively by putting a banal speech into his mouth. By inviting suitors to compete for his daughter, Cleisthenes made a display of magnificence. He had to; he was a self-made man.

Writers later than Herodotus made Cleisthenes a member of a dynasty, which took its name from its founder, Orthagoras, and which allegedly ruled Sicyon for 100 years; the dynasty may have lasted from *c.* 656 to *c.* 556, but modern research has not fully determined the chronology. Even if Cleisthenes had predecessors in the tyranny, he transformed its basis by the tribal reform. Perhaps the only fact of value provided by the later sources is that Cleisthenes joined in the First Sacred War against Crisa (see below *Relations among the Greek poleis to the end of the 6th century BC: The early Archaic period*), which stood on the north shore of the Corinthian Gulf. Sicyon on the south shore could regard Crisa as a commercial rival, and trade was the ultimate source of Sicyonian prosperity under the tyranny.

The Cypselids. The prosperity of Corinth, too, depended on overseas trade. In the 8th and early 7th centuries the city was ruled by a group called the Bacchiads, whose members traced descent from a legendary common ancestor and were so exclusive as to practice intermarriage within the group. But one of their daughters, Labda, was lame and thus was given in marriage to Aetion, who was not one of the Bacchiads; he claimed descent from the pre-Hellenic inhabitants of Greece and came from Petra, a village in Corinthian territory at some distance from the city. Their son was Cypselus; and when he came of age, he overthrew the Bacchiads and started the Cypselid tyranny of Corinth. He killed some of his opponents and exiled others. After ruling for 30 years he was succeeded by his son Periander, who continued and intensified his father's severities; Periander kept a bodyguard of spearmen, whereas none is attested for Cypselus. Periander was succeeded by his nephew Psammetichus, who was overthrown after a short reign. The Cypselid tyranny is commonly dated *c.* 657–*c.* 584 BC, although somewhat later dates (*c.* 610–*c.* 540 BC) can be defended.

The tradition recorded by Herodotus and Aristotle has not been kind to Cypselus or his dynasty; it stresses their overthrow of the Bacchiads and their alleged cruelties.

Yet a comparison of Cypselid policy with Bacchiad policy reveals not contrast but continuity. The export of Corinthian pottery flourished increasingly until *c.* 575; then the ware began to decline in quality, and after *c.* 550 it yielded to Attic competition, which may have had to do with the fall of the tyranny.

Corinthian colonies

Corinthian trade was helped by colonization. Under the Bacchiads, Corinthian colonies were established at Syracuse and Corcyra *c.* 733. Cypselus founded colonies at Leukas, Anactorium, and Ambracia; these commanded the Gulf of Ambracia and the approach to southern Epirus. Early in the time of Periander, the Corcyreans sent a colony northward to settle at Epidamnus on the coast of the Adriatic; they invited Corinth to send a leader of the colony, and the man sent was a Bacchiad. (This apparently indicates a reconciliation between the Cypselids and at least some of the Bacchiads.) Indeed, Periander controlled Corcyra and eventually sent one of his sons to rule it. A Corinthian colony was founded at Apollonia, somewhat south of Epidamnus, probably in the time of Periander; Epidamnus and Apollonia were situated at the two western termini of one of the few natural routes leading into the Balkan highlands and even into Macedonia. A son of Periander led a colony to Potidaea on the Isthmus of Pallene in the northeast; this, too, was well placed for opening up Macedonia.

Besides colonization, other items of Cypselid policy assisted trade. The tyrants, for example, built a stone causeway, or *diolkos*, across the Isthmus of Corinth. Commerce may have been among the considerations inducing Periander to maintain good relations with Lydia and Egypt.

Other tyrants. Little is recorded of other contemporary tyrants in or near the Peloponnese. Theagenes ruled Megara *c.* 632. According to Aristotle, he caught the flocks of the wealthy at pasture and slaughtered them; this may mark the overthrow of a wool-growing aristocracy. Procles, the tyrant of Epidaurus, is known only from his relations with Periander. The latter married Procles' daughter, but after her death he quarrelled with his father-in-law and overthrew him.

Conditions producing the tyrannies

The conditions producing the several tyrannies varied; no one common cause can be specified. Tension between the Dorian and pre-Dorian elements in the population, for example, was important in Sicyon and in Elis, but it has not been proved for Corinth. Admittedly, in virtue of his paternal descent Cypselus appears to have begun as a man on the fringe of the ruling group, and he may consequently have been motivated by jealous ambition; but his policy, so far as known, did not attempt any drastic change in the social structure of his city. Yet the note of magnificence, common to the different tyrants, may have had political significance. They set a new standard of wealth and power; the first stone buildings in Greece since the Bronze Age were the work of the tyrants. (For example, the Corinthian treasury at Delphi was built by Cypselus.) Such achievements marked a greater concentration of resources and authority than had been known before. It is often said that the tyrants overthrew traditional aristocracies; and this is not incorrect, but it can be misleading since "aristocracy" may suggest a regime within a fully organized city. In fact, under the traditional aristocracies the organization of the *polis* was still rudimentary; the Archaic tyrannies marked a stage in the growth of public power.

SPARTA

Early period. The territory south of the Arcadian highlands, which form the centre of the Peloponnese, consists of two plains bounded by mountain ranges. In ancient times, the eastern plain was Laconia and the western plain was Messenia. The plain of Laconia was bounded by Mt. Taygetus on the west and Mt. Parnon on the east, and it was drained by the Eurotas River. Dorian invaders reached Laconia in the age of migrations; they came as conquerors but mingled to some extent with the previous inhabitants. The earliest Laconian pottery of Proto-Geometric type belongs to the period *c.* 1000–*c.* 950 BC; it may mark the beginning of settled conditions, as a number of townships arose in the plain. One of these, Sparta, situated well up the valley of the Eurotas, consisted of four villages—Pitane, Mesoa, Limnae, and Conoura. The account given by Pausanias of the early expansion of Sparta is credible, although one cannot be sure whether it rests on genuine tradition or on plausible reconstruction. According to this account, the Spartans first moved northward, conquering Aegys at the approach to Arcadia. In the next generation they advanced southward against the strong town of Amyclae—a long war that led to the absorption of that town into the city of Sparta, which thenceforth consisted of five units instead of the original four—and against the towns of Pharis and Geronthrae, whose inhabitants were expelled. Pausanias' account continues by saying that, in the generation following these conquests, the Spartans overcame Helus on the coast; other coastal towns, such as Las and Gythium, probably fell about the same time. These gains should belong to the 9th and 8th centuries; they built up a large state, embracing the plain of Laconia.

Early expansion of Sparta

The Messenian Wars. The population of Greece was growing, and many states sought relief in colonization. Sparta founded only one colony, namely Tarentum in southern Italy; its earliest pottery belongs to the period *c.* 700–*c.* 650 BC. For the most part, Sparta sought land for its surplus population by expanding against its neighbours. Expansion was attempted in two directions. Northeast of Mt. Parnon lay the coastal plain of Thyrea; there Spartan ambitions came into conflict with those of Argos. In 669 the Argives defeated the Spartans at the Battle of Hysiae; this had the effect of excluding Spartan interests from the plain of Thyrea for more than a century.

The other direction of expansion led westward into Messenia. This plain had been settled by Dorians, and numerous townships had arisen there, probably in much the same way as in Laconia. A Spartan attack on Messenia led to protracted warfare. The poet Tyrtaeus, exhorting the Spartans at a late stage in the struggle, said that under King Theopompus "the fathers of our fathers" conquered Messenia; he added that they fought there for 19 years but in the 20th year the defenders fled from the mountains.

Conquest of Messenia

The chronology of the fighting is obscure. After 370–369 BC, when the Theban commander Epaminondas led a successful campaign against the Spartans and restored Messenian independence, a previously sparse tradition on early Messenian history became contaminated with romantic inventions. Many modern historians believe that there were two early Messenian Wars: the first (*c.* 735–*c.* 715), was the first Spartan conquest of Messenia; but a Messenian revolt (*c.* 660) precipitated a second war, in which the Spartans were ultimately successful. The date suggested for the first war rests largely on an argument from the list of Olympic victors: until 736 the victors are from the western Peloponnese and chiefly from Messenia; but after 736 there is no Messenian, and the first Spartan victor is recorded in 716. But the reliability of the early entries in the Olympic list is an unsolved problem. Moreover, it must be admitted that although writers after 369 distinguish between two early Messenian Wars, earlier writers such as Thucydides (writing in the late 5th century) merely mention fighting without specifying any number of wars. Perhaps the most valuable indication of date is a remark of the Theban general Epaminondas; in 369 he said that he had restored Messenia after 230 years. Evidently something concerning Messenia came to an end about 600.

The conquest of Messenia, completed by the end of the 7th century BC, gave the Spartans extensive land. Much of it was divided into allotments for Spartans. The whole state, consisting territorially of Laconia and Messenia, was properly called Lacedaemon; it was a federal state of somewhat authoritarian type. The institutions of the city of Sparta had sole direction of foreign policy and exercised supervision of all Lacedaemon.

Among the inhabitants of Lacedaemon, three main sta-

tuses should be distinguished. First there were Spartan citizens—the most privileged group. Each family of Spartan citizens had an allotment (*klēros*) of land in Laconia or Messenia; the allotments were deemed to be equal. Secondly there were the *perioikoi*, or free inhabitants of the other towns of Laconia and of those Messenian towns that were allowed to survive after the conquest; each perioecid town had some local autonomy. Finally there were the helots, whose name may be derived from the town of Helus. They were serfs bound permanently to the soil; a group of them was assigned to each Spartan allotment of land, and they tilled it for the Spartan masters as well as for their own subsistence.

The new political system. Aristotle, referring to Tyrtaeus as his source, says that during the Messenian War some people in Lacedaemon demanded a redistribution of land. The protracted warfare against the Messenians and the eventual conquest of their land probably caused extensive political changes within Sparta, but the nature of the change is obscure. Here it will be convenient to consider first the institutions of the city of Sparta as they were in the 6th century BC, and then to try to discern the changes that produced them.

Political institutions in the 6th century. The Spartans attributed their institutions to an early lawgiver called Lycurgus; scholars, however, have not determined whether he was real or, if real, what institutions should be attached to his name.

The dual kingship

Sparta had two hereditary kings, drawn from two royal families. It is often conjectured that the dual kingship arose from the early amalgamation of different communities, perhaps from some combination of the original four villages. The powers of each king included command in warfare, and during the Archaic period this carried with it a large share in determining foreign policy. But by the later decades of the 5th century, decisions on foreign policy were taken in the public assembly; and in their military capacities the kings became merely the executors of the assembly's will, although they retained discretionary control of strategy in the field. In home affairs the kings enjoyed some priesthoods and perquisites, they were members of the council, and they exercised judicial powers of three kinds: they judged cases concerning public roads, they arranged the marriages of unbetrothed heiresses, and they presided over all adoptions of sons within the community. These functions may at first seem slight, but in Archaic societies, where public authority was weak and the family was the strong unit of society, matters concerning heiresses and adoptions were important. Later the domestic powers of the kings did not decline, but the state undertook additional functions through other organs.

The council, or *gerousia* (*gerōn*, "elder"), consisted of 30 members including the kings. The other 28 councillors had to be at least 60 years old; they were chosen by acclamation in the public assembly and remained in office for the rest of their lives. The council could advise the kings, served as a court to try capital cases, and probably had the task of preparing business for the public assembly. The extent of its real influence in politics is a delicate question and requires careful attention.

Any adult male citizen of Sparta could attend the public assembly. In addition to electing councillors, this body elected officers called ephors and had the ultimate decision on questions of legislation and policy. Voting was usually by acclamation, a procedure that allowed some discretion to the presiding officer, but occasionally a division was taken.

The ephorate

A king, a council, and an assembly were the usual pattern in any Greek state; but Sparta developed a further major organ—the ephorate. The ephors were five in number and were chosen by acclamation in the public assembly; each group of five held office for only one year. In the 5th and 4th centuries the ephors transacted a good deal of day-to-day business. They presided at meetings of the assembly, and, because they formulated an issue for voting, they could sometimes influence the outcome. Probably they could preside at meetings of the council. They carried out the instructions issued by the assembly for beginning military campaigns, and they transmitted orders to commanders in the field. They received foreign envoys and provided for bringing their business before the assembly. They exercised supervision over the helots.

The ephorate may have existed in some form from very early times. Each month the ephors exchanged oaths with the kings; the kings swore to rule according to the laws; and the ephors, on behalf of the city, swore to preserve the kingship provided that the kings kept their oaths. Since the oath was monthly, not annual, it may have come from a time when the moon, but not the sun, was observed for calendric purposes. But the functions of the ephors were probably small at first and grew later. Spartan tradition held that the powers of the ephorate were much increased by Chilon, who was ephor for a year within the period 556–553, but nothing specific can be said about his work in internal affairs. Scholars once supposed that for periods of many years the ephors pursued a single corporate policy in opposition to the kings, but this is not plausible; because there were five ephors in office together and because they changed each year, it is more likely that rival political groups used every opportunity to get their own sympathizers elected to the ephorate. Indeed, instances are known where one or two ephors pursued a policy distinct from that of their colleagues and their immediate predecessors. To explain the growth in the significance of the ephorate, perhaps the most likely hypothesis is that public authority, as it grew stronger, increasingly assigned new functions to the ephors.

A special question concerns the relationship between the ephors and the kings. By the time of Thucydides, the ephors could imprison the kings; but this was merely a power of arrest with a view to trial, not a power to judge and condemn. An incident occurring about the mid-6th century is suggestive. King Anaxandrides had a wife but no children, so the ephors urged him to divorce her and take another wife in her place. He refused. As a next step the ephors and the councillors together urged Anaxandrides to take a second wife in addition to the first, and they suggested that the Spartan citizens might take action against him if he did not comply. Although he did comply, it is not clear whether he yielded to the authority of the councillors or to the suggested threat from Spartan citizens. But it is evident from the incident that the ephors alone could not coerce the king.

The genesis of Spartan institutions. Such were the main institutions of Sparta in and after the 6th century. Discussion of the way they developed in the previous century has concerned itself mainly with the relationship between the council and the assembly. Plutarch in his *Life of Lycurgus* preserves a document that may have a bearing on this. Although its text is uncertain at some points, probable readings give the following sense:

> After the people has set up a sanctuary of Zeus Syllanios and Athena Syllania, after the people has arranged itself by tribes and obes, after the people has set up a council of thirty together with the kings, let them gather from season to season for the festival of Apellae between Babyka and Knakion; let the elders introduce proposals and decline to introduce proposals; but let the people have the final decision.

Commenting on the text, Plutarch says that two later kings, Polydorus and Theopompus, added a further clause; this addition, now often called "the rider," says: "But if the people make a crooked utterance, let the elders and the kings decline it."

The Spartan *rhētra*

Plutarch says that the document was an oracle, which Lycurgus brought from Delphi, and that the Spartans called it a *rhētra*. The word *rhētra* means here "an enactment"; and the name suggests that the document was an enactment of the Spartan assembly, even if it was based on an oracular response. Its authenticity can be disputed on linguistic grounds. Even if it is authentic, it mentions a good many institutions that required further definition; so it can only have been part of a body of legislation or perhaps a summary of measures, which required further and more specific enactments.

The main clauses of the *rhētra* deal with the decision-

making functions of the council and the assembly, the festival of Apellai being the occasion for meetings of the assembly. The *rhētra* provides that measures are to be introduced by the council but that the final decision shall be taken by the assembly, which thus has at least a yea or nay competence. A two-stage procedure of this kind is attested later in Athens and in other Greek states.

According to a theory accepted by many present-day historians, the two-stage procedure was introduced by the *rhētra* and was a Spartan invention of the 7th century; it appeased strife between the aristocracy, whose organ was the council, and the wider class attending the assembly; and it achieved this by defining their respective shares in determining policy. The theory has also a social aspect. The hoplite phalanx is first attested *c.* 650, and adoption of hoplite tactics may have enabled the Spartans finally to overcome the Messenians; if their adoption spread the burden of fighting to a wider class than before, this may explain why that class demanded a share in making policy.

This theory is impressive. It is less than explicit, however, about the change made by the *rhētra* and about what the previous condition was. A public assembly of some kind must have existed since the age of the migrations, when it began as a parade of the army to receive marching orders; analogies with other Greek states suggest that in settled conditions such an assembly is never likely to have had less than a yea or nay competence. Moreover, a two-stage procedure for determining policy was not necessarily a deliberate Spartan invention adopted subsequently by other Greek states. Comparable procedures are attested in the Roman Republic—where it was customary to bring measures to the Senate before presenting them to the assembly—and among the ancient Germans. Indeed, in the *Iliad*, Agamemnon, having conceived a plan for withdrawal, first discusses it with a "council of elders" and then summons the army to hear it. From these parallels it appears that in Sparta and in other Greek states the two-stage procedure may have arisen from custom and convenience, without any specific enactment designed to appease a conflict.

The problem is complicated by a fragment of a poem of Tyrtaeus. It reads thus:

> After listening to Phoebus they brought home from Pytho the oracles of the god and perfect words: "Let the lead in deliberation be taken by the god-honored kings, who have care of the lovely city of Sparta, and by the elderly councillors, and then let the men of the people, answering with straight enactments. . . .

The fragment breaks off. It can be read as a paraphrase of Plutarch's *rhētra*, but even so it is much more monarchical in spirit. Alternatively, Tyrtaeus' remarks can be adequately understood even if the *rhētra* is inauthentic; in that case he would seek to commend the two-stage procedure in a monarchical spirit by alleging that it was sanctioned by an oracle. Indeed the fragment is good evidence for only one assertion, namely that the Spartans followed the two-stage procedure.

Plutarch introduces the *rhētra* to indicate the importance that Lycurgus attached to the council. If the rider is combined with the *rhētra*, the result gives the council unusual authority, for the rider grants the council some power of veto over the assembly. Obscure passages of Aristotle and Diodorus have been used to explain and illustrate the provision embodied in the rider, but their proper interpretation is not wholly clear. Until recently it was commonly believed that in Sparta the real policy decisions were usually taken in the council; the assembly, it was supposed, failed to assert itself and merely ratified decisions of the council. New research has led to a different conclusion. The best evidence on Spartan procedure concerns the 5th and 4th centuries and is provided by Thucydides and Xenophon; these two authors hardly ever mention the council, but they indicate regularly that real decisions on policy were taken in the assembly.

Discussion of 7th-century changes in Sparta must be largely inconclusive, but some probabilities can be indicated. It is likely that the struggle with Messenia caused internal tension. It is also likely that hoplite tactics were adopted by the Spartans during the 7th century and had political effects. Those effects need not have been an exacerbation of tension; on the contrary, hoplite discipline may have accustomed the citizens to accepting a higher degree of organization. By the early 6th century, Spartan institutions were relatively articulate and defined; a hundred years earlier they had probably been much more amorphous and fluid.

Social organization. Sparta developed for its citizens an elaborate system of public education and discipline, often called the *agōgē*. At the age of seven, children were taken from their mothers and began to pass through a series of annual stages. The content of education included music and poetry as well as physical exercises and games. Girls were freed from public control when they reached maturity, but for boys, education culminated in two years of intensive military training. Then at the age of 20 each man sought admission to one of the living and eating groups, called *phiditia*. Each *phidition* had 15 members and served as a unit on military campaigns; each member contributed to the common stock a fixed monthly amount of food drawn from his *klēros*, or land allotment. The members of each *phidition* lived together until they reached the age of 30, when they could live at home with their wives. For military purposes men were summoned according to year classes, and the first incidence of fighting fell on those between the ages of 20 and 30; but they could be required for service until they were 60 years old.

The agōgē

The *agōgē* probably takes its origin from changes carried out in the 7th or early 6th century BC. It included initiation ceremonies that may go back to an earlier date; what calls for explanation, however, is not the details of the *agōgē* but its incorporation into a well-designed system. As that system trained men for fighting as hoplites, it is not likely to have originated before the adoption of hoplite tactics. The success of the system is reflected in Sparta's military victories from the mid-6th century onward.

The adoption of the *agōgē* may be connected with the introduction of a new way of organizing the citizen body. Originally the Spartans were divided into the three Dorian tribes, membership being hereditary. Tyrtaeus mentions the three tribes in a passage that seems to be an exhortation to Spartans of his own 7th century. But eventually a different method of division, on a territorial basis, was adopted. Spartan inscriptions of Roman date attest citizen units called obes; and the names of these coincide for the most part with those of the five divisions of classical Sparta—with Amyclae and the four original villages. Further, Aristotle says that Sparta had five ancestral regiments; he names them, and one of the names—Mesoa—is that of one of the villages.

Territorial division of citizens

These facts led to the hypothesis that the division into three tribes was replaced by a division into five territorial units called obes. Because membership of the tribes was hereditary, they would have become unequal in size; parallels in other states suggest that the aim in adopting the new division was to distribute evenly the burdens of citizenship, including military service. The problem became complicated by recognition of an Archaic Spartan inscription that names a sixth obe, called Arkaloi. Theories have been propounded that combine the division into three hereditary tribes with a division into numerous territorial obes, but such schemes would leave the ultimate units unequal and thus frustrate the purpose of introducing the territorial division of citizens. The figure five was significant in Spartan institutions; both the ephorate and a board of public messengers called *agathoergoi* consisted of five men each year. This fact and Aristotle's reference to five ancestral regiments suggest that the territorial division of the citizens was fivefold; whether "obe" was the word for the five chief divisions or for subdivisions of them is another question.

The social structure of Sparta produced a strong military force that served both to maintain the city's ascendancy in Lacedaemon and to expand its power in the Peloponnese. It does not follow, as has sometimes been as-

Life in Lacedaemon

serted, that life in Lacedaemon became narrow, militaristic, and inartistic as early as the 6th century. Pindar praised Sparta as the place "Where the counsels of the old men and the spears of the young excel, as do dances and the Muse and splendor." The crafts and trade were conducted mainly by the *perioikoi* because Spartans were excluded from trade. Although the use of coinage was not adopted in Lacedaemon, archaeological finds show that the export of Laconian vases and bronzes continued through the 6th century. Laconian vase painting, though not among the best in Greece, was respectable and reached its highest level *c.* 560–550; its subsequent decline was not as acute as in the case of Corinthian pottery and can best be explained by the competition of Attic ware. Laconian craftsmen continued producing bronze figurines of good quality into the first half of the 5th century.

The condition of the helots was severe, but their misfortunes and discontent in the early period should not be exaggerated. Tyrtaeus indeed compares them to asses worn down by heavy burdens and says that they surrendered half of the produce of the soil. But he wrote during warfare, and "half" may be approximate; Plutarch says that each *klēros* had to supply its master with a fixed and uniform amount of produce, and this suggests that the helots could keep the surplus. For the Battle of Plataea in 479, Sparta supplied 5,000 citizen hoplites and 35,000 helots; the helots were servicing troops, not fighting troops, but it would have been foolish to surround every armed man with seven thoroughly disaffected men, even if the latter were unarmed. Probably relations between Spartans and helots grew much more tense after the helot revolt of the 460s.

The Peloponnesian League. After the reduction of Messenia, and perhaps in consequence of it, Sparta came into conflict with some cities of Arcadia. The fighting culminated in a long war (*c.* 580–*c.* 550 BC) against Tegea. Although the Spartans may have gained some superiority, they were not able to overwhelm the Tegeans, and instead they concluded the war by making an alliance with their enemy. Shortly afterward (*c.* 544) the Spartans resumed their former struggle against Argos; the immediate issue was control of the plain of Thyrea, and this time the Spartans won a decisive victory.

Spartan alliances in the Peloponnese

The alliance with Tegea began a new development. In the 7th century the Spartan policy toward Messenia had been one of conquest and absorption; in the second half of the 6th century, Sparta tried to bring more and more cities of the Peloponnese into an alliance. Moreover, on allying with Tegea, Sparta brought from there the cult of the pre-Dorian hero Orestes; although Sparta remained a Dorian state, by this and subsequent actions it posed as the champion of the pre-Dorian element in the population, thus exploiting an issue that some of the tyrants had used. After defeating Argos, Sparta was evidently the strongest state in the Peloponnese, and during the next decades it succeeded in winning most of the Peloponnesian cities as allies. The resulting organization is called the Peloponnesian League. The stages of its growth are obscure; in 510 the Spartan king Cleomenes I led a force by land against Athens, and this shows that he had the goodwill of Corinth and Megara.

The name Peloponnesian League is modern; the ancient name was "the Lacedaemonians and their allies." The league began from bilateral treaties between Sparta and the several allies. Eventually a congress of the allies met and contracted a multilateral agreement; this may have occurred in the closing years of the 6th century, when Spartan attempts to intervene in Athens precipitated a crisis among the allies (see below *Relations among the Greek poleis to the end of the 6th century BC: Spartan policy under Cleomenes*). The main provision of the multilateral agreement was that all members would be bound by a majority vote of the allies. The league lasted till 366, although its extent varied from time to time. It was a new experiment in interstate organization. Regional leagues elsewhere (*e.g.*, in Aetolia and Boeotia) relied on a fiction of common descent; the Peloponnesian League was an artificial creation of political power.

ATHENS

Early period. The city of Athens, with its Acropolis, was one of the few Greek sites where habitation was not interrupted at the end of the Bronze Age. Indeed, Athens was comparatively prosperous in the Proto-Geometric and Geometric periods. But it is not proven that any political institutions survived from the Bronze Age. Moreover, it is most unlikely that Athens retained in the "dark ages" any ascendancy over other parts of Attica that it may have previously exercised. Attica, Athens' territory in the Classical period, was large, but its unification was only completed at a late date.

Geographically, Attica consisted of several plains divided by hills. The central plain held the city of Athens and a strip of coast, including the bay of Phaleron. Farther west was the plain of Thria, on which the largest settlement was Eleusis. Eastern Attica consisted partly of low-lying territory and partly of hills; and it held several settlements, the most important being at Brauron and Marathon.

Unification of Attica

A lengthy process of Attic unification was probably composed of two stages. In the first, the strongest settlement in each plain acquired control over the other towns of its plain; thus three or four powerful states emerged. In the second stage, Athens, which already controlled the central plain, absorbed the communities of the outlying districts. This second stage was completed not later than the first half of the 7th century BC, but perhaps not much before.

The result was a unitary state; the political organs of the city of Athens were the sole focus of Attic political activity. But in the 7th century, public authority was still weak; real power belonged to locally influential families, each commanding numerous dependents. Recollections survived of conditions that prevailed before the final steps of unification; late in the 5th century, people remembered that there had once been warfare between Eleusis and Athens. Accordingly, at least some of the political conflicts of the 7th and 6th centuries were struggles of "regionalist" type—that is, struggles of leading families in the outer districts against the ascendancy of those in the central plain.

Athenian political institutions

The chief political institutions of the unitary state were an executive board called the "nine archons," a council, and a public assembly. Tradition held, credibly enough, that Athens had once been ruled by kings who retained office for life but whose powers had been diminished in two ways: other offices were created to take over some of their functions, and the term of all the officers was reduced to one year. Hence, the word king survived as the title of one of the nine archons; the others were the archon (or eponymous archon, since his name was used to designate the year), the *polemarchos* ("war commander") and the six *thesmothetai* ("setters of verdicts"). The evidence conflicts on the exact mode of selecting the nine archons, but it is safe to say that until 487 they were always elected by the assembly.

The council came to be called the "council of the Areopagus" because it held some of its meetings on that hill west of the Acropolis. The council consisted of all who had held any of the nine archonships. Doubtless it had begun as a council advising the kings. (The public assembly, too, probably existed in some form from time immemorial, and questions were brought before it for ultimate decision; but probably its meetings were infrequent in the Archaic period.) Athenian writers of the 5th and 4th centuries BC knew that the council had once been powerful, but their attempts to define its former functions produced little more than vague generalities. Probably its ascendancy in the Archaic period rested not on specific powers but on a personal factor. The minimum age for the nine archons was 30 years. As one of the archons, a man had only one year in office, but thereafter he could expect some 30 years of active life as a councillor. The archons were likely to respect the opinion of their seniors, to whose august company they would soon be admitted.

Cylon's uprising

Probably in 632 an uprising led by Cylon tested the stability of the unified state. Cylon had married the

daughter of Theagenes, tyrant of Megara. Encouraged by an oracle, Cylon gathered his friends and a force from his father-in-law and seized the Acropolis during the Olympic festival. But the Athenians came en masse from the fields, beseiged the Acropolis, and entrusted operations to the nine archons. Cylon escaped but his followers ran short of food and made terms with the archons, who agreed to spare their lives; then, as the Cylonians came down from the Acropolis, the archons killed them. Later those responsible for the breach of faith were exiled as being accursed, but later still they were allowed to return. The scandal of the curse was revived more than once against their descendants.

The nature of the rising is obscure. Evidently Cylon hoped to make himself tyrant, but attempts to regard him as a champion of popular discontent have no evidential support. A clue may perhaps be found in the identity of his opponents. Whenever the scandal of the curse was revived, it was used against the powerful family of the Alcmaeonids and their associates; apparently the Alcmaeonids predominated among the archons of 632. As will be seen later (see below *Peisistratus and his sons*), the Alcmaeonids belonged to the central plain or possibly to southeastern Attica. Cylon could get forces without impediment from Megara; and this suggests, though by no means proves, that he belonged to western Attica. Possibly his rising reflects a conflict between families powerful in different regions.

Still greater mystery surrounds the activities attributed to Draco (Dracon) in 621. Later Athenian tradition said that he wrote down all the laws but that subsequently Solon repealed the laws of Draco except those on homicide. This story sounds artificial, and it is more likely that Draco concerned himself only with homicide laws. In 409 the homicide laws of Draco were reinscribed on stone, and the part of the inscribed text dealing with unintentional homicide has been recovered. The law provided that if a man killed another man unintentionally his case should be judged by a "Court of Fifty-One," called the *ephetai;* if they upheld his plea, the relatives of the victim could grant him pardon by a unanimous vote, but if the relatives declined to do so, the state provided the killer with a safe-conduct to the border.

Laws of Draco

Supplemented by evidence from the 4th-century orators, the inscribed text throws a good deal of light on early Attic procedure. The blood feud is implied, and the complementary institution of wergild (payment to relatives of a slain person, according to a determined value for the deceased) is attested. By recognizing the killer's intention as distinct from his act, the law moved toward greater sophistication; by setting up a court to judge intention and by providing a safe-conduct to the border for the unintentional but unpardoned killer, the state asserted its authority to a modest but real degree. Unfortunately the inscribed text of Draco's laws is incomplete.

By the 4th century, Athens had no less than five homicide courts, including the Areopagite council, which dealt with intentional homicide. This system of courts observed other distinctions besides those based on intention; for example, it recognized a category of lawful homicide, and it treated the killer of a slave or foreigner differently from the killer of a citizen. So complex a system must have taken a long time to develop, even though all the provisions may have had an early origin. It is not clear whether all parts of the system were attributed to Draco, and it is not clear which items were innovations of 621. But at least it can be said that by 621 the state had asserted itself in the originally private sphere of killing and the blood feud.

The reforms of Solon. Athens in the Geometric period was prosperous, but in the 7th century BC its economy was comparatively backward. Its failure to found any colonies in that century indicates political weakness at a time when population was doubtless growing in Attica as in the rest of Greece.

Early in the 6th century, Solon—in political poems of which parts are preserved—spoke of acute antagonism between rich and poor. This economic crisis has often been regarded as the result of impoverishment, but the archaeological record suggests a more complex explanation. The export of Attic pottery decorated with black figures on a red background began in the period 620–600, and the ware reached farther afield until *c.* 520; thereafter Attic ware retained the predominance it had achieved, but pottery decorated with red figures on a black background eventually became more popular than the black-figure ware. This export trade was not only in vases but also in their contents, perhaps wine and olive oil. Thus the Athenian economy was already expanding in the early 6th century. The expansion may have brought complex dislocation; in particular, amid growing prosperity, many people may have come to resent conditions they had previously tolerated.

Solon was eponymous archon in 594; then or later he received a special commission to appease the strife and write down the law. Nothing is known about the political steps leading to his appointment; perhaps the normal rivalry between powerful families grew more acute, with the risk that one group of families might exploit the grievances of the poor. The nature of the economic predicament and of Solon's solution appears in part in one of his poems. He says that he removed the "marking stones" (*horoi*), which had held the land in slavery; he adds that he freed men who had formerly trembled at the whims of their masters. These freed men should be identified with the *hektēmoroi*, or "sixth-parters," who according to Athenian tradition tilled the soil for wealthy landowners before Solon's reforms and paid a sixth of the produce to the landowners.

Solon's solution to economic problems

Thus a tolerably clear picture emerges, but the origin of the dependent status of the *hektēmoroi* is another question. Athenian scholars of the 4th century BC held that this status arose from debt, and many modern theorists have made the same assumption. One of the more sophisticated of these theories says that, as population grew, the land was cultivated more and more intensively. Eventually the peasant found that his crop would not suffice for his family until the next harvest; he therefore borrowed from a wealthy neighbour and expected to reap a surplus in the coming year. But because the need to fertilize the soil or to let it lie fallow was not understood, the next year's crop was inadequate; soon the peasant contracted a permanent tie with his wealthy neighbour, paying him a sixth of the crop, and a marking stone was placed on the land to show that it was now under an obligation.

Solon, however, does not mention debt in the extant poems, and perhaps an alternative explanation should be preferred. This says that the position of the *hektēmoroi* was not a recent development due to economic pressures and debt but a traditional status of inferiority, which arose in the "dark ages," when humble men offered their services to powerful men in return for protection in unstable conditions. A dependent status of this kind is attested in several Greek states, including Argos, Sicyon, and Thessaly. Solon in any case abolished the status of the *hektēmoroi*, making them fully free citizens and probably granting them full ownership of the land from which they had previously paid part of the crop. The wealthy landowners thus lost their claim on the *hektēmoroi* and gained little beyond freedom from the threat of violent revolution. The reform was called the *seisachtheia*, or "throwing off of burdens."

Solon tried to draw up a comprehensive statement of the law, and his laws were written on wooden *axones* and *kyrbeis*, objects of disputed character. Some of his provisions are quoted by the Attic orators and by later writers. Athenians in Classical times, however, tended to attribute all their laws to Solon, and thus the question of authenticity is sometimes difficult.

The recent study of indisputably genuine provisions has shown that the penalties known to Solonian law were negative. The state could withdraw legal protection from a culprit, thereby making him an outlaw (*atimos*), or it could utter a curse against him; correspondingly, fines could be imposed only in an indirect way—that is, the culprit could buy himself free from the basic penalties of outlawry or the curse. Positive penalties such as death,

Penalties in Solonian law

imprisonment, and confiscation of property arose only later in Attica, when the state learned to intervene actively. Few things illustrate more clearly the weakness of public authority in the early 6th century.

Some of the laws may have been intended to improve the economy. One of them offered citizenship to those aliens who settled in Attica with their families in order to practice a craft. Another, cited from the first of Solon's *axones*, allowed export of olive oil but forbade export of any other agricultural produce. The effect of such measures should not be exaggerated. They may have been traditional rules, which Solon reduced to writing. Even if they were innovations, it would be naïve to attribute the growth of Athenian prosperity primarily to them; the *seisachtheia* may have increased freedom of movement, but the forces determining the growth of the 6th-century economy were more complex than contemporary legislation. The tradition that Solon changed the standard of Athenian coinage—from that current in much of the Peloponnese to that of Corinth and Euboea—should be rejected; Athenian coinage did not begin so early.

Division into property classes

Modern writers have often attached large significance to those Solonian measures that affected the constitution. The most important of these was the division of the Athenians into four property classes. The first class, the *pentakosiomedimnoi*, consisted of those whose estates produced at least 500 measures of grain, wine, and olive oil annually. Membership in the next two classes, the *hippeis* and the *zeugitai*, depended on annual production of 300 and 200 measures respectively. Those who lacked sufficient land to produce 200 measures annually constituted the fourth class, the *thētes*. Political privilege was graded according to the classes. Only *pentakosiomedimnoi* qualified for the office of treasurer of Athena; the nine archons were drawn probably from the first two classes; the *thētes* enjoyed no privilege beyond membership in the assembly.

A widely held theory maintains that the property classes were a major innovation. It is supposed that after the unification of Attica was completed, political privilege was restricted to a hereditary ruling class, or caste, the eupatrids; they alone, it is said, were admissible to the nine archonships and hence to membership on the council. By the early 6th century some families outside this caste had achieved comparable or even greater wealth and demanded a share in political power. Solon conceded this demand, replacing heredity with wealth as the qualification for office. By so doing he transformed Athens from a closed to an open society.

Several criticisms may be made of this theory. Some of the evidence for the existence of a eupatrid class is poor, and the better evidence does not show when that class monopolized political office; it is conceivable that the eupatrids ruled Athens only at a very early stage, when the long process of unification had scarcely begun. Moreover, as Athens gradually absorbed the rest of Attica, it must have made compromises with the groups ruling at least some of the other communities; the political institutions of Athens and the other towns must have been adjusted mutually to some extent. This process would be likely to break down the exclusiveness of a hereditary ruling class in Athens; and compromises between the leading families of different towns may have even led to the adoption of wealth as the criterion for political privilege. Indeed, study of the names of the property classes suggests a pre-Solonian origin. The name of the first class, *pentakosiomedimnoi*, refers etymologically to qualification by 500 measures, but the names of the other classes are not derived from words for numbers of measures: *hippeis* refers to cavalry service; *zeugitai*, to a yoke of animals. Because the names are thus disparate, Athenian law probably accepted them at different times; in other words, at least the three lower classes were probably recognized as the basis for graduated political privilege before the work of Solon.

The poem of Solon that mentions marking stones appears complete and provides a summary of his work. It devotes 17 lines to the *seisachtheia* but only two lines and a word to his other legislative work; in those two lines he says merely that he "wrote laws alike for the bad man and the good, fitting straight justice into each." If he had carried out a major innovation by introducing the property classes, surely he would have had something to say about it.

The same argument may be used against the attribution to Solon of another major innovation—the Council of Four Hundred. Aristotle says that Solon created such a council, drawing 100 members from each of the four tribes into which the Athenians were traditionally divided. Plutarch adds that the task of this council was to prepare business for the assembly. But neither Aristotle nor Plutarch explains how the members of the council were selected or how long they stayed in office; thus neither has any clear conception of the council. There is no certain reference to the activity of this council in the 6th century; and that is strange if Solon really created such a council, for so serious a step ought to have had some perceptible effect. The question of the authenticity of the Council of Four Hundred attributed to Solon has not been finally resolved; perhaps the easiest hypothesis is that the story of that council is a later fiction, designed perhaps in the 4th century to provide justificatory precedent for the Council of Five Hundred, which Cleisthenes introduced.

Authenticity of the Council of Four Hundred

As comparisons with other ancient states at a similar stage suggest, Solon may have conceived his task for the most part as one of writing the laws down, not of making new laws. Even in the *seisachtheia* he may have believed that he was restoring a pristine condition. Reducing customary law to writing is a major improvement because it diminishes the liberty of judges to favour their friends. It also brings new precision into the law. For example, lawsuits, in the time of Solon, except for cases of homicide, were heard by the nine archons separately (each of the nine archons decided one group of cases), but serious cases could then be referred to the assembly, which was called the *hēliaia* when it sat as a court. In writing down the laws, Solon had to specify how serious an issue had to be before it could be referred to the *hēliaia*. A written statement of the law was a more valuable contribution to the growth of the Athenian state than were the deliberate innovations sometimes attributed to Solon.

Peisistratus and his sons. Thanks to Herodotus, Athenian political conflicts become clear toward 560 BC. The main rivalry was between a group called "the men of the plain," led by Lycurgus, and one called "the men of the shore," led by Megacles. But Peisistratus, who had won popularity by commanding a successful campaign against Megara, organized a third group—"the men from beyond the hills." Pretending to have been wounded by his enemies, Peisistratus persuaded the assembly to grant him a bodyguard of club bearers; with these he seized the Acropolis in 560. But within a few months, Megacles and Lycurgus joined forces and drove out Peisistratus, who withdrew, probably not from Attica but to his private estates at Brauron.

The rise of Peisistratus

A few years later Megacles quarrelled with Lycurgus and allied himself with Peisistratus, who agreed to marry Megacles' daughter. The confederates dressed up a woman as Athena, and she conducted Peisistratus back to Athens. But in 556 Peisistratus quarrelled with Megacles, who withdrew his support; and so this second tyranny collapsed. This time Peisistratus left Attica and stayed away for ten years; he spent much of the time at Mt. Pangaeum in Thrace, where he exploited the silver mines. Several powers worked for his restoration. Thebes, among other cities, gave him money, Argos sent a force of troops, and Eretria let him use its territory as a base. In 546 he sailed from Eretria and landed above Athens, at Marathon, unopposed. When he began marching toward Athens his enemies brought their forces from the city against him; but he defeated them at Pallene and met no further resistance. The tyranny that Peisistratus thus won was passed on to his sons after his death.

Attempts have often been made to explain the conflict

The conflict between Peisistratus and his rivals

between Peisistratus and his rivals in terms of constitutional programs, class interests, and even differences of foreign policy; but such theories are not convincing. The readiness with which Megacles, Lycurgus, and Peisistratus could make and break alliances suggests that the followers of each were bound firmly to the person or family of their leaders rather than to any abstract principle, program, or interest. Moreover, any explanation of the rise of Peisistratus must do justice to the regional names that Herodotus records for the three parties. In Attic usage, the word plain meant the plain of the city of Athens; evidently Lycurgus' men belonged there. The "shore" was a long coastal strip, beginning perhaps within the city plain and extending southeast toward Sunium. Megacles came of the Alcmaeonid family, which belonged in the city during the 5th century; his party of the shore probably had its centre in or near the plain of the city. The party of Peisistratus is called by Herodotus "the men from beyond the hills." (Aristotle calls it "the men of the hills," but this term is less likely to be accurate because by Aristotle's time the tradition had become contaminated by speculation.) Brauron, the home of Peisistratus, was indeed separated by hills from the city of Athens, as was Marathon, where Peisistratus landed unopposed in 546.

The rise of Peisistratus is a clear instance of "regionalist" conflict. Previously there had been moderate rivalry between the parties of Lycurgus and Megacles, both based in or near the city, for the unification had brought ascendancy to men of that district. Bringing his local following to Athens, Peisistratus injected a new acerbity into the political struggle; he expressed the resentment toward leading families of the city felt by leading families of eastern Attica. To say this is not to deny that he won popularity outside his native district; that is proved by the grant of a bodyguard in 560, and after the Battle of Pallene he was accepted by the great majority of Athenians.

Rule of the Peisistratids

Peisistratus ruled until his death in 528 and was succeeded by his eldest son Hippias, who allowed some influence to his brothers. Later Athenian tradition was to speak favourably of the tyrants; Herodotus, Thucydides, and Aristotle said that they ruled well and observed the laws. The tyrants ensured, however, that each year some of their supporters should be among the men elected to office; hence, through the nine archonships there arose within the council of the Areopagus a group of men fully committed to the tyrants, and this may have been the key to their continued rule. Peisistratus and his sons tried to conciliate as many Athenians as possible. Herodotus says that the Alcmaeonids and some others withdrew into exile in 546, and this is credible. But the Alcmaeonids were persuaded to return before 525, when their leader, Cleisthenes, was archon.

It is often asserted that on winning his third tyranny, Peisistratus seized land from his leading opponents and distributed it to peasants. This is a modern conjecture unsupported by ancient evidence, although it is not unlikely; a possible redistribution of land was much discussed in 6th-century Athens. Peisistratus introduced the first direct taxation in Attica; he exacted a proportion of produce, recorded variously as one-twentieth (by Thucydides) and one-tenth (by Aristotle). Aristotle says that Peisistratus created judges to tour the countryside (*dikastai kata dēmous*), so that the state provided jurisdiction even in distant parts of its territory. The same office was certainly created in 453, and it is difficult to see how it could have lapsed before then. Thus the attribution to Peisistratus may be apocryphal; but if it is genuine, the establishment of this office in the 6th century was an important step toward increasing the unity and internal power of the state.

An ambitious foreign policy gave the Athenians an increased chance to take pride in their city (see below *Relations among the Greek poleis to the end of the 6th century BC: Athenian foreign policy*). The same can be said of the building policy of the tyrants; although little can be asserted about this with confidence, at least a temple was built on the north side of the Acropolis to house the ancient wooden image of Athena. Magnificence, a common note of tyranny, is apparent among the Peisistratids; such poets as Anacreon of Teos, Lasus of Hermione, and Simonides of Ceos were invited to sojourn at their court.

Fall of the tyranny

An incident occurred in 514 that began a series of events that brought about the fall of the tyranny. Two men—Harmodius and Aristogeiton—felt they had been slighted by Hipparchus, a younger brother of Hippias, and they plotted to overthrow the family. But their plans were ill conceived; they succeeded only in killing Hipparchus, and in the consequent investigation Hippias came to believe that discontent was widespread. He therefore made his rule more severe, killing many Athenians. The Alcmaeonids withdrew into exile again. A year or so later they led a party of exiles into Attica and fortified a position at Leipsydrium, but they were driven away by the forces of Hippias; evidently the bulk of the Athenians was not willing to turn against the tyranny.

The Alcmaeonids resorted to diplomacy. They won the favour of the Delphic oracle, which accordingly urged the Spartans repeatedly to overthrow Hippias. A first Lacedaemonian expedition was sent by sea under Anchimolius, but it was unsuccessful; Hippias had gained cavalry from his allies, the Thessalians, and when Anchimolius landed his force in the bay of Phaleron, the cavalry overcame him. In 510 the Spartans sent a land expedition under King Cleomenes I. He defeated the Thessalian cavalry and besieged the Peisistratids on the Acropolis until they yielded and agreed to leave Attica; they withdrew to Sigeum.

The above narrative of the overthrow of Hippias is the Athenian account, the only one to survive. It emphasizes the role of the Alcmaeonids, but it does not conceal the fact that the fall of the Peisistratids was due to foreign intervention; the step fatal to Peisistratid rule was a decision taken in Sparta, although the Delphic oracle was useful in providing religious justification.

Rivalry between Cleisthenes and Isagoras

The reforms of Cleisthenes. Some time after the fall of Hippias, rivalry developed between Cleisthenes and Isagoras, who was archon in 508. Cleisthenes was worsted, but he instituted reforms that gave him ascendancy. So Isagoras appealed to Cleomenes, with whom he had contracted friendship during the campaign that overthrew Hippias. Cleomenes sent a herald and then came himself to Athens; he demanded the expulsion of Cleisthenes and of 700 families descended from those who had become accursed by slaying the followers of Cylon (see above *Athens: Early period*). Further, he tried to dissolve the council and replace it with 300 supporters of Isagoras; but the Athenians rallied to the support of the council and expelled the Spartan king. Cleomenes tried once more to make Isagoras master of Athens: he brought the army of the Peloponnesian League to Eleusis and persuaded the Boeotians and Chalcidians to attack Attica from the north and the northeast, respectively. But at Eleusis the Corinthian contingent in the Peloponnesian force mutinied, and the force disbanded; then the Athenians defeated the Boeotians and Chalcidians successively.

It is difficult to ascertain with clarity the nature of the rivalry between Cleisthenes and Isagoras. Inferences from the measures of Cleisthenes are not easy to draw, because the program of Isagoras is not known. But Isagoras' local origin may provide a clue. Herodotus observed that the relatives of Isagoras worshipped Zeus Karios. In Attica the cult of Zeus Karios is attested only at the village of Icaria just north of Mt. Pentelicus. It is therefore likely that Isagoras came from an easterly district separated by hills from the city of Athens. The possibility must be considered that his conflict with Cleisthenes was "regionalist" in character. Perhaps Isagoras attracted the support of those Areopagites who had become councillors through their loyalty to the Peisistratids; not many men had been expelled with Hippias in 510.

New tribal system

The reforms of Cleisthenes introduced a new tribal system. Previously Athenian citizens had formed four hereditary tribes, but thenceforth these were retained only for some religious purposes. Cleisthenes introduced a divi-

sion of the citizens into ten new tribes, constructed in a complex way. The basic unit was the deme—a village or parish; all Attica, including Athens itself, was now divided into more than a hundred demes, which varied in size. To form tribes from the demes, Cleisthenes divided Attica into three large districts, called "the city," "the coast," and "the interior." Each of these districts was subdivided into ten *trittyes*, or "thirdings"; their boundaries were drawn so that each *trittys* held one or more complete demes. Then to form each tribe Cleisthenes put together three *trittyes*, drawing one from each of the three districts.

Each deme had its own assembly and its annual headman, or *dēmarchos*. It kept a list of its adult male members, and membership in a deme constituted proof of citizenship. In the first instance men were assigned to demes because of local residence; but for the future, membership was hereditary. The *trittyes* failed to develop any extensive functions and served essentially as a means of distributing demes among tribes. Each tribe had its own officers and assembly. Moreover, the tribes were used for distributing the burden of military service; thenceforth the main Athenian land force was ten regiments of hoplites.

Other ancient states underwent a change from hereditary tribes to territorial tribes when the former became seriously unequal in size and could not absorb new immigrants; the introduction of territorial tribes served to equalize burdens. Considerations of this kind may have influenced Cleisthenes, but they do not explain the peculiar tripartite nature of the new Attic tribes. Some historians, unable to discover any precise purpose, have followed Aristotle in saying that Cleisthenes sought merely to mingle together citizens of different districts and thus break down local barriers. But this explanation is inadequate; the reforms gave Cleisthenes ascendancy over Isagoras, and so they must have had a partisan character. Some recent historians have observed that in outlying parts of Attica the boundaries of some *trittyes* were drawn in such a way as to break down powerful units. For example, in the northeast a unit of four townships—called the "Tetrapolis" and including Marathon—continued to exercise religious functions, but it was split between two *trittyes* and hence between two tribes. Such steps were doubtless deliberate, but they could have been achieved through unitary tribes; they do not explain why each tribe had demes from different districts.

Explanations for tribal systems

No fully satisfactory explanation for the Cleisthenic tribal system has been offered. Yet it must be admitted that the tripartite composition of the tribes looks like a measure taken in a "regionalist" struggle; it may have promoted the ascendancy of the leading families residing in or near the city. How it might serve this end can only be discerned in part. With one exception the assemblies of the several tribes met in the city; thus influential families of the city could bring their dependents to vote more easily and in greater numbers than could influential families of outlying districts. Again the rule of candidature for office may be significant. In 501 the office of general (*stratēgos*) was introduced or reformed; thenceforth each year ten generals were elected by the assembly of all citizens, but one had to be chosen from each tribe. Likewise the number of archons was increased from nine to ten through the addition of the secretary of the *thesmothetai*, although they continued to be called the nine archons. If one member of the board of ten generals was to be chosen from each tribe, then, because each tribe held some city territory, all members of the board might be city men; and this was likely to happen because the electoral assembly met in the city.

Council of Five Hundred

The other major reform of Cleisthenes was the introduction of the Council of Five Hundred. The members of this were chosen by lot and held office for one year; in the 4th century one second term was permitted, but at first even this may not have been allowed. Each tribe supplied 50 members to the council, and within the tribes each deme supplied a number of councillors roughly proportionate to its size. The task of the council was called *probouleusis* and consisted of preparing business for the assembly; the rule was that no item could come before the assembly unless it had been considered previously by the council.

The reason for establishing the new council is not difficult to discern. As Aristotle observed, there had to be a smaller council to prepare items for a public assembly. A mass meeting of several thousand could not deal with business in the raw. But the preparatory council, if it had the will, was in a position to influence the outcome of the assembly's vote by its formulation of the issues. The members of the Council of Five Hundred, however, were not men of particular talent or ambition because they were chosen by lot; and they did not acquire expertise because they served only for one year or, at most, two. Thus by virtue of its personnel the council was not likely to assert itself. In practice it prepared business efficiently but left the real decision to the assembly; its existence ensured that no other body would encroach on the assembly's authority.

Aristotle, perhaps correctly, attributed to Cleisthenes a further measure—the introduction of ostracism. By this practice, in late winter of each year the assembly was asked whether or not it would hold an ostracism; if it answered affirmatively, at the next possible meeting each citizen deposited a potsherd on which he had written the name of the man who he would most wish to have leave Attica. A valid ostracism required 6,000 votes. The man against whom the most votes were cast had to leave Attica for ten years but retained possession of his property. Attempts to link the presumed reasons for this custom with the other measures of Cleisthenes are not persuasive. The origin and purposes of ostracism remain opaque.

The practice of ostracism

In the second half of the 5th century, Cleisthenes was regarded as the founder of the democracy. The introduction of the new council may have been even more effective than the tribal reform, both for the outcome of the struggle with Isagoras and for the further development of Athens. The council ensured that real decisions were taken in the assembly, and in a sense this was democratic. But the word *dēmokratia* had not yet been coined in the time of Cleisthenes, and there was no synonym; his purposes should not be sought in democratic ideas. The assembly met in the city, and so it could most readily be influenced by families of the city and its immediate neighbourhood. Cleisthenes brought "regionalist" struggles to an end by winning victory for the city.

Society and the economy. In antiquity the Athenian economy depended upon the natural resources of Attica. Noteworthy mineral resources included silver in the southeast and clay for pottery. Attica's pronouncedly Mediterranean-type climate—mild temperatures and less rainfall than in many parts of Greece (the annual average is 16 inches, but there is considerable variation from year to year, and virtually all the rain comes in winter)—suited the territory to fruits such as the vine, the olive, and the fig, which can send roots deep down into the soil and in summer draw on moisture stored up from the previous winter.

Economy in the late 7th century

In the late 7th century BC, Attica was predominantly a land of subsistence farming. The main crop was grain—especially barley, which is hardier than wheat—but vines were also grown. There was some fishing. Relatively few animals were kept. Marginal land was brought under cultivation to support the growing population. But as the record of black-figure ware shows, exports were beginning to expand even before 600. Furthermore, there was contact with the flourishing commercial cities of Corinth, Megara, and Aegina. The crisis facing Solon was due not only to the poverty that arose when population grew beyond the local means of subsistence but also to the desire to achieve better conditions.

The measures taken by Solon are now obscure, but something can be said about their effects. Thenceforth rural Attica was a land of free peasants; there is no later trace of any quasi-servile status, such as that of the *hektēmoroi*. On the other hand, wealthy men could still command a large following of humble dependents. (As late as 480, at the Battle of Artemisium, Cleinias supplied his own trireme, which he equipped and manned at his

own expense; this means that he could command about 200 able-bodied men.) By abolishing the status of the *hektēmoroi*, Solon increased the mobility of rural manpower; thenceforth a subsistence farmer who failed might drift to the city. A tradition about Peisistratus may indicate some growth of an urban proletariate; Aristotle says that the tyrant lent money to poor men so that they could settle as farmers. Such a scheme probably concerned newly cleared land on the hills, where vines and olives could be grown but required capital.

Growth of trade

The distribution of black-figure ware is the clearest indication of the growth of the Athenian economy. In the period 560–520, Attic vases reached Chios, Lesbos, Cyprus, Asia Minor, the coasts of the Black Sea, Egypt, south Italy, Etruria, and southern France. (Many of the vases were of high artistic quality and prove the existence of skilled potters in Athens.)

Some political developments helped the trade. Sometime between 600 and 560 the Athenians captured the island of Salamis from Megara. Thus, the bay of Phaleron became safe from attack, and trade with Corinth was facilitated. Athenian goods reaching the west may have travelled in Corinthian ships. After Peisistratus had finally made himself tyrant, an Athenian expedition under Miltiades settled the Thracian Chersonese on the European shore of the Hellespont, and the tyrants kept up the connection with the settlement into the next generation. Peisistratus captured Sigeum in the Troad from the Mytilenaeans and sent one of his sons to rule there as tyrant. The expeditions to Sigeum and the Chersonese suggest that Peisistratus was interested in the trade route through the straits to the shores of the Black Sea.

The effect of coinage on the Athenian economy is difficult to determine. Studies conducted since World War II have shown that the earliest coins were struck in Lydia *c.* 640–*c.* 630, not *c.* 700 as previously supposed. The invention was adopted gradually by many Greek cities. The earliest Attic coins, called by numismatists "Wappenmünzen," bore a variety of heraldic designs, which may stem from different families; these were struck not before *c.* 575. They were superseded by coins bearing the design of the owl, the emblem characteristic of Athens, and these began not before *c.* 527. Some scholars indeed would put the earliest "Wappenmünzen" soon after 550 and the earliest "owls" *c.* 510.

While in exile, Peisistratus drew silver and perhaps gold from the mines of Mt. Pangaeum; this enabled him to raise forces for his restoration and to maintain a bodyguard thereafter. He may have begun working the silver mines at Laurium in southeast Attica. The Peisistratids were not the only family to interest themselves in foreign sources of wealth. A legend of the Alcmaeonids told how one of their ancestors, Alcmaeon, visited the Lydian king Croesus and came away laden with gold. Part of a grave monument honouring an Athenian named Croesus and erected *c.* 530 has been found to the southeast of Mt. Hymettus and has been claimed as indicating the home of the Alcmaeonids. This inference is not necessary; there is no reason to suppose that the Alcmaeonids were the only Athenian family to develop relations with the Lydian kingdom.

Foreign sources of wealth

Although the stages of economic growth are obscure, by 500 Attica was a very different place than it was a century before. It had become the leading exporter of vases among the Greek cities; its other exports included oil, wine, and silver. It imported grain, fish, and ship timber. The population was growing, and economic expansion had attracted immigrants. According to Herodotus, toward 500 the number of adult male citizens was 30,000; the reorganization by Cleisthenes had entailed some survey of the citizen body, and this figure may be approximately correct.

OTHER IMPORTANT POLEIS

The Greek *polis* was usually small, and therefore special interest attaches to those *poleis* that achieved rule over a comparatively large territory. Chief among these were Sparta, Athens, and Thebes. Lacedaemon was, strictly speaking, a federal state, although most power belonged to the dominant city, Sparta. Athens organized Attica as a unitary state. Boeotia, however, in which Thebes was located, was a federation of about a dozen cities that varied greatly in size and power.

Thebes. Geographically, most of Boeotia consisted of two river basins—those of the Asopus in the south, where Thebes predominated, and of the Cephissus in the north, where Orchomenus was the strongest city. There was rivalry between Thebes and Orchomenus, and usually Thebes had the superiority. Starting in or before the 6th century BC, the Thebans built up a federation of Boeotia, which was dissolved in 479. Nothing is recorded about its constitution, but it is fair to infer from the structure of a restored confederacy (447–386) that it was organized on relatively liberal principles. (The constitution of the restored Boeotian confederacy is known and has remarkable features: each city managed local affairs, and the supreme federal organ was a representative council of 660 members; in 395 Thebes supplied 240 of these—more than any other city but less than a majority.) Apparently by 519 the early federation embraced much of Boeotia. Orchomenus, however, did not belong, and on the southern border Plataea refused to join; attacked by the Thebans in 519 (or possibly *c.* 509), the Plataeans gained the alliance of Athens and hence preserved their independence.

Boeotian confederacy

Phocis and Thessaly. Northwest of Boeotia, Phocis lay as a broad wedge separating Locris into two parts. As late as the early 5th century, the men of East and West Locris were still aware of their mutual bond and cooperated to found Naupactus on the north shore of the Corinthian Gulf. For a long time in the 6th century, Phocis was on the defensive in an intermittent conflict with Thessaly; finally about 500 the Thessalians invaded Phocis, but the Phocians inflicted heavy defeats on the Thessalian infantry and cavalry in two successive battles.

Thessaly itself, a large plain surrounded by four ranges of mountains, stood apart from the political development of the rest of Greece. It was dominated by a landowning aristocracy; indeed its several cities may have been brought into existence by their local dynasties, such as the Aleuads of Larissa and the Scopads of Crannon. These cities were strongholds or market towns rather than centres of any lively political development. Intermittently, the leader of a local dynasty gained recognition of his ascendancy in all Thessaly and took the title of *tagos;* but at other times federal organization was dormant.

Overseas settlements. The Aegean islands and the cities of western Asia Minor were the scene of a particularly brilliant development in the Archaic period. The affairs of Mytilene, the chief town of Lesbos, are known in part from the fragmentary poems of Alcaeus (*c.* 620–*c.* 580). His lifetime saw the final overthrow of the Penthilids, a clan that had once controlled the city. A complex series of factional struggles within the aristocracy followed; there is no sufficient warrant for the common assumption that Pittacus, one of the contenders, led a popular movement. Alcaeus was at first an ally but later an enemy of Pittacus and was driven into exile, probably more than once. Eventually the Mytilenaeans chose Pittacus as *aisymnētēs*, a position which Aristotle described as an elective tyranny. Pittacus' primary task was to preserve the city from the threat posed by the exiles. He seems to have succeeded in bringing civil strife to an end, and he then laid down his office. Hellenistic scholars said that Pittacus ruled for ten years (590–580), but the chronology of Mytilene is difficult. Alcaeus in exile said that he longed for the assembly (*agora*) and the council (*bolla*).

Developments in the Aegean islands

Farther south, on the island of Chios, an inscription of *c.* 575–*c.* 550 attests a *bole demosie*, a term often taken to mean a "popular council" as distinct from an "aristocratic council," although other interpretations are possible; for example, the term may mean a "state council" as distinct from "local councils." At least the occurrence of the epithet suggests that there was more than one council in Chios, and this indicates some sophistication of political development.

On the west coast of the Asiatic mainland, the Ionian

Ionian settlements of Asia

settlements were among the most prosperous Greek cities of the Archaic period. They held the alluvial plains at the mouths of major rivers, which flow westward into the sea; so they could expand considerably before feeling the pressure of population on local means of subsistence. Moreover, the rise of the Mermnad dynasty in Lydia meant that the Ionian settlements had a strong, stable, and wealthy power as their neighbour; excavations at Sardis show that Lydian trade with Greek cities increased steadily during the Mermnad period. It is not surprising that the 6th-century Ionians were ahead of other Greeks in some spheres; for example, Thales of Miletus, who predicted the year of the solar eclipse of 585, began the series of Greek scientists. Miletus, the richest Greek city of the Asiatic mainland, had a tyrant named Thrasybulus, who exchanged friendly messages with Periander. Afterward the city suffered a long period of internal strife, but sometime before 500 it achieved stability by inviting mediators from the island of Paros; by the settlement these mediators made, power appears to have been entrusted to men of landed wealth.

Phocaea, the northernmost city of mainland Ionia, did much to open the western part of the Mediterranean to Greek enterprise in the 6th century. The Phocaeans developed commercial relations with the flourishing kingdom of Tartessus in Spain (*c.* 620–*c.* 540). On the south coast of France the Phocaeans founded the colony of Massilia (the modern Marseille) *c.* 600; the Massiliotes in turn founded dependent colonies as trading stations in a westward line reaching into the northeast coast of Spain. The Phocaeans also founded a settlement in Corsica called Alalia *c.* 560. When Persian forces attacked Phocaea *c.* 540, the Phocaeans withdrew by sea and, although some returned, a large party sailed to Corsica and joined the colony of Alalia. But after practicing piracy for five years, the people of Alalia were attacked by the combined fleets of the Etruscans and the Carthaginians; although the Greeks claimed to have won the battle, they had to withdraw from Corsica, and, after a sojourn at Rhegium, they settled at Elea (Velia) in southern Italy. The Etruscans took Corsica, and the Carthaginians took Sardinia, which they subdued after a struggle with the natives; together the two powers exercised ascendancy in the western basin of the Mediterranean.

Greeks in Sicily and southern Italy

In Sicily the 6th century marked the beginning of a protracted conflict between Greeks and Phoenicians. The latter held three bases—at Motya, Panormus, and Solus—in the western part of the island; they were often joined by the people of Segesta in their opposition to the Greek advance from the east. The first known conflict occurred *c.* 580, when Pentathlus of Cnidus led a force from Cnidus and Rhodes to settle at Lilybaeum—a position commanding the harbour of Motya; he was driven out by the Phoenicians and Segestans. In the next decades Carthage gained leadership over the Phoenician settlements in Sicily. In 510 Dorieus, the half-brother of Cleomenes of Sparta, led a party of perhaps 1,000 men to settle at the foot of Mt. Eryx, but he was overcome by the combined forces of Carthage, Segesta, and the Phoenicians of Sicily.

In southern Italy Sybaris, an Achaean colony, built up a loose federation of Achaean colonies from *c.* 550. Croton, another Achaean city, became its chief rival. Soon after 530 Pythagoras the philosopher settled in Croton; his political teaching is not clear, but his followers gained predominant influence in the city and set about expanding its rule. In 510 the Crotoniates took advantage of an internal dispute in Sybaris to attack and destroy the latter; at the time of its destruction Sybaris was perhaps the most thriving of Greek colonies. Croton then became the leading power in southern Italy, although the alliance it created was not so extensive as that formerly led by Sybaris. The Italiote cities played a prominent part in Greek thought and culture in the late 6th century; they produced such men as the philosopher Parmenides of Elea and the poet Ibycus of Rhegium. Croton had a flourishing school of medicine; and one of its doctors, Democedes, served for a time at the court of Darius I of Persia.

RELATIONS AMONG THE GREEK POLEIS TO THE END OF THE 6TH CENTURY BC

The early Archaic period. Early in the Archaic period the structure of the *polis* was weak and therefore there was little official intercourse between *poleis*. Thucydides remarked that in early Greece there were hardly any wars sufficiently extensive to bring in many cities; he noted as an exception the war between Chalcis and Eretria, neighbouring cities on the island of Euboea, in which many other Greeks supported one side or the other. This struggle is called the War of the Lelantine Plain because it was in part a conflict for control of the plain of the Lelanton River in Euboea. The allies of Eretria included Miletus, and those of Chalcis included Samos. Chalcis also gained help from Pharsalus in Thessaly. (Some of the issues were local; Samos and Miletus went to war for control of Priene as late as 441.) The date, duration, character, and outcome of the Lelantine War are far from clear; a good case can be made for a date in the late 8th or early 7th century BC, although a later date can also be reasonably defended. If the earlier date is correct, rivalry for colonial sites probably exacerbated the struggle.

Local leagues

Although dealings between cities were slight, local leagues arose at an early date. These were predominantly religious, but they sometimes acquired a political character. An important example was the League of Amphictyones, comprising several powers of central and northern Greece from Boeotia to Thessaly. It arose to protect the cult of Demeter at Anthela near Thermopylae, but it acquired control over the sanctuary of Apollo at Delphi. The oath of its members included a promise not to destroy or cut off the water supply of any member-city, even during warfare.

Another local league included 12 Ionian cities—Samos, Chios, and ten cities on the Asiatic mainland. Members built a temple—the Panionion—on the northern slope of Mt. Mycale and gathered there for periodic festivals. When the Ionians felt themselves threatened by the Persian advance *c.* 544, they gathered at the temple for political deliberation.

The early tyrants in the Peloponnese brought greater cohesion to their cities and asserted themselves abroad. The colonies founded by the Cypselids and the dealings of Periander with Procles of Epidaurus have been noted above. Periander maintained good relations with Oriental powers. For example, after seizing 300 Corcyrean boys as hostages during a dispute with the island, he sent them as a present to King Alyattes of Lydia (the present did not arrive; the convoy put in at Samos and the Samians enabled the hostages to escape). Periander's nephew was called Psammetichus—a non-Greek name (Psamtik) borne by some pharaohs of the Saite dynasty (664–525), which suggests Corinthian dealings with Egypt.

First Sacred War

Cleisthenes of Sicyon fought a war against the Argives, but his great opportunity was provided by the First Sacred War. This war (*c.* 595–*c.* 586) concerned the Delphic sanctuary and the Phocian town of Crisa. The latter had a good harbour on the northern shore of the Corinthian Gulf; it controlled a fertile plain, through which ran the route taken by pilgrims to Delphi. According to the tradition—known from texts written more than two centuries later—the Crisaeans levied tolls on these pilgrims; and the Amphictyonic League, exhorted by a response of the oracle, went to war against the Crisaeans because of their impiety. The war developed into a siege of Crisa; the northern powers blockaded it on the landward side, but it received supplies by sea. Thereupon Cleisthenes joined in the war; he built a fleet and hence stopped the transport of supplies to Crisa. The town fell in 591–590, although pockets of resistance still held out. Cleisthenes received a share of the spoils and won a chariot victory at the Pythian festival of 582. The war brought him distinction in Greece, and his fleet brought him power.

The war led to reorganization of Delphic affairs. The previous condition is not clear. Thenceforth the Amphictyonic League had ultimate control over the sanctuary; the town of Delphi managed its own affairs and supplied the priestess and interpreters for the oracle. The Pythian

festival was reorganized on a larger scale, with athletic and other competitions; it was to be held every four years, and the year 582 was reckoned as the first Pythiad.

Reorganization of festivals

The Pythian reorganization seems to have prompted the reorganization or foundation of other panhellenic festivals. About 580 the Eleans finally gained ascendancy over the Pisatans and reorganized the Olympic festival, held likewise every four years; there had been a festival at Olympia before, but its early history is obscure. (In the second half of the 5th century, Hippias of Elis published a list of Olympic victors; he assigned the first celebration of the festival to 776 and all subsequent scholarship followed him, but the reliability of the entries in the early part of his list is an unsolved problem.) The reorganization of *c.* 580 may have caused the festival to attract visitors from a wider area than before. In Corinthian territory the Isthmian Games were founded *c.* 580 and celebrated every two years. In the territory of Cleonae the Nemean Games were founded in 573 and likewise celebrated every two years; it is not clear whether Cleonae, which lay between Sicyon and Argos, was under the influence of Cleisthenes or of the Argives at the time of the foundation.

In general it is disputed whether panhellenic festivals owed their expansion in the early 6th century to the tyrants or to other regimes. The fact emerges, however, that the expansion of the festivals reflects the same growth of wealth and ambition as do the early tyrannies. The four festivals—Pythian, Olympic, Isthmian, and Nemean—were panhellenic in that they admitted all Greeks without regard to such distinctions as that between Dorians and Ionians. They attracted visitors and competitors of all kinds, but their contribution to promoting panhellenic consciousness should not be exaggerated; they did not acquire political significance.

The later Archaic period. In the 7th and 6th centuries BC, the overseas activities of Greeks on the European mainland increased. The fortunes of the Asiatic Greeks, and later of those on the European mainland, were affected increasingly by the non-Greek powers of Asia Minor.

Lydian influence

Influence of non-Greeks. The kingdom of Lydia was rich in gold and silver and in an alloy of the two—electrum—which occurred there naturally. In founding the Mermnad dynasty at Lydia, Gyges (*c.* 680–648) seems to have filled a power vacuum. He made indecisive attacks on some Ionian cities; he raided Miletus and Smyrna, and he captured the citadel of Colophon. This policy was continued by his successors: the third and fourth members of the dynasty, Sadyattes and Alyattes, fought a war of 12 years against Miletus. In each of the first 11 years the Lydian forces plundered the Milesian crops before the harvest; but the sea routes to Miletus remained open, and in the 12th year Alyattes concluded a treaty of friendship and alliance with the city. His successor Croesus attacked Ephesus and then the other Greek cities of the Asiatic mainland and made them pay tribute, although Miletus retained its privileged status; regular tribute was more satisfactory to all parties than were intermittent raids. Under Croesus the Lydian kingdom reached its widest extent, stretching as far as the Halys River in the east.

In spite of the raids and the tribute, the Mermnad dynasty brought advantages to Greeks. Trade with Lydia allowed a steady growth in Ionian prosperity. Greeks seem to have travelled freely in Lydia; later tradition told of the hospitality shown by Croesus to the Athenians Solon and Alcmaeon; and although details are suspect, such stories attest the King's philhellenism. No Lydian garrisons are recorded in Greek cities. Looking westward beyond the mainland, Croesus achieved treaties of friendship with Ionians of the offshore islands. He sent gifts for dedication in Greek sanctuaries at Ephesus, at Branchidae near Miletus, at Thebes, and at Delphi. He exchanged presents with Sparta.

About 544 (the precise date is disputed) Croesus went to war against Cyrus, the founder of the Persian Empire. Cyrus drove the Lydian forces back to Sardis, besieged the city, and overcame Croesus. The Greek cities of the Asiatic mainland had to consider their attitude toward the new power approaching them from the east. Before Cyrus left Sardis, the Greeks sent envoys and asked him to let them continue in the same status they had enjoyed under Croesus. Cyrus confirmed the Milesians in their previous condition, but he would make no undertaking to the others because they had refused his request that they desert Croesus while the war was in progress.

After a short stay in Sardis, Cyrus hastened to Ecbatana in Media. The Lydians promptly rose in revolt and were joined by the Greek cities of the Asiatic mainland. Cyrus sent forces that reduced Lydia anew and then captured the Greek cities piecemeal; Miletus alone continued to enjoy special treatment. Thus, by *c.* 540 the Persian Empire had acquired the Greek cities of Asia Minor. These cities became subject to the nearest satrap, or provincial governor, such as the satraps resident at Sardis and Dascylium; and the imperial administration usually tried to maintain order by supporting a local man as tyrant in each city. The cities were required to pay tribute, but generally their burdens were not severe.

Persian rule

From the mainland of Asia Minor the Persian administration might have hoped to extend its authority to the offshore islands, but the steps taken are obscure. Details are available only for Samos. This island was ruled from *c.* 535 by the tyrant Polycrates, who raised it to unprecedented power; he practiced piracy extensively and maintained the strongest fleet in the Aegean. Polycrates seems to have taken advantage of a change in naval engineering; probably during his time the trireme was invented. His contacts were not only with Greeks. He exchanged presents with Ahmose II, the pharaoh of Egypt; later (in 525) Cambyses, the son and successor of Cyrus, launched his successful expedition to conquer Egypt, and Polycrates sent 40 triremes to help Cambyses. Toward 522 Oroetes, the satrap of Sardis, conceived a design for conquering Samos; he enticed Polycrates into a conference on the mainland and crucified him. But before Oroetes could proceed further, Cambyses was overthrown, and, after a struggle for the succession, the usurper Darius made himself king. Darius had Oroetes assassinated. Meanwhile, Maeandrius, the second in command to Polycrates, ruled Samos; but shortly afterward Darius sent a special force under Otanes to attack the island. The force was guided by Syloson, the exiled brother of Polycrates. Otanes sacked Samos and then entrusted it to Syloson.

How far Persian rule extended over Greeks *c.* 514 is known from Herodotus' account of the expedition that Darius led into Scythia. Darius conducted his forces across the Bosporus and thence northward and across the Danube; meanwhile the Greek tyrants from cities under his control sailed ahead to the Danube to bridge the river, and Herodotus gives a list of them. They included men from such cities of the Ionian and Aeolian mainland as Miletus, Phocaea, and Cyme, and also tyrants from several cities on the Asiatic shore of the Hellespont and Propontis, notably from Abydos, Lampsacus, Parium, Cyzicus, and the neighbouring island of Proconnesus. The only Aegean islands represented in the list were Samos and Chios; the tyrants present from the European mainland were Miltiades, the Athenian ruler of the Thracian Chersonese, and one from Byzantium. Elsewhere Herodotus remarks that Coes brought a force from Mytilene and that Darius made him tyrant of that city as a reward for his services on the expedition.

Spartan foreign policy. The foreign policy of Sparta in the 6th century was determined first by the city's position in the Peloponnese and later by its attitude toward developments across the Aegean. Sparta's wars with Tegea (*c.* 580–*c.* 550) and Argos (*c.* 544) and the growth of the Peloponnesian League were noted previously (see above *Sparta: The Peloponnesian League*). In its dealings with powers outside Greece, Sparta at first followed—consciously or otherwise—the practice of tyrants like Periander. Relations were opened with Croesus toward the middle of the century. The Lacedaemonians sent to Sardis to buy gold for a proposed statue of Apollo, and Croesus gave them the gold free. Thereafter the Lacedae-

Spartan relations with Lydia

monians sent a large ornamental bowl of bronze as a present to Croesus; the bowl only travelled as far as Samos, and the Lacedaemonians said that it was seized there by Samian pirates. Relations were opened likewise with Ahmose of Egypt. Shortly before the dispatch of the bowl to Croesus, Ahmose sent the Lacedaemonians an embroidered corslet of linen, decorated with gold; this, too, travelled no farther than Samos.

The tradition reaching Herodotus portrayed the dispatch of the bowl as the Spartan response to a request from Croesus for a military alliance against Cyrus. Likewise it said that when Croesus was besieged in Sardis by Cyrus, he appealed to Sparta for help, and, although the Lacedaemonians were engaged in their war with Argos, they prepared to send a fleet, but that the news arrived of Sardis' capture before the fleet could sail. These stories may have gained something in the telling. The historian should beware of statements of unrealized intentions; the fact emerging from the alleged preparation of a fleet is that Sparta did not send aid to Croesus. Even so, the exchanges of presents may have had political implications. There may have been more presents than are recorded, and at the least they indicate commercial possibilities.

Herodotus tells a tantalizing story of action taken by the Ionians immediately after Cyrus overcame Croesus. Allegedly they gathered at the Panionion and sent envoys to Sparta for help. The Spartans decided to give no help, but they sent a penteconter (50-oared galley) to Phocaea and one of its crew proceeded to Sardis. There he gave Cyrus the message of the Lacedaemonians—that they would not tolerate the Persians ravaging any Greek city—and Cyrus made a contemptuous reply. In this story it is not difficult to recognize an element of fiction. The account of the Lacedaemonian penteconter and message to Cyrus was surely invented later to gloss over the Spartan refusal of aid to the Ionians. But the Ionian appeal to Sparta may well be authentic; it attests the standing that the Lacedaemonians had achieved in Greece and the extent to which they had interested themselves in affairs beyond the Aegean.

Decline of Spartan overseas operations

Such a development of Lacedaemonian interests had been possible in the time of Croesus; accordingly it was checked by the Persian conquest of Lydia. A further check to Spartan enterprise in the eastern Aegean occurred *c.* 525. In response to the appeal of some Samian exiles, the Lacedaemonians and the Corinthians sent a fleet to try to overthrow Polycrates. The force besieged Samos for 40 days unsuccessfully; then it withdrew.

Modern historians have sometimes supposed that in consequence of the final subjugation of Messenia, Sparta concentrated on developing its power by land and declined to acquire interests overseas. This is incorrect. In the middle of the 6th century Sparta concerned itself with developments across the Aegean, but by 520 its attempts to look eastward had suffered more than one setback. Thereafter until 412, it made little attempt at naval enterprise and concentrated on extending its influence by land operations. Such a policy appears in the reign of Cleomenes I (*c.* 520–*c.* 490), but before considering his early activities, attention must be given to the foreign policy of Athens.

Athenian foreign policy. Late in the 7th and early in the 6th century, the Athenians scarcely developed foreign interests beyond border warfare with their immediate neighbours. There was a long struggle with Megara, the main issue being control of the island of Salamis. During this war Peisistratus led a campaign sometime before 560, in which the Athenians captured Nisaea, the eastern harbour of the Megarid; it is not clear when the Megarians recovered Nisaea. The main outcome of the fighting was that the Athenians won Salamis.

There are slight traces of Athenian enterprise farther afield. An Athenian contingent, commanded by Alcmaeon, took part in the First Sacred War; thenceforth, Athens had a vote in the Amphictyonic Council. A less reliable tradition said that the original resolution of the Amphictyones against Crisa was proposed by Solon, but this may be a later invention of Athenian pride. Many historians believe that late in the 7th century an Athenian force commanded by Phrynon fought the Mytilenaeans for control of Sigeum in the Troad. The historicity of this war depends on a complex problem of chronology: if indeed there was such fighting at so early a date, then it was probably not an official act of the Athenian state but the work of some Athenians. Certainly the question of Sigeum arose, or arose again, in the time of Peisistratus.

Athenian foreign policy may be said to begin with the final tyranny of Peisistratus and his sons (546–510). Peisistratus established an Athenian presence in the neighbourhood of the Hellespont, he asserted Athenian influence in the heart of the Aegean, and he maintained good relations with several powers of the Greek homeland. In the first of these spheres he captured Sigeum from the Mytilenaeans and entrusted it to Hegesistratus, one of his sons. Hegesistratus had to conduct further fighting against the Mytilenaeans, but he remained in control of Sigeum; indeed, the Peisistratids withdrew there when they were driven from Athens in 510.

Athenian presence in the Hellespont

The foundation of an Athenian settlement in the Chersonese on the shore of the Hellespont was still more ambitious and was probably carried out soon after 546. The settlement was invited by the natives of the Chersonese—the Thracian Dolonci—who desired reinforcements against the attacks of their neighbours, the Apsinthii. The settlement was led by Miltiades of the Philaid clan, which apparently belonged to Brauron because the deme in which Brauron stood was called Philaidae. After bringing Athenian settlers, Miltiades built a wall across the isthmus of the Chersonese and waged war indecisively against Lampsacus on the other shore of the Hellespont. At one stage in the war Miltiades was taken prisoner by the Lampsacenes, but he was released through the intercession of Croesus. At his death Miltiades entrusted the Chersonese to Stesagoras, the son of his uterine brother Cimon. The war with Lampsacus continued and in consequence of it Stesagoras was assassinated. Thereupon (*c.* 516) the sons of Peisistratus sent out from Athens another son of Cimon, who bore the name of his uncle Miltiades. The dispatch of the younger Miltiades shows that the Peisistratids were keen to preserve their connection with the Chersonese.

In the heart of the Aegean, Peisistratus had a political debt to pay. In 546 Lygdamis, an adventurer from Naxos, helped him with men and money in his final restoration. In return Peisistratus subdued Naxos and entrusted it to Lygdamis. (The latter helped Polycrates to make himself tyrant of Samos.) Peisistratus also carried out a ceremonial purification of the island of Delos, the main sanctuary of the Ionians, by digging up and removing all corpses buried within sight of the temple; by the purification Peisistratus took advantage of the Athenian claim—first attested in the poems of Solon—to be the metropolis of all other Ionian cities.

Several cities in the Greek homeland aided Peisistratus in his final restoration; they included Thebes, Argos, and Eretria. One of his sons bore the name Thessalus, which suggests that the family had ties with the Thessalians; Hippias' alliance with Thessaly was apparent in the last years of his rule, when Thessalian cavalry helped him against Lacedaemonian attacks. The Peisistratids seem to have also enjoyed the friendship of Sparta for a time; some years after the overthrow of Hippias the Spartans conceived a plan for restoring him and remembered that the Peisistratids had been their friends. (International "friendship" was a formal matter, akin to alliance, and Spartan recollection may have been correct.)

Athenian alliances

In large part the tyranny owed its foundation and its preservation for many years to the favour of other cities; neighbours of Attica, like Thebes and Eretria, and powers farther afield were keen to see a *persona grata* in control of Athens. Developments abroad eventually weakened the Peisistratids. A Spartan force overthrew Lygdamis of Naxos, perhaps during the return of the unsuccessful Spartan expedition against Polycrates of Samos, the friend of Lygdamis; the attack on Lygdamis may not have been intended to harm the Peisistratids, but it removed a friend of theirs. A few years later the Persians overthrew Polycrates.

Some time later the Plataeans were attacked by the Thebans; on the advice of Cleomenes, who happened to be near with a Lacedaemonian force, the Plataeans sought and gained the alliance of Athens. (The traditional date of concluding this alliance is 519; it has been disputed but cannot be proved wrong.) Herodotus, recounting the alliance, says that the Lacedaemonian aim was to create discord between Athens and Thebes. That may be an anachronistic inference from conditions obtaining in his own time, but the conclusion of the alliance did create discord between Athens and Thebes; the Athenians promptly sent a force that defeated the Boeotians.

If the Plataean alliance was concluded in 519, then the international position of Hippias had been impaired well before 514. After the assassination of Hipparchus, Hippias sought new friends abroad; he gave his daughter in marriage to the son of Hippoclus, the tyrant of Lampsacus. According to Thucydides' recording of the marriage, Hippias knew that the Lampsacene tyrants had influence with Darius. Hippias may also have recognized that the makers of Persian policy were showing a new interest in the West. At least Hippias secured the prospect of a refuge in the Persian Empire, should he be overthrown.

Spartan policy under Cleomenes. Spartan policy, as it developed in the early years of Cleomenes, followed recognizable principles. On the one hand Sparta gave up its attempts to extend its influence by naval enterprise across the Aegean. This meant that Sparta would not oppose the expansion of the Persian Empire, provided that it did not reach European Greece. Thus when the Persians under Otanes finally attacked Samos, Maeandrius sailed to Lacedaemon and appealed to Cleomenes for help; instead, Cleomenes advised the ephors to expel Maeandrius from Lacedaemon, and this was done. Again, when Darius had attacked the Scythians, the latter sent envoys to Sparta and offered an alliance for military operations against the Persian Empire, or so Herodotus says; the account may be imprecise, and the only result was that Cleomenes learned from the envoys the habit of drinking unmixed wine.

On the other hand, the Spartans tried to enlarge their ascendancy by land operations on the Greek mainland; this led them into conflict with friends of Persia, when Persian influence appeared to encroach on European Greece. The alliance of Hippias with Hippoclus of Lampsacus—a friend of the Persians—may well have induced the Spartans to adopt extreme measures against Hippias, measures that culminated in the expulsion of the Peisistratids from Athens. Whether Athens became a member of the Peloponnesian League in 510 is perhaps a question of terminology. Certainly in the ensuing decade the Spartans tried repeatedly to ensure that there should be a favourable regime in Athens. As related above (see above *Athens: The reforms of Cleisthenes*), some time after the Peisistratids' fall Cleomenes tried to support Isagoras against Cleisthenes, first by sending a herald and then by coming himself to Athens, where he attempted to dissolve the council; but the Athenians drove him out. Thus the Athenians had quarrelled with Sparta, and so they turned to the alternative source of support; they sent envoys to Sardis to seek an alliance with Persia, and, at the request of the satrap, the envoys gave earth and water, the tokens of submission. Hence for a time Athens was technically a vassel of Persia.

Spartan–Athenian antagonism

As also mentioned previously, Cleomenes responded to the rebuff he received at Athens by calling out the army of the Peloponnesian League and leading it against Attica, but at Eleusis the Corinthian contingent mutinied and the force disbanded. This operation caused a crisis in the league, and a congress of envoys from the member-states was summoned. Doubtless questions of authority and command were discussed; indeed, this may have been the occasion when a multilateral agreement was first accepted, superseding the bilateral treaties from which the league had arisen. For the congress the Spartans invited Hippias from Sigeum, and they proposed to the allies that he should be restored to Athens. But first the Corinthians and then the other allies rejected the proposal; it was abandoned and Hippias returned to Sigeum.

As a result of the congress, the Spartans gave up trying to control Athenian affairs. The Athenians, however, soon became perturbed because asylum was granted to Hippias within the Persian Empire, and they sent envoys to Sardis to complain. The satrap replied with a brusque instruction: they should restore Hippias, if they cared for their safety. The Athenians were not willing to comply, and thus their close relationship with the Persians came to an end.

6th-century changes in Greece. The 6th century saw two major changes in the balance of power in European Greece. One of these was the development of the Peloponnesian League; by 500 it was the largest single force in interstate affairs, capable of providing a nucleus for resisting a foreign invader and fit to serve as a model for other alliances. The other change was the rise of Athens to mature statehood. Before 600 the Athenians had done little to assert themselves beyond the borders of Attica, and for a long time afterward Athens was much subject to foreign influence; other cities played a large part in the final restoration of Peisistratus and even in the overthrow of Hippias. But the ambitious policy of Peisistratus had aroused new aspirations among the Athenians; in the decade following 510 they came to assert themselves against Sparta and Sardis, and thus they achieved full independence. (B.R.S.)

III. The 5th century BC

THE PERIOD OF THE GRECO-PERSIAN WARS

The Ionian revolt and its consequences. The sequel to the conquest of Lydia by Cyrus the Great was the subjection of the Ionians and other east Greeks. The Persians were completely alien and their capital more than two months distant. Not understanding Greek political attitudes, they based their control on local tyrants, despite the fact that any form of absolute rule had become an anachronism to the Greeks. It was thus more for political than for economic reasons that the Ionians revolted in 499 BC. This revolt, led by Miletus, spread from Caria to the Hellespont and, for the first two years, resulted in a remarkable degree of unity. It began with the expulsion of the pro-Persian puppet rulers and an appeal to the mainland for help. Sparta, the dominant military power in Greece, rejected the appeal, but Athens, which claimed to be the mother-city of the Ionian cities, sent 20 ships, which were joined by five more from Eretria. The united forces sacked the Lydian city of Sardis but were caught on their return near Ephesus, suffering heavy losses. The Athenians seem to have become convinced that the military prospect was hopeless, and consequently they withdrew; the Ionians, however, fought on. Not until 494 was their revolt crushed—by a naval defeat at Lade, followed by the reduction of Miletus.

The punishment of Athens now became a natural Persian objective, and in 492 a strong force was sent around the north coast of the Aegean; but the fleet was wrecked off Mt. Athos and the expedition was abandoned. Two years later a second force was sent, this time through the islands across the Aegean. Traitors opened the gates of Eretria, and the population was enslaved; the Persians then landed in the Bay of Marathon. The Athenians, roused by Miltiades, the most dynamic of their generals, decided to march out and meet the Persians on the Plain of Marathon. There they were joined by the Plataeans and by clever tactics and tough fighting won a decisive victory against more than twice their number. When they marched out they had sent a messenger to Sparta, but the Spartans waited for the full moon and arrived the day after the battle.

The Battle of Marathon (490 BC)

In the decade following Marathon there were keen political rivalries at Athens. Miltiades, the hero of Marathon, was discredited in the following year when an expedition against Paros, which he had advocated and led, failed miserably. The death sentence was commuted to a heavy fine, but he died shortly afterward. There followed a quick succession of ostracisms; in 487 and the next two years the victims were friends of the Peisistratids, who were suspected of treason at the time of

Marathon. Later, other issues became more important: the most critical was a struggle between Aristides and Themistocles, probably over naval policy, which resulted in the ostracism of Aristides in 482. This struggle was associated closely with a dramatic decision that, more than any other, influenced the course of Greek history in the 5th century. A particularly rich vein of silver had been discovered in the mining area of Laurium, and it was at first intended to distribute this windfall among the citizens. Themistocles, fully aware of the growing danger from Persia, persuaded the people to use the money to build triremes. His new Athenian navy was to prove one of the decisive factors in the repulse of the Persians.

The return of the Persians under Xerxes. The Persian force that was broken at Marathon had limited objectives and may not have exceeded 30,000: the aim of the next, much larger, expedition was the conquest of Greece. The Greek states met at the Isthmus in 481 to concert plans for their defense. But not all of the Greeks were prepared to join in active resistance. The Peloponnesian League followed Sparta's lead, but Argos remained neutral and suspected, and north of the Isthmus only Athens could be firmly relied on; Thessaly even reached an accord with the Persians; and the other states were not likely to expose themselves unless it became clear that they could be protected. The Greek army was eventually posted at Thermopylae to hold the narrow pass between mountain and sea, while the fleet, stationed at Artemisium in north Euboea, was to prevent the Persian fleet from moving south and landing troops behind the Greek lines. Some 4,000 men held up the vastly superior masses of Persians that Xerxes threw into battle, but the Greek position became untenable when a small Persian force was guided along a mountain path that brought them down behind the Greek position.

Victory of Xerxes at Thermopylae (480 BC)

The failure to hold Thermopylae meant the abandonment of Greece north of the Isthmus. Athens was evacuated, though a small company remained to defend the Acropolis. The Greek fleet assembled at the island of Salamis, while Xerxes moved south, ravaged Attica, and, after a short siege, took and sacked the Acropolis. Themistocles, who commanded the Athenian fleet, persuaded the Greeks that they could only hope to win at sea if they fought in narrow waters. He then sent a messenger to Xerxes, assuring him that the Greeks were on the point of flight and in no state to resist; Xerxes closed the western channel and sent his main fleet into the straits between Salamis and the mainland, expecting wholesale surrender. Instead, the Persians found the Greeks fully prepared for battle; in the narrow waters numbers were a handicap, ramming and boarding was the general pattern, and the Greek marines had better arms and a more robust morale. The victory was complete, and Xerxes withdrew with his fleet.

The Persian general Mardonius was left with a substantial army in Greece, and the next year, after raiding and ravaging as far south as the Isthmus, withdrew to a camp on the Asopus near Thebes. In August a Greek army of some 35,000 hoplites and more than 50,000 light-armed troops, led by the Spartan regent Pausanias, advanced to give battle. The decisive engagement was fought near Plataea, when Pausanias, having failed to draw Mardonius, ordered a general retreat by night from a position that the Persian cavalry had made untenable. Mardonius attacked in the early morning, expecting an easy victory, but the Spartan wing took the main shock and their superior arms and spirit decided the issue. Mardonius was killed. The remnants of the Persian army fled in disorder, and their camp was sacked. At roughly the same time news was brought that the Greek fleet, considerably reduced after Salamis, had won a great victory. King Leotychides, the Spartan commander, having heard that the Persians guarding Ionia were demoralized and that the Ionians were eager to revolt, crossed the Aegean to Samos. The Persians were encamped on Cape Mycale, where the Greeks landed unopposed and stormed the camp. The victory was the signal for a widespread revolt in Ionia.

The western Greeks. The Greeks of the west had not been able to help in the defense of Greece because they had troubles of their own. While the Phoenicians of the east formed the backbone of Xerxes' fleet, the Phoenicians of Carthage were hoping to drive the Greeks of Sicily out of the island. In the late 8th and early 7th centuries the Greeks had established colonies in the east of the island and the Phoenicians in the west at Motya, Solus, and Panormus. For more than 100 years no hostilities are recorded between the two peoples, but friction developed in the 6th century when Carthage—which by then was the strongest and richest Phoenician city in the west and which had made an alliance with the Etruscans—attempted to exclude Greeks from the western Mediterranean.

By 500 it was merely a question of time before a serious crisis arose. It came in the form of a quarrel between tyrants. Gelon of Syracuse had marriage alliances with Theron of Acragas. Their main enemy was Anaxilas of Rhegium, on the toe of Italy. He had seized Zancle on the opposite Sicilian coast and was allied with Terillus, tyrant of Himera, who was on good terms with the Phoenicians. Theron drove out Terillus and made his son ruler of Himera. Terillus appealed to Carthage, and the Carthaginians sent Hamilcar, one of their chief magistrates, with an army that was probably much larger than the Greeks could muster, though Herodotus' figures (300,000 men) are no doubt exaggerated. The Carthaginians landed near Panormus and proceeded to besiege Theron in Himera. Gelon marched across Sicily with 50,000 foot soldiers and 5,000 cavalry, and in a great pitched battle Hamilcar was killed (according to one account he sacrificed himself) and his army routed. The two tyrants were left in possession of a rich booty and thousands of prisoners. As a result of this decisive victory Sicily was to be safe from Carthaginian invaders for more than 70 years. But in 409 the Carthaginians returned and took their revenge. Himera was destroyed and Acragas sacked. Syracuse was saved by another tyrant, Dionysius I, who for 30 years maintained his power by equivocal tactics and saved the east of Sicily from Carthaginian control.

Battle of Himera (480 BC)

Gelon died two years after his victory, and the rule of Syracuse passed to his brother Hieron, who in 474 won a great naval victory off Cumae against the other main enemy of the Greeks in the west, the Etruscans. This victory was of much more than local significance. The Etruscans, who in the 6th century had extended their rule south of the Tiber and into the rich lands of Campania, had lost control of their land communications when they were defeated in a great battle near Aricia by an army of Cumaeans and Latins, which was soon followed by the expulsion of the last Etruscan king from Rome. The destruction of their fleet off Cumae meant that the Etruscans were cut off from Campania and would no longer be a threat to the Greeks of the Bay of Naples and farther south.

Greek offensives against Persia in Asia Minor. After the Battle of Plataea, the Persian army had abandoned Europe and was not to return. The battle of Mycale was the beginning of another chapter. What was to be the future of the Ionians, many of whom were now in revolt? The Spartans proposed that they should be transferred to the Greek mainland from which they had originally come and be settled on the lands of the states who had medized (collaborated with the Persians). From a military point of view this was good sense. History had shown that without strong help from outside they were subject to the power that controlled Asia Minor. The Ionians, however, had no wish to leave their good lands; they were supported by the Athenians, and the plan was abandoned. It was now near the end of summer, but the Athenians insisted on proceeding to siege Sestos, which was the Persian headquarters in the Chersonese; the Peloponnesians preferred to sail home. By the end of winter Sestos had been starved into surrender and the Chersonese, which was vital to the protection of the corn route from the Euxine, was once again under Athenian control.

Between the battle of Plataea and the summer of 478 the Greeks had decided to continue operations and take the offensive against Persia. The Spartan regent Pausanias was again given the command and led a force of some 100 ships to the island of Cyprus, where Persian garrisons were driven out and the Greeks of Cyprus were encouraged to take control of the island. From Cyprus the fleet sailed up to Byzantium, which commanded one of the two easiest passages from Asia to Europe and the entry to the Euxine. At Byzantium Pausanias became increasingly unpopular with the allies, and there were even rumours that he favoured the Persians, which led to his recall to Sparta. Meanwhile the Athenian commander Aristides had been invited by the Ionians to take over command of the united forces. History had shown that Sparta was ill-adapted to overseas commitments; Athens could claim to be the mother-city of the Ionians and had sent ships to them in the Ionian revolt. Admittedly, they had been withdrawn at the first serious reverse, but it was now a very different Athens. The victory of Marathon, the creation of the largest and most modern navy in Greece, and its major contribution to the destruction of Persian sea power had increased its confidence and stimulated its ambitions. The Spartan army was still the strongest in Greece, but in 479–478 the Athenians had succeeded in rebuilding their fortifications, despite strong Spartan protests. Considering these developments, then, Aristides probably had a shrewd appreciation of Athens' interests when he persuaded the Athenians to accept the leadership of a new league.

Formation of the Delian League

In the winter of 478–477 Aristides, no doubt in consultation with the leaders of the contingents from Lesbos, Chios, and Samos (who provided the largest of the allied fleets), worked out the basic structure of the new league; all Greeks who were interested were invited to attend an inaugural meeting in the summer of 477 on the island of Delos.

Thucydides describes the essentials of the organization:

> The Athenians decided which cities were to contribute money and which ships . . . The *hellēnotamiai* were then first instituted as an Athenian office and they received the *phoros*, which was the name for the money that was collected. The first assessment was four hundred and sixty talents; Delos was their treasury and the meetings were held in the sanctuary. The allies were at first autonomous and policy was discussed at meetings which all shared.

To symbolize the permanence of the association lumps of metal were sunk in the sea and oaths were exchanged. Aristides, on behalf of Athens, and representatives of the allies bound their states by oath "to have the same friends and enemies." From the outset, the influence of Athens, as the freely accepted leader, was paramount. It was understood that it would always provide the commander for military operations and, probably, the chairman of league meetings. It is doubtful whether any formal guarantee of independence was laid down, for the allies were not in a suspicious mood. They realized that without Athenian leadership they would again become a fringe of the Persian Empire, and their feeling was one of undiluted gratitude.

The original alliance included the Cyclades with the exception of Melos and Thera, which had been settled from the Peloponnese; most of the Ionian and Aeolian coastal towns from Miletus to Byzantium; and probably the Dorians to the south, comprising the three cities of Rhodes (Ialysus, Lindus, Camirus), Cos, and Cnidus. In southwest Asia Minor some Carian communities joined at once, but most of the Carians from the south coast and the interior were not won over until later. The rich cities of the Thracian coast may not have joined the alliance until the middle 470s, but most of the Greek colonies of Chalcidice seem to have followed the example of their mother-cities in Euboea and Andros. By the middle of the century the large islands of Lesbos, Chios, and Samos were the only ship contributors. This was because a much larger number had by then taken the easier option of paying an annual tribute instead.

For some ten years the Athenians continued to lead league forces against the Persians, liberating Greek cities and extending membership of the league. The operations, culminating in a spectacular victory at the mouth of the River Eurymedon (*c.* 466 BC) were carried out under the command of the Athenian general Cimon, and these years could fairly be called the age of Cimon. His series of victories over the national enemy made him the hero of the day, and his family background and his associates gave him a powerful political base. Meanwhile, Themistocles had steadily lost ground. Looking ahead, he advocated anti-Spartan policies, which had little appeal while Cimon, who believed in a dual leadership of Greece, was winning victories in the east.

Cultural developments. The cultural revival of Athens after the Battle of the Eurymedon was at least encouraged, if not led, by aristocratic patronage. The years immediately following the expulsion of the Persians from the Greek mainland had been very bleak. The Acropolis and the lower city had been razed to the ground; public and private buildings had been reduced to rubble and ashes. The main concentration at first must have been on new living quarters and the restoration of the buildings that were essential to the carrying on of public business, but before long architects had new opportunities. When Cimon brought back the bones of the legendary hero Theseus from the island of Scyros a new special shrine was built for them; and from the spoils of the Battle of the Eurymedon he built a new south wall for the Acropolis, enlarging its monumental area.

It is significant that the architects of the Cimonian buildings are now nameless, while the painters whose work they accommodated are remembered. For this was the period of the first blossoming of monumental painting in Greece, a period of bold experiment breaking away from the restricting formulas of the Archaic period. To later generations, Polygnotus was the greatest of these painters. Born in Thasos, the son of a painter, he was probably attracted to Athens by Cimon, who had campaigned in Thracian waters.

Achievements in sculpture

In the sculpture of the period 480–450 BC the same spirit can be seen as in the painting of Polygnotus and his contemporaries. Archaic conventions are discarded and the main concern of the sculptor moves from the surface and linear pattern to a study of solid form. The Archaic smile disappears, and the unreal Archaic symmetry is replaced by more realistic standing figures with the weight of the body unevenly distributed as in nature. A wider and more enterprising range of subjects is chosen, including, toward the end of the period, Myron's bronze figure of an athlete, with body tensed, ready to throw the discus—a bold and difficult subject that was still widely copied in the Roman period. The crown of this generation's achievement was on the Temple of Zeus at Olympia, which was completed by 456. The two pediments, with their contrasting scenes of calm and conflict, and the Labours of Hercules on the metopes, were the finest architectural sculpture that Greece had yet seen.

This was a leaner period for poets at Athens. There was much trivial versifying in honour of Cimon but nothing to match the brilliant circle of poets who had been attracted by the Peisistratids. Simonides of Ceos retained his association with Athens sufficiently long to celebrate its praises in verses on the battles of Artemisium and Salamis, but in 476 he moved to the court of Hieron at Syracuse. Pindar was still in his prime. Before the Persian invasion he had lived and learned at Athens, and he paid a handsome tribute to its share in the preservation of freedom, but the Medism of his native Boeotia strained his feelings toward Athens, and he saw dangers ahead in the development of Athenian policy. His world of aristocratic virtues and privilege was passing away, but there is magic in the rhythm and the imagery of the victory odes for his wealthy patrons in Sicily and Cyrene and for athletes from Aegina and other mainland states.

Athenian tragedy

The most exciting literary development in this period was the growing maturity of tragedy, which was to remain the particular pride of Athens. Aeschylus, who was always regarded later as the greatest of the early trage-

dians, had fought at Marathon and Salamis and was profoundly moved by what seemed to him the divine salvation of Greece. In his *Persae*, produced in 472, he presented the Persian defeat as a godsent punishment for the hubris of Xerxes, and throughout his career he remained convinced that the divine power who controlled the universe must be fundamentally just. His main concern was to explain the ways of God to man; his characters are cast in the heroic mold, and his language is vivid and colourful. Most of his plays were produced during the Cimonian period, but by the year of his *Oresteia* (first performed in Athens in 458) Cimon had been ostracized, and new democratic forces were transforming Athens.

THE ATHENIAN EMPIRE

The changing nature of the Delian League. Beneath the political calm of the Cimonian period there had been undercurrents that were soon to come to the surface. At the time of the Eurymedon victory there was already a growing change in the attitude of Athens to the alliance. In *c.* 472 Carystus, at the south end of Euboea, which was not an original member of the Delian League, had been forced by military action to join. When, a little later, Naxos wished to secede, it was besieged. Naxos capitulated, and terms were dictated that were incompatible with its independence. It is possible that the revolt of Naxos was instigated by medizing oligarchs, and its reduction may have been in the genuine interest of the league, but when Thasos revolted in 465 the main issue seems to have been Athenian encroachment on what Thasos regarded as its own economic interests on the mainland. Its navy was defeated and, after a two-year siege, the island surrendered. Thasos had to give up its fleet, pay a war indemnity, and bring annual tribute. Athens took over Thasian mining rights and other interests on the mainland, but the attempt to settle a colony of Athenians and allies at Ennea Hodoi on the Strymon ended disastrously. Penetrating inland, the settlers were defeated by a combined Thracian force, and the colony had to be abandoned.

When Thasos revolted, it had appealed to Sparta and, according to Thucydides, Sparta promised to help by invading Attica. This was one of the first signs that the peaceful coexistence of Athens and Sparta was becoming strained. Hitherto the Spartans had been too preoccupied with their own troubles to worry about the growing strength of Athens. After their resignation of the leadership against Persia, Sparta had sent King Leotychides, the victor of Mycale, to Thessaly to unseat the medizing dynasty of the Aleuads, but the expedition was ineffective and ended with the trial of the King for bribery and his exile from Sparta. There followed a period of political agitation in the Peloponnese. Sparta had first to face an alliance of Argos and Tegea and then an Arcadian alliance that only Mantineia refused to join.

Pausanias, after being acquitted of treachery at Sparta on his first recall, in 478, had sailed out privately and occupied Byzantium, from which he was forcibly expelled by Cimon. He then settled at Colonae in the Troad and from there seems to have made approaches to the Persian satrap at Dascylium. He was recalled to Sparta and, after a surprisingly long delay, incriminating correspondence with Xerxes was alleged to have been found; he was also suspected of stirring up trouble with the helots. Pausanias fled for refuge to the temple of Athena of the Brazen House, from which he was taken when on the point of death from starvation. It is impossible now to recover the true story of Pausanias' last years, but it is probable that the Spartan authorities were more concerned with his possible future plans for Sparta than for his alleged Medism. The Spartans also claimed to have found evidence implicating Themistocles in the designs of Pausanias, and his enemies in Athens brought a charge against him in his absence. He fled from Argos and eventually reached the Persian king, who made him governor of the Greek city of Magnesta on the Maeander. There he died.

When Thasos revolted against Athens in 465 BC, Sparta promised help but was distracted by a serious earthquake, which provided the occasion for the most serious helot revolt since the 7th century. The Spartans were able to subdue the helots in Laconia, but the main strength of the revolt was beyond Mt. Taygetus in Messenia. There the helots took refuge on the very defensible slopes of Mt. Ithome, from which the Spartans could not dislodge them. At this point they called for help to their erstwhile allies in the war against Persia and, among them, to Athens. In the Athenian assembly, Cimon, well-known for his sympathy with Sparta, urged support, but there was strong opposition from new men who were working for a more democratic Athens free from any association with Sparta. Cimon narrowly won the vote and led a strong hoplite force to Sparta, but the Spartans found the attitude of the Athenians suspicious and dismissed them, alone of their allies.

Revolt of the helots and the dismissal of the Athenian force by Sparta

Whatever reasons the Spartans may have had for dismissing the Athenians, the results of their action were far-reaching. Cimon was discredited and in 461 he was ostracized. His main opponent, Ephialtes, was able to carry a series of laws that vitally changed the character of the Athenian democracy. The aspect of these reforms that is most emphasized in tradition is the curtailment of the power of the Areopagus, or Areopagite council, which the archons joined at the end of their year of office, but neither Aristotle nor any of his successors was able to define specifically what powers had been lost. In the 6th century the Areopagus had been the main guardian of the constitution, but the reforms of Cleisthenes shifted the balance of power to the Council of Five Hundred and the Assembly, and when, in 487, the lot was introduced in the election of archons, the prestige of the Areopagus probably declined. The Areopagus must previously have been exercising powers that were considered undemocratic. The definition of the very limited cases, all of a religious nature, that they were henceforward entitled to try implies that they had been acting more widely as a judicial court. The emphasis in the new, more developed democracy of Athens on the people's control of the executive suggests that magistrates had previously been influenced by the Areopagus. New powers were now explicitly given to the people. The jury courts were reorganized on a popular basis, and court decisions no longer rested with the magistrates but with the jurors, who were paid for their services, a principle that was later applied to the Council of Five Hundred and to all offices of state; no citizen was to be barred from state service by poverty. All officials had to undergo an examination in a people's court at the end of their year of office, and public auditors were annually appointed to examine accounts. It is significant that from the 450s important decrees of the Assembly and the summary accounts of public works were normally inscribed on stone and set up on the Acropolis or in some other public place where all could see them.

The dismissal of Cimon's force of Athenians by the Spartans led to a reversal of foreign policy. The immediate Athenian reaction was to conclude an alliance with Argos, Sparta's traditional enemy, and with Thessaly, with which Athens had been allied in the 6th century. A dangerous situation became more tense when Megara, a member of the Peloponnesian League, under pressure from its neighbour, Corinth, was accepted into alliance by Athens.

Developments in the west. The 460s were a period of change in Sicily also. For two generations Sicilian history had been dominated by the tyrants of Gela, Acragas, and Syracuse—their cities had enjoyed a period of great prosperity, which was based primarily on military strength. The tyrant Gelon, who came to power in 491, had succeeded in making Syracuse the strongest and richest city in Sicily. A network of marriage alliances linked him with Theron, who had become tyrant of Acragas in *c.* 488, and he proceeded to build up an army and navy to extend his power. He built a wall to defend the extension of Syracuse on the mainland and increased the population at the expense of smaller cities.

Syracuse under the tyrant Gelon

With the exception of a northern pocket around Mes-

sana, Gelon came to control the east of Sicily, and by 480 his army was sufficiently powerful and strong to prove decisive at the Battle of Himera when, with his ally Theron of Acragas, he defeated a force of Carthaginian invaders led by Hamilcar. The rich booty from this victory enabled the two cities to build impressive new temples, and the magnificence of the two tyrants' courts became a byword throughout the Greek world; it was reflected in their chariot victories at Olympia and in their dedications at Olympia and Delphi. Gelon died two years after the great victory. Hieron succeeded to the tyranny of Syracuse, being replaced at Gela by the third of the brothers, Polyzelus. It is probable that Hieron was less popular than Gelon and more oppressive in his rule. In his relations with the other Greek cities of east Sicily he was as ruthless as his brother. He drove out the populations of Naxos and Catana and forced them to settle in Leontini, and near Catana he established his own new city of Aetna with settlers from the Peloponnese and Syracuse. His relations with his brother Polyzelus, however, became strained, and when Polyzelus took refuge with Theron at Acragas, the two allies came near to war. When Theron died in 472 his son Thrasydaeus seems to have thought that the best way to counter internal opposition was by war. He assembled a large army, but Hieron struck first, and in a hard-fought battle, in which 4,000 men of Acragas and 2,000 Syracusans are said to have been killed, Thrasydaeus was defeated and was forced to leave Acragas. He took refuge in mainland Megara but was there condemned to death. Hieron made peace and, though Acragas was to make another challenge in 446, Syracuse never lost the primacy in Sicily that Gelon had established.

The tyrant Hieron as patron of the arts

Hieron was not merely a man of war. Greek literature benefitted greatly from his generosity and taste. Poets need patronage, and Hieron followed the precedent of earlier tyrants, the Cypselids at Corinth, the Peisistratids at Athens, and Polycrates at Samos, on a more magnificent scale. In the desolation following the Persian sack, Athens was in no position to attract a literary circle. Pindar was glad to visit Hieron's court and write odes to celebrate his victories at Olympia. Simonides of Ceos, who had commemorated the victories over Persia, came to Syracuse in 476 and stayed there as a court poet; Bacchylides, his nephew, joined him. Aeschylus was also attracted and gave a special production of his *Persae* in the theatre that Hieron had built; his *Aetnaeae* was no doubt intended to commemorate Hieron's new foundation of Aetna.

Hieron died in 467 at his own Aetna. The youngest of the four brothers, Thrasybulus, who succeeded him, had neither his ability nor the support to maintain his position, and was before long forced to withdraw to Locri.

The Syracusans celebrated the end of their tyranny by establishing an annual Freedom Festival and setting up a colossal statue of Zeus the Liberator. There seems to have been no serious opposition to democracy, and a democratic crusade was led against the few remaining tyrants in the island. The fighting in Syracuse had repercussions in many other cities, but finally a general agreement was reached that those who had been uprooted by the tyrants should be restored to their own cities and that the mercenaries were to settle in Messana.

Of relations between the Greek cities of south Italy very little is known. After the destruction of Sybaris in 510, Locri was the strongest of the states, but the archaeological evidence suggests a fairly widespread prosperity. The descendants of the refugees from Sybaris attempted near the mid-century to resettle their old home, but they lacked the strength to hold their own and appealed to the mainland for help. The Athenian settlers who came to join them became a source of friction rather than of strength, and the result was the planting by Athens of a new Panhellenic colony on a new site at Thurii, in 443.

In the first half of the 5th century there was a considerable loss of life in both Sicily and Italy in battles between Greeks and between Greeks and barbarians and, in Sicily, great uprootings of populations; but in this same period the achievement of the western Greeks in peaceful fields was very remarkable and comparable with that of the east Greeks in the 6th century. Following the incorporation of the eastern Greeks in the Persian Empire, refugees from Ionia probably played an important part in stimulating the western Greeks. When the Persians came, a large body of Phocaeans sailed to Corsica and, when driven out, settled at Elea (Velia) on the west coast of Italy. Xenophanes of Colophon, a vigorous free-thinker, after much wandering, also settled in Italy. A little later Pythagoras left Samos for Croton, where his compound of philosophy and religion made a deep impression.

Pythagoras and Parmenides

During this period, the west was the leader in intellectual speculation. Rival schools flourished. Pythagoras was a considerable mathematician and seems to have thought that numbers were the ultimate reality, but he was also strongly influenced by mystery religions. It is from this source that he based his belief in transmigration of souls, one of the few doctrines that can be certainly attributed to him.

His followers, observing a rule of life that combined principles drawn from philosophy with a religious mysticism, exercised a strong influence on politics. Their stern morality was effective at first, especially at Croton, where Pythagoreanism had evolved toward the end of the 6th century. The movement spread to Metapontum and other cities, but by the middle of the 5th century the emphasis on moral reform savoured too much of oligarchy, and there was a wholesale reaction against Pythagoreans.

Parmenides of Elea was more important in the history of philosophy. The Milesian speculators had tried to explain the nature of the physical universe, but Parmenides turned from physics to metaphysics. The basic question that puzzled him was what lies behind what one says or thinks: what is the corresponding reality? He dismissed sense perceptions and by involved but logical deductive arguments concluded that reality was a unity—indivisible and unchanging. He established a considerable following, among whom Zeno was his most persuasive supporter, and his arguments, in verse, greatly influenced Plato.

Empedocles of Acragas was a colourful contemporary who, like Parmenides, explained his views in verse. He rejected Parmenides' approach and defended the senses, explaining their distortions in physical terms. He was closer to the Milesians; but instead of searching for a single substance to explain the world, he found four basic elements—earth, air, fire, and water—which, under the changing pressures of love and strife, took varying forms. But there was also a strain of mysticism in his teaching, for, like the Pythagoreans, he believed in the transmigration of souls and would not eat the flesh of any animal.

The poets of the 6th century BC, Stesichorus of Himera and Ibycus of Rhegium, had no distinguished successors, although Epicharmus established comedy in Syracuse independently of Athens. None of his plays survive, but in later tradition he was better known than his Athenian contemporaries. With democracy came an increase in the importance of public speaking; Corax and his pupil Tisias, both Syracusans, were regarded as the pioneers in the study and teaching of the art of rhetoric.

Visual arts of the western Greeks

Unlike the Greeks of the mainland, the western Greeks had no easily accessible marble suitable for building, but their shelly limestone is an attractive material in texture and colour, and their temples in a severe Doric style are impressive for their size and proportions; the temple of Olympian Zeus at Acragas was considerably larger than the Athenian Parthenon. The lack of marble helps to explain why little is heard of west Greek sculpture, and it is significant that Pythagoras (not to be identified with the philosopher), the only sculptor of the west who was later regarded as one of the great masters, came from Samos before settling at Rhegium. But there was one field in which the west Greeks were supreme. The coin engravers of Sicily deserve a place in any anthology of Greek art. There are few coins, ancient or modern, more beautiful than the Demareteion, a ten-drachma piece issued to commemorate the Battle of Himera and named for Gelon's wife.

The Greek legacy owes not a little to the west Greeks of the first half of the 5th century.

The first Peloponnesian War (460–445 BC). The ten years that followed the dismissal of the Athenians from Ithome by Sparta (462 BC) and the ostracism of Cimon witnessed a dramatic rise and fall in Athenian fortunes as the Athenians consolidated their democratic reforms and fought on two fronts—against Peloponnesians and Persians. In 460, 200 Athenian and allied ships sailed to Cyprus, which, since the expedition led by Pausanias in 478, had been partly recovered by the Persians. This fleet, however, was soon diverted to what seemed an even more promising opportunity. For Inaros, a Libyan prince who had raised a revolt in the Nile Delta soon after the accession of the new Persian king, appealed to Athens for support. The restoration of Egyptian independence would have been a more serious blow to Persia than the liberation of Cyprus. The rich grain resources of Egypt may also have influenced the Athenian decision to intervene. At first all went well. The Greeks defeated the small Persian fleet and won control of the Nile. The remaining Persian troops fell back on Memphis at the head of the Delta and were besieged; but, though the Greeks quickly occupied most of the town, a determined force of Persians held out stubbornly in the White Fort.

Meanwhile, hostilities had broken out in Greece. The brief surviving record does not explain the circumstances that led to the fighting. The narrative begins with an Athenian attempt to take Halieis, a port on the Peloponnesian coast beyond Troezen, which could have proved a useful naval base for the Athenian fleet for raids on the Peloponnese and for strengthening communication with Argos; but forces from Corinth and Epidaurus drove the Athenians off, though the attempt to destroy the fleet on its return journey resulted in an Athenian victory off the island of Cecryphaleia. At this point Athens' old rival Aegina determined to put out its full strength, but the Athenians, reinforced by league allies, won a decisive victory in a great naval battle. The Aeginetans were driven from the sea and, after a long siege, their city was captured. Aegina was forced to become a member of the Delian League and pay the high annual tribute of 30 talents, the highest of the assessments, shared only by Thasos. During the siege, Cornith had tried to relieve the pressure on Aegina by invading the Megarid but met with no success.

Direct conflict between Sparta and Athens (457 BC)

Hitherto the fighting had been between Athens and the members of the Peloponnesian League near the Isthmus. In 457 the Spartans intervened. Until 460 (and possibly longer) Sparta had been preoccupied with the resistance of the helots on Mt. Ithome, but they had now surrendered on condition that they could freely leave the Peloponnese. The Athenians gave them a home in Naupactus on the north shore of the Corinthian Gulf, which they had recently captured. In this strategic site, commanding the entrance to the gulf which Corinth regarded as its own, the helots were to repay abundantly the generosity of Athens. The desertion of Megara from the Peloponnesian League and its subsequent alliance with Athens had severely limited the range of Spartan action, for the Athenians had built a wall between Megara and Nisaea, effectively barring the only practical route for a Peloponnesian army invading Attica. The Athenians had also strengthened their position by building long walls from the city to the Piraeus and to Phaleron, within which the population of Attica could take refuge in the event of invasion. To demonstrate their power outside the Peloponnese, the Spartans took a strong force of some 11,500 Peloponnesians across the Corinthian Gulf. Their immediate objective was to defend the small state of Doris, a task that was soon accomplished. They now found that their way home was barred. The Athenians had sent a fleet into the Corinthian Gulf and had closed the mountain route over the Isthmus. The Spartans moved into Boeotia and encamped at Tanagra, threatening Attica and negotiating with extreme oligarchs in Athens, who were anxious to overthrow the democracy. The Athenians marched out, and the pitched battle that followed was evenly fought, but the Thessalian cavalry deserted and the Athenians had to admit defeat. The Spartans, however, did not feel strong enough to attack Athens itself and were content to return through the mountain passes of the Isthmus. Athenian morale was not badly shaken and two months later they returned under their general Myronides and won a decisive victory over the Boeotians at Oenophyta (457 BC). This victory gave them control over all Boeotia, with the possible exception of Thebes, and their control was soon extended over Phocis and Locris. They could now turn to the offensive against the Spartans; the Spartan dockyards at Gythium were burnt and Chalcis on the Gulf of Calydon was captured. This marked the high tide of Athenian success.

The tide first turned in the east. The Persian king had tried to relieve the pressure on Egypt by bribing the Spartans to act more vigorously against Athens. When this failed, a large army and navy were sent to crush the revolt. A squadron of 50 ships, sent out to relieve the Greek fleet or part of it, entered the Nile and was largely destroyed. The Greeks were driven from Memphis and fell back on Prosopitis, an island between two branches of the Nile. After 18 months the Persians dried up the canal linking the two branches, and the Greek position became untenable. Under an armistice the survivors were allowed to cross the western desert to Cyrene. The loss in men and ships was serious, and the effect on Athenian expansion decisive. During the early 450s Athens had aimed at bringing into the league Cyprus and all cities in the eastern Mediterranean that had Greek roots. The collapse in Egypt meant the abandonment of Cyprus and the eastern Mediterranean. In Greece itself the offensive against the Peloponnese was halted. Over the next three years the Athenians concentrated on building up their strength, but before a new expedition sailed east they wisely decided to secure themselves against Sparta. No one could have been a more acceptable mediator than Cimon, who had always advocated friendship with Sparta. Cimon was, therefore, specially recalled, shortly before his ten years' ostracism was due to expire, and he succeeded in negotiating a five years' truce early in 451.

Cimon could now resume command of league forces against Persia, and in 451 (or possibly 450) he sailed east with 200 ships. His main objective was the recovery of Cyprus, but he detached 60 ships to Egypt, where revolt was still smoldering. In Cyprus the campaign started well, but Cimon died while besieging Citium, the strongest Phoenician city in the island, and his colleagues in the command decided to take the fleet home. The Greek withdrawal from Cyprus marked the end of fighting between the Delian League and Persia. Both sides could now leave the struggle with some dignity, and both had good reason to do so. The Athenians realized that they had overstrained their resources and felt nervous for the stability of the league; the Persians had recovered control over the eastern Mediterranean and Egypt, although they were faced with a revolt in Syria. It is probable that a formal peace, the so-called Peace of Callais, was now made, in which the Athenians abandoned their ambitions in the eastern Mediterranean and the Persians agreed not to attack the members of the Delian League and to keep their fleet out of the Aegean; but there is no firm evidence, and the peace remains controversial. It is certain, however, that Athens decided to terminate hostilities.

Cimon's expedition to Cyprus

The years that followed the end of fighting against Persia were the most critical period for Athens, both at home and in relations with the allies, since the invasion of Xerxes. The most important problem was the future of the Delian League now that Persia could no longer be regarded as an active enemy. Athens was determined to maintain the league and did not hesitate to make this clear to the members. The Athenians' relations with the allies had, by 465, changed considerably, and they had used force to put down revolts. In the 450s, some of the cities were called upon to send ships or troops to fight against the Peloponnesians—something that had certainly not been anticipated earlier. Athens removed the league treasury from Delos to Athens. Henceforth the annual tribute was to be brought to Athens in early spring at the time of the Dionysia. The transfer of the treasury could be justified when it was made, for the Persian victory in Egypt might have encouraged the Phoenician fleet

to sail into the Aegean and raid Delos. But however honest the motive may have been, the transfer made it much easier for the Athenians to use the money for their own purposes.

After the death of Cimon, Thucydides, son of Melesias (not the historian), and probably Cimon's brother-in-law, set out to organize into a coherent party the extreme oligarchs and all those who might have an interest in reversing the tendencies of the democratic leadership. When the Assembly decided to use the tribute from the allies to build Athena's temple, Thucydides tried to persuade the people to repudiate the immoral use of the allies' money:

> Greece appears to be the victim of monstrous violence and manifest tyranny when she sees that with the money contributed under compulsion for the war we are gilding our city like a wanton woman, adorning her with extravagant stones, statues and thousand-talent temples.

But the pride of a new temple and the prospect of employment that it offered were sufficient to dull any tender consciences. Thucydides continued to attack his democratic opponents, but Athenian policies were not significantly modified.

Difficulties within the Delian League

Athens also had trouble from the allies. The tribute lists of the years from 450 to 446 show that this was an unstable period. Many cities paid late in the year; others made incomplete payments and some, including probably Aegina, did not pay at all. It was probably in this period that Colophon revolted and was suppressed. There had been trouble in Erythrae and Miletus before 450, and the Milesian oligarchs, with whom the Athenians at first tried to cooperate, now carried out a purge of their opposition. But by 440 democrats were in control of Miletus; they had almost certainly been installed in power by the Athenians.

Two important decrees that illustrate the new imperial tone are probably to be dated in these years. The first requires all the allies to use Athenian coins, weights, and measures and bans all minting of silver: now that the tribute was to be spent mainly in Athens, crews, contractors, and workers would naturally want to be paid in Athenian coin. But the move was also an open expression of empire. Another attempt to bind the cities together in an empire is alluded to in a second decree, the main purpose of which was to tighten up the collection of tribute but which refers in passing to a more important decree, recently passed, requiring all the allies to bring to the festival of the Great Panathenaea—held every four years—a cow and a suit of armour: colonies were traditionally expected to recognize in such ways the main festivals of their mother-cities. Athens also hoped to encourage loyalty to the empire by spreading the cult of "Athena, queen of Athens" in the cities.

Another important feature of this period was the settlement overseas of Athenian citizens of the two poorest classes. This policy of establishing what were called "cleruchies" in the territories of allied cities was one of the main grievances still remembered against Athens in the 4th century, although not all such settlements were unpopular. Some cleruchies were intended as garrisons and settled as penalties for anti-Athenian attitudes or actions.

Periclean democracy and imperialism. By the various measures taken between 450 and 446, Athens had succeeded in converting a league into an empire. The opposition in the Assembly led by Thucydides had been active but ineffective, and allied discontent had been met by firmness. There was no attempt to disguise the change of status, and the Athenians freely admitted that they no longer led autonomous allies. The strength of the new Athenian position was fully confirmed when Athens successfully weathered a succession of serious crises in *c.* 446, the year when the five years' truce with Sparta was due to expire. In spring 446 or summer 447 the Athenian general Tolmides underestimated the seriousness of a rising in Boeotia. He was surprised and defeated near Coronea, and to save the army the Athenians had to undertake to withdraw entirely from Boeotia, undoing all that had been won at Oenophyta. The Boeotians had been helped by Euboean exiles and revolt now spread to the main cities of Euboea. Pericles, now the leading Athenian politician and general, had taken an army over to the island when he learned that the Megarians had massacred most of their Athenian garrison and that a Peloponnesian army was marching toward Attica. He brought his army back and negotiated a Spartan withdrawal (allegedly by bribery). He could then return to Euboea, where the rebel cities were quickly reduced.

The Thirty Years' Peace

Meanwhile, an Athenian embassy to Sparta had negotiated the terms of a peace which was to last 30 years. The Athenians gave up Nisaea, Troezen, and Achaea. Aegina had also hoped to regain its independence but had to be satisfied with a clause guaranteeing freedom from Athenian interference provided that an annual tribute was paid. Athens' most important gain was the recognition of its empire by Sparta. Both sides undertook not to interfere with the alliance of the other, though neutrals could join either.

During the peace negotiations the Athenians had to draw up new tribute assessments. Some 30 cities (and probably more, for the evidence is not complete) had their assessments reduced, while only three increases are known; the natural inference is that Athens was reacting to the crises from which it had emerged by a policy of concessions. There is no firm evidence of the attitude taken by Thucydides and his followers to the events of these critical years, but in 443 there was sufficient feeling against him to secure his ostracism. With his fall, the party he had built up collapsed. The extreme oligarchs remained implacably hostile to the democracy but were not strong enough to come out into the open; most of the moderates were content to accept the leadership of Pericles, who, in spite of his political views, was an aristocrat.

Pericles, early in his political career, had been personally responsible for the introduction of state pay for state service, first to the jurors and later to the Boule (city council) and all branches of the executive. By the early 440s he was the most influential politician in Athens. He was the main architect of the conversion of league into empire and was almost certainly mainly responsible for ending the fighting against Persia.

Periclean building in Athens

With his firm belief in the need for maintaining a tribute-paying empire, Pericles combined imaginative ideas on the quality of Athenian democracy and culture. He began the transformation of Athenian public architecture, and it was he who formally proposed that the tribute reserve should be used for the rebuilding of Athenian temples. The public buildings of the Cimonian period had been in limestone; the Periclean temples were in marble from the quarries of Mt. Pentelicus. The most impressive of the new buildings were the Parthenon, with its splendid gold and ivory cult statue of Athena, and the new monumental western entrance to the Acropolis. Other fine temples were built in the same generation, among them, one to Hephaestus overlooking the Agora, to Poseidon on the headland of Sunium, and to Nemesis at Rhamnous. The program was not confined to religious buildings. A large concert hall, with its tentlike roof supported by a forest of columns, was built near the theatre on the southeast slope of the Acropolis, and to make the amenities of life available to poor at well as rich, public money was used to build public gymnasia and baths.

Pericles was said to have been a pupil of Anaxagoras, a philosopher who spent the greater part of his active life in Athens. To him he owed his rigorous grounding in the study of the physical universe and a lasting interest in philosophy; together they freed him from the restricting conventions of contemporary religion and superstition.

Pericles was reserved and did not mix freely with the people; but he was a commanding orator, and the Assembly learned to accept his judgment and follow his lead. After the ostracism of Thucydides he was elected general for 15 years in succession. But the real source of his power lay more in his personal authority. The historian Thucydides, in recording the funeral speech he delivered in 431 to commemorate those who died in the first year of the Peloponnesian War, outlines the main qualities of the democracy which Pericles did so much to create. Un-

Nature of Periclean democracy

like Sparta, inward-looking and strictly disciplined, Athens was a city wide open to the world. Pericles had himself stiffened the qualifications for citizenship by requiring Athenian parentage on both sides, but foreigners were encouraged to settle in the Piraeus or Athens itself and were better treated and had more profitable opportunities there than in any other Greek city. Athens welcomed them, unskilled as well as skilled, and they played a dominant part in Athenian trade. In the Periclean democracy, public service was not confined to a minority of activists but was the natural duty of all citizens, poor as well as rich, and most offices were filled by the lot. Five hundred men had to be found each year for the Boule, and, as no one could serve more than twice nor be under the age of 30, a large proportion of the citizen body must have served at least once. The Boule reviewed all state business and controlled the executive; and no business could be discussed in the Assembly unless it had been put on the agenda by the Boule. A year in the Boule was an intensive political experience, and for a tenth of the year every member had to be continuously available. The sovereign body was the Assembly, in which every citizen had an equal vote. Policies were normally discussed on the basis of resolutions drafted in the Boule (*probouleumata*), but it was also open to the Boule to make no formal recommendation and merely to bring a matter forward for debate. In either case there was a free vote; anyone could address the meeting; and amendments could be proposed from the floor. Perhaps the most surprising feature of the proceedings of the Assembly was the appointment by lot of the chairman, who changed daily and was chosen from the members of the Boule. That important policies could be discussed at large open-air meetings—at some of which a quorum of 6,000 was required—and result in rational decrees is a great tribute to the tolerance of the average citizen. Thucydides, looking back on what he regarded as the golden age of Pericles said that "it was in name a democracy, but in reality the rule of the first man." This should not obscure the fact that Pericles had to persuade the Assembly and that the Assembly was not an ignorant mob; he could not act arbitrarily without authority. For his actions in office he had to submit to the same public examination as others, and at any time his proposals could be rejected by vote.

Importance of the empire to Athenian prosperity

The high standard of living that the Athenian democracy provided for its citizens depended in no small part on the stability of the empire, the tribute of which paid directly or indirectly for much of Athenian public spending. Without the tribute reserve, Pericles' ambitious building program would have been out of the question, and without the annual tribute, Athens would have been unable to afford both state pay for state service and a large and active fleet. There is, however, no evidence that the allies were financially oppressed. There was no general increase in assessments before the Peloponnesian War, and more than half the cities kept the same assessment from 453 to 431 BC. For most cities the tribute was probably a small price to pay for the policing of the seas by the Athenian fleet. Too little is known of the pattern and volume of trading during this period to draw firm conclusions, but it is clear that the Piraeus became the busiest trading port in the Aegean and probably in the Mediterranean. A wide range of goods came in from the Euxine, Phoenicia, Egypt, and Carthage, as well as from the Greek world. There seems, indeed, to have been a considerable growth of trade following the end of hostilities against Persia, and, though Athens was the main beneficiary, the security of shipping benefitted the allies as well. The grievances of the allies were political rather than economic. They had lost their independence and were no longer allies but subjects.

In order to control subject cities, Athenian officials, *archontes*, were installed, sometimes singly, sometimes as many as five; with them went a small security guard, stationed on the local acropolis, capable of dealing with small-scale trouble but not strong enough to crush a serious revolt. Athens also sent out commissioners, *episkopoi*, to cities where trouble had arisen to examine the situation, make recommendations on the spot, and report back. Such officials were widespread by the early 440s. More important was the control exercised by the people's law courts at Athens. All cases involving the major penalties of death, exile, loss of citizen rights, and total confiscation of property had to be referred to Athens. But the most reliable means of control was the encouragement or imposition of democratic governments. It was the oligarchs who most resented Athenian rule; the burden of tribute fell mainly on the wealthy, and the powers of local dynasts were restricted. Most of the revolts for which there is evidence were engineered by oligarchs, and in the settlements that followed power was normally given to the demos. It was inevitable that oligarchs should look elsewhere for support, to Persia or Sparta, and when it came to a choice between rule by their own extreme oligarchs or Athens, the common people would tend to prefer Athenian control.

The revolt of Samos

After the Thirty Years' Peace had been negotiated, the Athenian Empire seemed secure. It was a considerable shock when, in 440, Samos, one of the three islands off the Asiatic shore that still contributed ships, revolted. Fighting had broken out between Samos, which at the time was controlled by oligarchs, and Miletus, where Athens had recently installed a democracy. Priene, which at various times had needed to fight against both for its independence, was the bone of contention. When an Athenian attempt at a political settlement failed, war followed, and the Athenians had to deploy their full naval resources and summon contingents from Chios and Lesbos. After a nine-month siege, Samos had to surrender and was made to give up its fleet, raze its walls, and agree to pay a substantial indemnity in installments. Samos had appealed for help to Sparta, and the Spartans went so far as to summon a meeting of the Peloponnesian League; but the peace had stipulated that neither side should interfere with the other's alliance, and the league voted against war. The Samian oligarchs had also appealed to the Persian satrap at Sardis and had received some mercenaries from him. Of the Athenian allies, Byzantium alone followed Samos but was recovered without serious fighting.

THE GREAT PELOPONNESIAN WAR (431–404 BC)

Five years after the surrender of Samos a train of events started in the northwest that led to the Great Peloponnesian War. Though the record is thin, it seems that in these years Athens was in an expansive mood, a mood likely to cause misgivings in Sparta. In 436 Athens succeeded in planting a large colony at Ennea Hodoi, now to be renamed Amphipolis, on the Strymon, a splendid base for the exploitation of the rich mineral and timber resources of the region. But Athens could not, without weakening its manpower dangerously, provide more than a nucleus of the colonists; the mixed character of the population, drawn largely from the Greek colonies of the area, was to prove a serious weakness. A little later Pericles himself led a spectacular expedition into the Euxine to demonstrate Athenian strength and their interest in the Greek cities of the south as well as the north coasts. At Sinope the local tyrant was expelled, and 600 Athenians were sent out to strengthen the Greek population.

Athenian intervention between Corinth and Corcyra

By 435 Athenian restlessness was causing general concern, and when fighting broke out between Corinth and its old colony Corcyra there was soon a danger that it would spread much further. For Corinth, humiliated by defeat in a naval battle, began to make intensive preparations for a larger fleet to teach Corcyra a lesson, and Corcyra appealed to Athens. Since Corcyra belonged to neither alliance, the Athenians would not be breaking the terms of the peace if they accepted Corcyra as an ally, although such an action would clearly invite the bitter resentment of Corinth. After an even debate in the Assembly, the Athenians decided to make a defensive alliance: Athens would help Corcyra if it was attacked and was in danger of defeat. Ten ships, shortly followed by a second squadron of 20, were sent out. In the Battle of Sybota that followed, the Athenian ships intervened when the Corcyraeans were in grave danger; the Corinthians withdrew. More bitterness was caused when the Athe-

nians, suspecting trouble from Potidaea in Chalcidice, a Corinthian colony but a member of the Athenian Empire, required the city to pull down its main defensive wall and give hostages. The Potidaeans, after failing to negotiate concessions from Athens, and relying on a Spartan undertaking to invade Attica if Potidaea was attacked, came out in open revolt. It was probably soon after this development that the Athenians charged Megara with cultivating borderland that belonged to the Eleusinian goddesses and with giving a refuge to runaway Athenian slaves. When no satisfaction was given, Pericles carried a decree banning the Megarians from the Athenian Agora (civic centre and market) and from all the ports of the Athenian Empire.

In 432, while an Athenian force was blockading Potidaea, the Corinthians and other members of the Peloponnesian League sent representatives to Sparta to press for action. Sparta was divided; King Archidamus pleaded for caution and delay, but a clear majority preferred the rousing demand for war from the ephor Sthenelaidas. To involve its allies Sparta needed a majority vote at a formal meeting of the Peloponnesian League. In the late summer of 432 the league was committed to war.

Thucydides' view of the war's origin

In his detailed account of these events Thucydides emphasizes the initiative of Corinth in bringing matters to a head, but he saw clearly that the cause of the war went far deeper than the trading interests of Corinth. "The true cause of the war . . . was the growth of Athenian power which compelled Sparta to fight." Other writers laid the main stress on the Megarian decree and on the refusal of Pericles to compromise. To Thucydides the real issue was not the trade rivalry between Athens and Corinth but whether Sparta or Athens should be the leading power in Greece. In the war, which opened in 431, the Spartans were technically the aggressors. They sent ultimatums to Athens over the winter, the terms of which were such that they cannot have expected Athens to accept them. Peace, however, could have been maintained if the Athenians had been more conciliatory. But Pericles had long since decided that Athens could not reach its full development while Sparta controlled the Peloponnese. His vision of Athens included the eclipse of Sparta.

Pericles led Athens into the war deliberately and with full confidence. In spite of the ambitious building program there was still a reserve of some 6,000 talents on the Acropolis which, when combined with annual income, seemed ample to sustain even a long war. The Athenian fleet, with 300 ships, was immensely superior to that of the Peloponnesians in numbers, experience, and skill. The Athenian fleet, Pericles thought, could maintain a firm hold on the empire and attack the enemy's coastline. The Athenian army was weaker in numbers and military reputation than that of the Peloponnesians, but it was sufficient to man the defenses and raid the Megarid. Pericles argued that the army should not be used to fight the Spartans and should avoid pitched battle; Attica must be abandoned and the country population take refuge in Athens, where they would be protected by the city walls and the long walls running to Piraeus. There should be no attempt to expand the empire during the war. It is difficult to see how Sparta could have been crushed by Pericles' strategy, but under continuous pressure the Peloponnesian League might have disintegrated, forcing Sparta to accept an unequal peace, virtually admitting Athenian primacy.

The great weaknesses on the Spartan side were their fleet and their finances. The Spartans still dealt in iron bars for currency, and most of their allies were agricultural states with lean resources. The coastal towns, particularly Corinth, were economically stronger, but they could not afford to raise and pay large fleets. The Peloponnesian army, however, could easily muster 30,000 (against an Athenian limit of 13,000).

The pattern of warfare

The first year of the war established the pattern that was to be expected. The Peloponnesian army, led by King Archidamus, invaded and ravaged Attica, while the Athenians watched helplessly from their walls. Meanwhile, a fleet of 100 Athenian ships sailed around the Peloponnese, raiding selected points. In 430 the strategies were repeated, but the most important event of the year was the arrival in Athens of a plague from the east that swept through the overcrowded city, killing roughly a quarter of the fighting men and an even larger proportion of women, children, and slaves. This had a depressing effect on Athenian morale; an embassy was sent to Sparta to negotiate a settlement and Pericles was dismissed from his command. Sparta, however, would not offer acceptable terms; the Athenian mood recovered, and Pericles was shortly re-elected. But in 429 he died.

The demagogic element in Athens. Pericles' death was followed by a change in the political climate. Among those who had been generals with Pericles there was no outstanding politician, and, as a consequence, policies became less consistent. The most disturbing factor was the emergence of new political leaders who no longer depended for their influence on their prestige as generals or on the distinction of their families but on their shrewd understanding of the common people and their debating skill in the Assembly. Cleon was the first of these new leaders. He first appears in the record in 431, attacking Pericles when the latter would not allow the Athenians to go out and fight. Like his successors, he rose to power as a critic, prosecuting oligarchs in the courts, seizing any opportunities that were offered when magistrates were publicly examined at the end of their year of office, and claiming to be the champion of the common man. Aristophanes caricatures him as a leather seller, as if he were some small-scale retail trader, but Cleon's father had supplied a chorus for a public festival and must therefore have been a man of more than average means. By his background he was better qualified than most of his political rivals to handle financial problems.

The rise of Cleon

Neither Thucydides nor Aristophanes was sympathetic to the new style, but even Thucydides admits that by 427 Cleon was "by far the most persuasive speaker in the Assembly." Thucydides also describes him as "most violent," and from Aristophanes it is known that he had a loud voice "like a mountain torrent." His qualities were demonstrated when Mytilene, which had been in revolt, surrendered in 427 and the Assembly had to decide what penalties to impose. Cleon convinced the people that the right policy was to put all adult males to death and enslave the women and children. At a second meeting, called to reconsider the decision, he stood firm by his original motion and was only narrowly outvoted.

Cleon's influence became even stronger in 425. An Athenian fleet sailing to Sicily had left a force on the promontory of Pylos in Messenia, hoping to encourage Spartan helots to desert. The Spartans, in alarm, brought back an invading army from Attica; in their operations against Pylos they occupied the neighbouring off-shore island of Sphacteria in order to deny it to the Athenians. But Pylos held out, and the Athenian fleet blockaded Sphacteria. When the blockade proved ineffective Cleon violently attacked the generals in the Assembly, demanding more vigorous action, with the surprising result that he was himself given the command. He took light-armed troops and used the advice of Demosthenes, the most enterprising of the generals. They landed a force on the island and finally forced the Spartans to surrender. Now Athens had hostages against further invasions of Attica, and Cleon's followers could introduce an extraordinary assessment in which the tributes of the allies were sharply increased to bring in a total of nearly 1,500 talents, more than three times the prewar figure.

The conduct of the war now became more vigorous. A raid was made on Corinthian territory, and in the summer of 424 the island of Cythera off the south coast of Laconia was captured, providing a useful base for raids on the mainland and for intercepting merchantmen coming to the Peloponnese from Libya and Egypt. In 424, however, the Athenians over-reached themselves. Cleon had been elected general, but it was Demosthenes and Hippocrates who planned the campaigns. Their first objective was to gain control of Megara by a combination of military action and political propaganda, but in this they failed. Demosthenes and Hippocrates then proceeded with an elaborate attempt to regain control of

Boeotia. There were to be democratic risings in the cities; Demosthenes was to land with a force at Siphae on the Corinthian Gulf, while Hippocrates led out the main Athenian army and occupied Delium. The plan was too complicated and could not be kept secret. Siphae was garrisoned when Demosthenes arrived; the revolts in the cities misfired; the army from Athens was isolated and, in a pitched battle near Delium, was decisively defeated. At roughly the same time, the Spartan Brasidas and a small force had reached Chalcidice and was attempting to undermine Athenian control of the area. Brasidas met with rather less success than might have been expected, but he succeeded in his most important objective by capturing Amphipolis, a very serious loss to Athens.

The Peace of Nicias. Athenian prospects had now completely lost the promise that the victory at Pylos had brought, but the Spartans were more anxious to recover their prisoners than to continue the fighting in the north. A year's armistice was therefore arranged from April 423; it was intended to pave the way to a permanent peace. But within a few days of the armistice Mende and Scione revolted, and Brasidas supported them. Mende was quickly recovered, but Scione held out into 421. In 422 Cleon was given command in the hope that he would repeat his success of 425; he did indeed start well, winning back Torone and some smaller places, but at Amphipolis he failed. He led out his force to high ground beyond the town in order to reconnoitre, and was caught by a surprise attack while returning. His army was beaten and he himself killed, but Brasidas also fell in the battle. Their deaths made it easier for the Spartans and Athenians to negotiate seriously for peace.

Terms of the peace

The main terms agreed between the two sides were that each should return its prisoners and that Sparta should recover Pylos, Athens Amphipolis. No provision was made for the restoration of Sollium and Anactorium, which the Athenians had taken from Corinth, nor of Nisaea, which they had occupied in 424 when they failed to win Megara. Boeotia was required to hand back Panactum, an Athenian border fort captured after the Battle of Delium. The Spartans had ignored the interests of their strongest allies and, when the terms were submitted to the Peloponnesian League for formal ratification, Corinth, Megara, Elis, and Boeotia all opposed acceptance; but they were outvoted by the smaller states, which had nothing to gain from war; nor were their interests seriously affected by Athenian imperialism.

The peace was named after Nicias, the most influential of the Periclean politicians since the death of Pericles, and the main enemy of Cleon. Nicias lacked Pericles' authority, however, and, though popular, was indecisive and ill-suited to lead Athens firmly to friendly coexistence with Sparta. His peace was doomed from the outset because the war had settled nothing, and too many cities were dissatisfied with the terms—especially Corinth, which had lost most and gained nothing. Internally, neither Sparta nor Athens was stable. When new ephors were appointed in the autumn at Sparta the board of five included two men who were anxious to renew the war. In Athens, Nicias had two powerful opponents who were not prepared to accept his policy. One, Hyperbolus, had inherited the mantle of Cleon; he had the same background, relied on the same following in the common people, and advocated similar policies. Cleon's fortune came from leather, Hyperbolus' from lamps; in Aristophanes' plays they are both warmongers. More dangerous was the young Alcibiades, perhaps the most brilliant Athenian of his generation. His father had died at the Battle of Coronea in 446 when he was only five years old, and he had been brought up in the house of his guardian Pericles. Strikingly handsome, with a quick intelligence and a cheerful contempt for convention, he fascinated women and could lead men. He combined great abilities with a fundamental lack of responsibility and, as Thucydides saw, his private life ruined his public career. When the peace was made he looked for opportunities to discredit Nicias and win the kind of position that Pericles had held. In so doing he roused the bitter hostility of Hyperbolus and his associates because his colourful conduct appealed to the common people. So developed a feud that was to be one of the main causes of Athens' ruin.

Breakdown of the peace: Athens' Peloponnesian alliances

The failure of the Peace of Nicias was soon manifest. Prisoners were exchanged, but Amphipolis was not handed back to Athens because the inhabitants were determined to remain independent, and so Athens continued to hold Pylos. In the uneasy years that followed there was a confusing pattern of negotiations and alliances. First Corinth attempted to build up a third power, winning over Argos, whose Thirty Years' Peace with Sparta was due to expire toward the end of 421, and some of the members of the Peloponnesian League; but an alliance between an oligarchic Corinth and a democratic Argos was bound to be insecure, and Corinth lacked the power and confidence to hold the alliance together. At the time, however, it seemed sufficiently dangerous to persuade Athens and Sparta to join forces in a new alliance to protect themselves. Before the end of 420, however, the influence of Nicias had been seriously eroded and Alcibiades was able to lead the Assembly away from Sparta into an alliance with Argos and two members of the Peloponnesian League that resented Sparta's high-handedness—Mantineia and Elis. This alliance was put to the test in 418 when the two armies faced one another outside Mantineia and Sparta was cut off from Corinth and Boeotia. The battle was hard fought, and at one point a picked corps of Argives broke through the Spartan line; but Spartan discipline finally prevailed, though the defeat did not become a rout.

The combination of Athens and Argos with discontented Peloponnesians near Sparta offered the only practical prospect of defeating Sparta decisively. But the policy failed because Athens' support was halfhearted. Its contribution to the allies' army—only 1,000 hoplites and 300 cavalry—was inadequate, and it was led by Laches and Nicostratus, who had been consistently associated with Nicias; Alcibiades, who originated the policy, had not even been elected general for the year. Elis also was to blame; when it disagreed with the priorities for the campaign, it withdrew 3,000 men.

Having failed in its Peloponnesian policy, Athens turned elsewhere. In 416 it sent a strong force to bring the island of Melos into the empire. Melos had been neutral at the beginning of the war and had resisted when Nicias landed in 426 with a strong force and ravaged the land. Though the Athenians now promised that there would be no victimization if Melos would join the other islands as a tribute-paying member of the Athenian Empire, the Melians stubbornly refused. After a stiff resistance they were forced to surrender; the men of military age were put to death, and the women and children enslaved. This was the fate that Cleon had advocated for Mytilene and had succeeded in imposing on Scione. But these two cities had revolted; Melos had merely tried to remain independent. From a playful allusion to the "Melian famine" in Aristophanes' *Birds*, produced in 414, it seems that the savage punishment of Melos did not weigh heavily on the Athenian conscience at the time. But Thucydides emphasized its significance, and it was widely remembered against Athens in the 4th century.

The expedition to Sicily and its aftermath. In 415 the Athenians embarked on an ambitious attempt to gain control of Sicily. The occasion was an appeal for help from Segesta in the west of the island, which was under pressure from neighbouring Selinus. Alcibiades saw in a Sicilian expedition the opportunity he needed for spectacular successes, and the people were attracted by the wealth of Sicily. Nicias pleaded for caution but was outvoted. In spite of his strong opposition he was elected to a joint command with Lamachus and Alcibiades. It was a bad compromise. The expedition was the most impressive that had left the Piraeus since the middle of the century. There were nearly 100 battle triremes, including 34 from the allies, with 40 more used as transports, and 5,100 hoplites, of whom only 1,700 were Athenians.

But the expedition sailed under a cloud. One night, shortly before it was due to depart, most of the figures of Hermes in the city were defaced. The outrage was widely regarded as a sinister omen, and rewards were

offered to informers who brought evidence about the Hermae or any other acts of sacrilege. Nothing was immediately discovered about the Hermae, but reports were brought about other offenses, including parodies of the Eleusinian Mysteries in private houses, in which Alcibiades was implicated. This opportunity was eagerly seized by Androcles and other political enemies of Alcibiades, who roused popular feeling against him. Alcibiades urged that his case should be tried at once, but his enemies, who were afraid that he would have the support of the expeditionary force, insisted that the case should be postponed until further evidence was collected. Soon after operations had begun in Sicily Alcibiades was summoned home. Realizing that the scales were weighted against him, he gave his escort the slip and soon was in Sparta putting his knowledge and advice at its disposal.

Defacement of the Hermae and the defection of Alcibiades

It was this action against Alcibiades that was largely responsible for the Athenian disaster in Sicily. Morale was badly shaken by the recall of the driving force behind the expedition, and not long afterward Lamachus was killed in battle. Nicias, whose indecisiveness was made worse by his poor health, was left in control. Progress had at first been promising, although too slow. The Athenian fleet controlled the sea, and a wall around the city was begun in order to complete the blockade. But Gylippus, a vigorous Spartan general, brought new confidence to the defenders, and Corinth also sent help. The circumvallation wall was cut off and the Athenians fell back on the defensive. Nicias, frightened to return home unsuccessful, sent for reinforcements, which arrived under Demosthenes. He made a bold night attack on Epipolae, the rocky plateau overlooking Syracuse; but his units became confused in the dark, and the attack had to be called off. Demosthenes was now anxious to leave; but Nicias hesitated, and an eclipse of the moon persuaded him to wait further. A great naval battle in the harbour was lost, and the only hope was to lead the forces overland to friendly country. But the roads were blocked, and after grim harrassment the survivors of both divisions surrendered. Nicias and Demosthenes were put to death and the men were imprisoned in quarries from which some were released for reciting verses from Euripides; others were ransomed, but many died.

Consequence of the defeat in Sicily

The almost complete destruction of the fleet and army that went to Sicily was a crippling blow to Athens. Few ships and little money were left. Urged on by Alcibiades, Agis, the Spartan king, had brought a force into Attica and was encamped at Decelea, only 12 miles from Athens itself. As a result Attica had to be completely abandoned; the silver mines at Laurium could not be worked, and some 20,000 slaves took refuge with the Spartans. Widespread revolt could be expected among the allies. The survival of Athens for eight further years is a great tribute to its courage and determination, but the Spartans, as so often before, were slow to press home their advantage. The Athenians were still safe within their walls provided that they could import enough food; this the Spartans could not prevent until they achieved decisive superiority at sea.

In order to preserve what they could of their empire, the Athenians made Samos their chief naval base, and most of the fighting shifted to the east side of the Aegean. The Spartans soon found that their problems were more complex than they had anticipated. While they were able by 412 to build up, with the help of their allies, a fleet sufficiently large to match that of the Athenians, they lacked the financial resources to maintain their crews and had to appeal to the Persians for help. This, however, posed an embarrassing dilemma. The Persian price was naturally the recovery of the Greek cities that had once belonged to the Persian Empire, and Sparta had claimed to be the liberator of the Greeks. It is not surprising that the first agreement with Persia, which recognized the Persian king's right to all that he had formerly ruled, was modified twice within six months, and that Persian pay became very irregular. Spartan relations with Persia also made the east Greeks less anxious to revolt, and Athens still had considerable support from democratic parties that knew that both Sparta and Persia favoured oligarchy.

Oligarchy in Athens. Sparta lost its best chance of an early victory in 411 when Athens was torn by political upheavals. Before the Syracusan disaster there had been no serious threat to the democracy, but conditions were now greatly changed. Periclean democracy depended on state pay for state service, and there would soon be no money to pay. The need for some restraint on the democracy had been recognized in 413 when the Assembly agreed to the appointment of ten elderly commissioners to ensure economies and guide policy. Toward the end of 412 there was open talk of the need for more substantial reform. In the fleet at Samos some of the generals and officers began to plot and approached Alcibiades, who had outstayed his welcome at Sparta and was now exercising his charm on Tissaphernes, the Persian satrap at Sardis. Alcibiades knew that Androcles and his other enemies were irreconcilable, but his family background assured him some support from his own social class. His main hope, however, rested on his assurance that he could persuade Tissaphernes to support the Athenians financially, if they changed their form of government. The oligarch Pisander led representatives of the fleet to put the case for reform before the people, and the prospect of Persian money proved decisive. The Assembly agreed in principle to constitutional changes; Pisander was sent back to complete negotiations with Alcibiades.

Cooperation of the oligarchs with Persia

When it came to the test Alcibiades could not secure the flow of Persian money, but the oligarchs had gone too far to draw back. Pisander returned to Athens, where a revolutionary situation was developing. There was much talk of a government of the 5,000 best qualified in purse and person to serve the state, but in the streets the most determined enemies of oligarchy, including Androcles, were assassinated. Pisander carried a motion that ten commissioners should be appointed to submit proposals to the Assembly on a specified day. The meeting was called at Colonus outside the walls, where, owing to the Spartan presence at Decelea, the common people would feel insecure. The commissioners merely recommended that there should be no state pay for state services, and that the normal safeguard against unconstitutional proposals should be waived. At this point, Pisander, by prearrangement, proposed that a council of 400, appointed by a combination of nomination and co-optation, should have full powers for the duration of the war with authority to convene the Five Thousand for consultation whenever they thought fit. There was no opposition, but, since delay might be dangerous, the Four Hundred entered the council house before the day appointed, with followers prepared to use force, and dismissed the democratic Five Hundred after giving them their salary for the remainder of the year.

The seizure of power was cleverly manipulated by Antiphon, an intellectual who had kept out of the public eye in politics but had helped oligarchs in the courts and in the pursuit of office. The revolution, however, had inherent weaknesses. The voting power behind it was provided by moderates who had no wish to confine real power to a small minority; the government of the Four Hundred was created by and for extreme oligarchs who would have wished to overthrow the democracy even if the treasury was full. They had no intention of sharing power with the Five Thousand. But the Four Hundred could only remain in power if they had the sympathy of the fleet at Samos, on which Athens depended for the protection of its corn ships and the collection of tribute.

They looked to the enemy for support and planned to admit a Spartan fleet into the Piraeus, but this proved the signal for a counter-revolution led by Theramenes, who was to become the recognized leader of the moderates. The Spartan fleet sailed on to Euboea, and the small Athenian squadron guarding the straits between the mainland and the island was overwhelmed; Euboea, which had been in touch with the Spartans, now openly revolted. This was a catastrophic disaster, for a significant part of Athenian food supplies came from the island; it meant the end of the Four Hundred, whose government was dissolved by a special meeting of the Assembly. In

Revolt of the Euboeans

further meetings a moderate constitution was built up, which Thucydides considered the best form of government he had known, although it lasted less than a year.

Meanwhile, the fleet from Samos had won a surprising victory at Cynossema, in the Hellespont. This success was largely due to Alcibiades, who had been recalled by the fleet. After dividing forces over the winter, the united Athenian fleet sailed into the Hellespont and surprised the Spartan fleet outside Cyzicus. The victory was complete, Sparta was swept from the seas, and the Athenians proceeded to recover the cities of the Hellespont and Propontis, over which they had lost control. To raise money quickly a station was established at Chrysopolis near the mouth of the Euxine, which exacted a toll of 10 percent on ships entering and leaving the Euxine. A further consequence was a restoration at Athens of the full democracy in the summer of 410.

By 407 Alcibiades felt sufficiently confident to return to Athens, where his reinstatement had been formally decreed. His first conspicuous gesture was to lead by land the procession to the celebration of the sacred mysteries at Eleusis, which in previous years had gone by sea in fear of the Spartan force at Decelea. He was now given supreme command with full authority to carry on the war against Sparta. Less than a year later a subordinate whom he had left in command of the fleet at Notium, with instructions not to give battle, was drawn into an engagement in which 15 ships were lost. His enemies at home seized the opportunity to stir up feeling against him, and the Assembly, which was becoming increasingly unstable as nerves became more strained, elected a new board of generals. Alcibiades retired to a castle in the Chersonese and from there watched the ruin of Athens.

Lysander and the victory of Sparta. The Battle of Notium (406 BC) had been won by Lysander, the ablest Spartan commander of the war. Intelligent, incisive, and ruthless, he succeeded, where his predecessors had conspicuously failed, in assuring adequate financial support from Persia. He won the confidence of Cyrus, the Persian king's younger brother, who was sent down to supersede the two western satraps, and with his generous help was able to restore the strength and spirit of the Spartan fleet. The Spartan naval command lasted only a year, however, and Lysander's successor was a more typical Spartan. Resenting the humiliation of dependence on Persia, he broke off relations with Cyrus. In a major sea battle off the Arginusae islands, with some 200 ships on each side, he lost his life, and his fleet was crushingly defeated. But 12 Athenian triremes went down with most of their men, and it was thought at Athens that more should have been done to save them. A stormy sea may have been the decisive factor, but scapegoats were needed, and Theramenes and Thrasybulus, who commanded triremes in the battle and had been sent to save the men, helped to steer the blame onto the generals. The Assembly, in a fit of hysteria, insisted that instead of having the individual trials that the law required, they should all be tried together. All six who were present were condemned to death.

The mishandling of this affair aroused bitter feelings at Athens and undermined the confidence of the fleet at Samos; the new generals elected for the next year commanded little respect. The Spartans had meanwhile sent Lysander back to the fleet, nominally as a subordinate but actually to command. He moved up to Abydos in the Hellespont to threaten Athens' lifeline from the Euxine; the Athenians had to follow. They encamped near the mouth of the Goat River opposite Abydos, but after failing to bring the Spartans to battle they grew careless, and Lysander was able to make a surprise attack when most of the crews were dispersed on land. All of the Athenian ships but nine were lost. Sparta now closed the route of the corn ships and the fall of Athens was inevitable. Of the Athenian allies only Samos remained loyal and refused to surrender. A generous decree was passed by the Athenian Assembly offering Athenian citizenship to all Samians, but both cities were doomed. When Athens was near starvation Theramenes was sent to negotiate with Sparta. The terms he brought back were humiliating, but Sparta had resisted the demand of Thebes and Corinth that Athens should be destroyed. The long walls were to be pulled down; all triremes except 12 were to be surrendered; Athens was to follow the same foreign policy as Sparta; and a commission of 30 was to be appointed, of whom ten were to be nominated by Lysander. This commission was to work out the details of a more oligarchic form of government that would be acceptable to Sparta.

Athens besieged: negotiations for peace

The oligarchs at Athens now had a second opportunity, but once again there was a divergence of aim between extremists and moderates, and the extremists, led by Critias, dictated events. The lesson they thought they had learned from 411 was that they must be more ruthless with the opposition, and they extended their murders from political opponents to rich citizens whose money they needed. Realizing their insecurity, they appealed to Sparta to send a garrison, but this increased their unpopularity. Meanwhile, some democrats had found refuge in Boeotia, where the leading politicians were now hostile to Sparta, and a small force under Thrasybulus occupied the border fort of Phyle as a rallying point and from there made a surprise attack on the Piraeus, which as a trading centre was firmly democratic. The population of the Piraeus welcomed them, and they successfully resisted an attack from the city; but what seemed to be the beginning of a civil war was ended by the mediation of the Spartan king Pausanias, whose primary concern was to prevent Lysander from taking control. In the settlement that he encouraged, the leading oligarchs were allowed to withdraw to Eleusis, and the Assembly resolved not to be vindictive but to bury the past. Their bitter experience under the Thirty had ensured that there would be no internal challenge again to democracy for a long time; but the Athenians had to resign themselves to the loss of their empire and to rely once again on their own resources.

THE CULTURAL LIFE OF ATHENS DURING THE WAR

The continuing tensions of the Peloponnesian War hastened changes in the intellectual climate of Athens, the beginnings of which could be discerned in the 430s. This was the generation of the Sophists, who disturbed long-standing traditions and challenged their contemporaries to extend the rationalism that already had been applied to the physical world to man and the state. The Sophists were primarily teachers who taught for pay. They professed to teach the art of politics, and the study of rhetoric was one of their main tools. The development of democracy had increased the importance of public speaking in the Assembly and in the law courts and encouraged the study of rhetoric, which included the right use of words, the emotional rhythm of a speech, and the presentation of arguments. The Sophists knew, as Aristophanes stated in a clever parody in his *Clouds,* how to make the wrong seem right and the right, wrong. Some of the Sophists were little more than fluent salesmen, but others encouraged serious thinking. By raising such basic questions as the difference between what derives from nature and what is man-made convention they questioned traditional morality and undermined faith in established religion, but they paved the way for the constructive philosophers of the 4th century.

Influence of Socrates

The general influences that produced the Sophists also produced Socrates, who was to have more influence on succeeding generations than any of his Greek contemporaries. Unlike the Sophists, who professed to equip their pupils for the practical problems of politics, Socrates, who was himself a poor man, did not charge fees and was concerned to make men work out for themselves true values and to consider ends before means; his method was to move toward the truth by rigorous cross-examination. Though despising current political practice, he regarded state service as a citizen's duty and himself served in the Boule; but, true to his character, when the Assembly, against existing law, insisted on trying six generals together after the Battle of Arginusae, he refused to put the proposal forward. He attracted some of the ablest young Athenians of the day, including Critias,

Alcibiades, and his cousin Charmides. The cruelties of Critias when he directed the oppressive policies of the Thirty and the selfish ambitions of Alcibiades must have been a bitter disappointment to him—their association with him was not forgotten when they were dead and he was brought to trial.

Thucydides, the historian, though very different from Socrates in character and interests, reflects with him the intellectual boldness of this generation. His passion for antithesis shows his keen interest in the rhetoric that the Sophists were teaching, but his mind was tougher and his thinking more profound than that of the Sophists. His *History of the Peloponnesian War* is a penetrating analysis of human nature and political motivation. He was particularly fascinated by the study of power, and from the outset in his summary of early history he emphasizes that it is inevitable that the strong should control the weak; and this assumption pervades his attitude to the Athenian Empire. In interstate relations expediency alone governs policies. He discarded the gods from his history, and for him oracles were only for the superstitious. Though less than a full generation separates him from Herodotus, he thus belongs to a different world.

Thucydides attached considerable importance to accuracy in observation and recording. In this he shared the mental attitude of Hippocrates and the medical writers who followed him. They too had removed the religious framework from their calling and looked to physical causes for physical symptoms. Epilepsy was not a "sacred disease" but a maladjustment of the brain. They built their medicine on a detailed study of cases.

Euripidean tragedy

Euripides, the third of the great Athenian tragedians, was a contemporary of Thucydides and members of the new school of medicine, and he had much in common with them. Like them, he could not be satisfied with traditional religion, but whereas Thucydides quietly ignored the gods, Euripides argued with them; the considered piety of Sophocles was replaced by a restless agnosticism. Euripides was profoundly interested in the roots of human behaviour, and in such characters as Medea and Phaedra he provided psychological studies of considerable power. But Aristotle was not being merely old-fashioned when he complained that tragedy lost something when heroic characters were modernized. Euripides also modified the form of tragedy. His plays tended to be less organic, and the chorus, whose importance was already declining, sometimes served simply as a musical interlude. The same loosening of structure can be seen in his use of the dithyramb, which adopted a freer form and developed the lyric solo, thus giving more play to the emotions.

Similar tendencies toward realism and individualism can be seen in contemporary sculpture, which tended to react against the disciplined restraint of the full Classical style and became increasingly concerned with the decorative treatment of drapery and a wider range of subjects. Of architecture too little remains to judge, but the emphasis on exquisite decoration in the Erechtheum is significant; it would have been out of place in Periclean temples. It is also no coincidence that the most famous painters of the late 5th century, Parrhasius and Zeuxis, carried realism to extremes.

Rationalism, realism, and individualism were the keynotes of this generation, but the attacks on the free-thinkers Anaxagoras, Protagoras, Diagoras, and Socrates are a reminder that conservative religious forces were still strong. Aristophanes made great fun, not always good-humoured, of the academic cleverness of the Sophists in his *Clouds* and *Wasps* and looked back wistfully on the simple instinctive virtues of the men of Marathon, but he knew in his heart that the men of Marathon would have been dull company and he delighted in his own cleverness, which was a match for any Sophist; he would have liked to have the best of both worlds. The new attitudes weakened the close bonds of the city-state, but they were to have a fertilizing influence far beyond the city-state. In Abdera, meanwhile, which had been the home of Protagoras (the greatest of the early Sophists), Leucippus and Democritus laid the foundations for an atomic theory that owed more to the Ionian physicists of the 6th century than to the Sophists. It is astounding that while Greece was torn by the most destructive war in its history, it should have produced such a wealth of creative achievement. (Ru.M.)

IV. The Greek world from 404 to 323 BC

THE SPARTAN HEGEMONY AND DECLINE

Spartan policy toward the Greek states. After 27 years the Peloponnesian War had wound to a close. Sparta had overcome Athens' maritime empire and had destroyed the Delian League. The decisive battle had been fought at Aegospotami, on the European side of the Hellespont, and the defeated Athenians saw their naval and military power melt away. Blockaded and starving within their walls, they sent the wily negotiator Theramenes to Sparta to sue for peace. Peace had been rejected several times during the war by each side, and the so-called Peace of Nicias (421) had barely lasted a few deceptive years. But finally, in the spring of 404, peace was signed by the warring states of Greece at a moment when the vast Persian Empire began to be troubled by the death of Darius II, the man who had finally tipped the balance of the war in favour of Sparta. In 404 the Greek world, divided as never before, despite the peace, entered a new stage of its history.

The peace of 404

Would Sparta, victorious but weakened, succeed where Athens had failed? Sparta was a town without walls whose defense lay in the strength, moral and physical, of its soldiers. Its salient characteristics were toughness, tenacity, and insularity. The city was ruled by a narrow-minded oligarchy that totally lacked imagination. But the Athenians, with all their boast of intelligence and creative energy, had shown themselves incapable of pursuing a long-term policy. Fortunately for the Athenians, they had an oligarchic party traditionally devoted to a Spartan alliance, while Sparta itself was divided into two factions: that of Lysander, victor of Aegospotami, the implacable war leader, obstinately hostile to Athens, and that of Pausanias, the king, perhaps jealous of his rival's prestige but also wise or generous enough to realize that the total destruction of Athens would create a dangerous imbalance of power in Greece.

Sparta's allies, especially Corinth and Thebes, wanted Athens annihilated once and for all. The Spartans would not go so far, but their terms were nevertheless severe. Athens kept Attica and Salamis but had to surrender what was left of its fleet and demolish the fortifications of the Piraeus, the city's port. Further, the city was to raze the long walls that defended it, recall the exiles, abandon all foreign possessions, and conclude an offensive and defensive alliance with Sparta. Once the head of a maritime empire, Athens was henceforth merely a member of the Peloponnesian League under the leadership of Sparta.

As soon as the peace was signed, Lysander made a triumphal entry into the Piraeus. He ordered the immediate demolition of the walls and called upon the pro-Spartan party to take over the government of Athens. In September 404, backed by a vote of the terrorized Assembly, he put in power a group of thirty Athenians, which included the peace negotiator Theramenes and Critias, Socrates' former pupil and first cousin of Plato's mother.

The Thirty and their regime were destined to leave behind unpleasant memories. They reduced the citizen body to 3,000 privileged persons and for the rest behaved as tyrants. Lysander had chosen the right men to complete the dissolution of the city he hated, but Sparta sent him on a new campaign overseas, and the Thirty held power only for eight months. They reigned through terror, and their rule was one long succession of arrests, political trials, and profitable confiscations. Fifteen hundred citizens were put to death, guilty only of political opposition, and pressure was put upon the Persian satrap Pharnabazus to have the exiled Athenian leader Alcibiades murdered.

Complicity in crime makes men stand together—for a time. But, while dissension was growing among the Thirty, while Theramenes was falling out with Critias, civil

war was on the point of breaking out in Attica and being loosed upon Athens. Opposition to the Thirty was led by the exile Thrasybulus, chief of the democratic party at Athens. Surreptitiously helped by Thebes, his little army grew day to day, and in May 403 he won a decisive victory over the Thirty in the Piraeus, Critias perishing in the battle.

Deposition of the Thirty

The surviving tyrants were deposed and took refuge at Eleusis, where two years later they were massacred in an ambush. Lysander reappeared and for a moment was ready to blockade the Piraeus, until Pausanias hurried from Sparta to re-establish order. After defeating Thrasybulus' forces, Pausanias was wise enough to set himself up as arbitrator between the two parties: that of the city, as it was called, the more or less faithful followers of the Thirty, which had Athens as its seat of government; and that of the Piraeus, the democrats united around Thrasybulus. Pausanias managed to bring the two parties together, and in autumn 403, under his auspices, the Athenians voted for a general amnesty, from which only the Thirty and the most heavily compromised of their officials were excluded. The old regime was restored, and in the following year (403–402) the Athenians undertook a general revision of the laws and a reorganization of the citizen body, which had suffered much from the two wars, external and civil. In 401–400 the amnesty was confirmed. Socrates, however, did not benefit from it, and he in his turn drank hemlock in 399.

Athens had thus lost its empire and its leadership of the Greek world. But, if the Greek cities thought that the time had come for them to recover or obtain their freedom, they were soon disabused. Sparta, short both of ideas and of manpower, impoverished by a long war, and haunted by the spectre of a helot revolt, proved incapable of fashioning a new system and merely imposed its own empire in place of that of Athens.

In Thessaly the Spartans put down a civil war; in the Peloponnese they strengthened their own domination. In particular, they invaded and destroyed Elis, leaving that city nothing but its traditional responsibility for the Olympic Games. Sparta, however, was directly threatened by trouble at home. In 397 a plot led to a massacre of citizens. Supported by Lysander, who was used to plots, the new king, Agesilaus II, destined to become the great man of Sparta and friend of Xenophon, drowned the conspiracy in blood, and Sparta was free to complete the task of establishing its own domination over the Greeks. Lysander made them realize that they had merely changed their master.

It was Lysander who, since the peace of 404, had been organizing a Spartan empire through blood and war methods. He brought under Spartan hegemony nearly every Aegean government that had once been controlled by Athens. Athenian cleruchs (colonies placed in conquered territory) were replaced by Spartan garrisons, Athenian *episkopoi* by harsh governors known as "harmosts." The islands, including the sacred island of Delos, were reduced to servitude, and in the cities that had formerly belonged to the Delian League he installed oligarchic governments loyal to Sparta; ten magistrates were invested with supreme power and instructed to exact the same tribute the islands had paid to Athens. In Greece the new system was merely a copy of the old, under a different master. In Persia, however, there lived a king, not a Greek, who might have something to say about Greek affairs.

Relations with Persia. The services that the younger Cyrus had rendered to Lysander from 408 onward had not engendered any gratitude toward the Great King, and when Darius II died, in 404, the victorious Spartans no longer needed either his financial or his military support. The treaties between Sparta and the Persian king, concluded without good faith on either side, had put the Ionian Greeks at the mercy of Persia. Now Sparta, wishing to consolidate its new empire, wanted to rally under its banner all Greeks, including those of Asia. On this point the Spartans risked the hostility of Persia.

Caught between two obligations, Sparta proceeded cautiously, profiting from Persian dynastic quarrels over the succession to the throne. The younger Cyrus had mounted an expedition in Lydia to depose his brother, the new king of Persia, Artaxerxes II. By extravagant promises he had persuaded 10,000 Greek mercenaries to go with him across the deserts, first against the satrap Tissaphernes but ultimately aiming to reach the capital city, Susa, in the heart of the empire. Sparta helped Cyrus in his attempt (401), sending a small fleet and a few hundred hoplites. It was a ridiculous escapade, though it allowed Xenophon to distinguish himself by leading the survivors back to Ionia after two years of hardships. The real end of the enterprise came only eight months after the expedition set out, at the Battle of Cunaxa, where, although the King's army was defeated, Cyrus himself fell, thus depriving the expedition of any point it may have had. In any case, a local wound had no effect on the huge body of the Persian Empire. The Ten Thousand had shown, however, that with determination Greeks could venture far from the sea, cross unknown lands, and challenge the Great King with impunity. The lesson would not be forgotten.

The Ten Thousand

Nevertheless, the Greek cities of Ionia had asked Sparta, in which they saw the champion of their liberties, to defend them against the satrap Tissaphernes. The return of what remained of the Ten Thousand allowed Sparta to economize on its own citizens and to send only a small body of troops into Asia, with an inexperienced general, Thibron, who soon had to be replaced by Dercyllidas. A considerable number of Greek cities were thus freed from Persian control. The King, however, finally tired of these provocations and prepared for full-scale war. In spring 396 Sparta had to send a large army, commanded by Agesilaus. He was a brilliant tactician, and Xenophon, who served under him, invests him with every virtue both as a man and as a general. He created a body of cavalry, organized fleet and infantry, and in a two-year campaign in Ionia went from success to success at the expense of the satraps Tissaphernes and, later, Tithraustes. But Susa was still far away, across the desert, and the empire could not be shaken. Though too big to be easily defended, the empire was also too big to be conquered.

Agesilaus, moreover, was obliged to retrace his steps, furious at heart, and had to abandon any attempt to follow up his victories: ten years after Aegospotami, in 395, Sparta was again threatened in Greece itself and recalled the King to meet the new emergency.

The Corinthian War (395–387). The peace of 404, far from satisfying Thebes, had merely intensified its bitterness against its Spartan ally. Discontented, Thebes naturally sought a rapprochement with Athens, which could now be of assistance in upsetting the unbearable domination of Sparta. While Persian gold once more flowed into Greece to stir up the Greeks against Sparta, the old enemies Thebes and Athens entered into an alliance and launched a campaign in central Greece that cost Sparta the life of Lysander, drove Pausanias into exile (395), and brought into being against Sparta a coalition of Argos, Corinth, Thebes, and Athens, whose interests for the moment coincided.

Alliance of Thebes and Athens

Agesilaus, hastening back from Asia, won a brilliant victory over the forces of the coalition at Coronea in Boeotia in 394. The victory, however, was not followed up, because, at the same time, the Athenian admiral Conon, the defeated commander at Aegospotami who had since entered the King's service, crushed Sparta's fleet at Cnidus, in Ionia. On the strength of his victory, Conon returned in triumph to Athens, where he set about rebuilding the walls in defiance of the peace of 404. Thus, both in Greece and in Asia Minor, Sparta's empire suffered severe blows at the hands of the Greeks, while that of Athens seemed capable of rising from its ashes.

To save its empire, Sparta took the easy way out. Without further hesitation, it changed policy once again and sought the help of the barbarian against the Greeks.

The King's Peace (386). Once again the internal conflicts among the Greek cities made the Great King the arbiter of Greece. There was a rush of embassies to Sardis, on the banks of the Pactolus, capital of the southern Ionian satrapy—embassies from Sparta, on the one

side, and from Athens, Thebes, and Argos, on the other. To stop the Athenians, who dreamed of re-establishing their empire, the Spartan envoys supported the autonomy of the Greek cities but at the same time proposed that the Greek cities of Asia Minor come under the Great King's authority. Thus, in 392 they proclaimed their pro-Persian policy. The following year another congress was held, this time at Sparta, and again the Spartans proposed autonomy for all the Greeks. This proposal was deceptive and was not carried out, since it aimed simply at separating Athens from its new allies. The Corinthian War flared up again, and hostilities began again both in Greece and in Asia. Athens, however, was not taken unawares; under the influence of Iphicrates the city restored its fleet and modernized its army.

While Agesilaus was occupied with a hard struggle against the Argives and Corinthians in Acarnania, Iphicrates went from success to success. Another Athenian, Chabrias, was sent in 390 to Cyprus to support the petty king Evagoras against the Great King. This interference was unwise at a time when money was scarce and the courts were working full-time to make citizens pay their war tax. It presumed upon Athens' strength and at the same time pushed the Great King more toward the side of Sparta, which, in addition, received help from the west, that of the elder Dionysius, tyrant of Syracuse.

Sparta, now confident of success, sent Antalcidas on a second embassy to Sardis, then to Susa (387). The result was the King's Peace (or the Peace of Antalcidas; spring 386). It bore the stamp of Sparta and Persia, which had agreed to divide up the world. Representatives from Athens, Argos, Corinth, Thebes, and, for form's sake, Sparta were summoned to Sardis to hear the terms of the peace that the King in Susa was pleased to grant. The west was to belong to Dionysius of Syracuse; the east, with the Ionian Greeks, betrayed by Sparta, to Artaxerxes; and Greece itself passed once again under Spartan hegemony. Finally, if any city refused to sign, the King would make war on it.

Aftermath of the King's Peace; Sparta's decline. Thus, the power of the Achaemenids was imposed on all the Greek cities, both in Europe and in Asia, and even on Sparta, whose complicity merely concealed its subservience. This proud city could maintain its precarious hegemony over the Greeks only by showing itself pliable enough to deserve the Persian king's support.

But, if Sparta was pliable toward Persia, it showed a new rigidity toward the Greeks. Doubtless the King had laid down that all Greek cities should be "autonomous." But if Sparta could accept this clause, which really meant autonomy for itself and for no one else, the other cities could not accept a system that stopped them banding together against Spartan tyranny.

Spartan lack of leadership

Sparta itself, if only it had been able to escape from its ossified principles and take a clear and balanced view of the realities of world politics, would have realized that its own mistakes, aggravated by its own deficiencies, made the King's Peace unworkable. Sparta was perpetually short of men, of citizens, even of generals. Agesilaus was still on hand, no doubt, wherever his services were needed, but the hidebound system provoked inevitable conflicts between the ephors and the kings. Sparta, the military city *par excellence*, had produced no general capable of modernizing its methods of warfare, as Iphicrates had done at Athens. The obscure figures who directed Sparta's destiny could imagine no way save that of brute force to keep order in Greece. They did not realize that the Athenians were silently working to rebuild their forces. Time itself was working against Sparta, blinded by its deceptive and doomed hegemony.

The Greek cities of Asia Minor had been under Persian rule for a half century—indifferent to liberty and slumbering in commercial prosperity. The King intervened in Egypt to put down disturbances and in Cyprus to reduce Evagoras and thus deprive Athens of its last supporter in the east. Meanwhile, Agesilaus, in Greece, applied the most inflexible principles of Lysander. He laid an iron hand on Corinth, Mantineia, and Megara, and suppressed any movement for independence outside the Peloponnesian League. Conversely, autonomy was granted to the cities of Boeotia, to undermine the supremacy of Thebes.

Sparta did not realize that its successes in fact masked serious defects. Passing through Boeotia on his way north to conduct operations in Chalcidice, the Spartiate Phoebidas, secretly backed by the Spartan government, seized the Cadmea, the citadel of Thebes. This outrage, which caused a sensation in Greece, was followed by a revolution in Thebes, where Sparta took the measures necessary to ensure that no help came from Athens. The presence of Peloponnesian armies in Chalcidice was equally directed against Athens, which had to be prevented from helping the Chalcidian League. Ever since the time of Brasidas, Sparta knew how important it was to Athens to maintain its presence in these Thraceward regions. The Spartan generals sent north at this time met many difficulties, but in the summer of 379 they secured the surrender of Olynthus, a city destined to play a large role in the future history of Athens and of Macedonia.

Sparta had never appeared more powerful, but it was only in appearance. In Xenophon's words, the Spartans "could at this time believe that their complete domination was gloriously and solidly established," but, he adds, "the gods do not forget men who violate both divine and human laws . . .; those who had sworn to leave cities autonomous were about to suffer their first punishment." Excess in fact brings nemesis.

Theban revival

In December 379 Thebes awoke with a start. Two Thebans, Epaminondas and Pelopidas, organized a conspiracy against the Spartan garrison of the Cadmea. The Spartans, taken by surprise, had to surrender. An army promptly came up from Sparta, but it, too, was defeated in front of Thebes. One of its generals decided to imitate Phoebidas' surprise attack by risking an attack on the Piraeus, which he almost captured. Greece was scandalized by this second attempt contrary to international law, and the acquittal of Sphodrias when he was put on trial at Sparta opened the eyes of all those who had blinded themselves to Spartan intentions.

Animated by a common hatred of Sparta, Athens and Thebes again concluded a defensive alliance (378). Agesilaus appeared before Thebes with a powerful army, but his failure merely strengthened the tie between Athens and Thebes that marked the end of the Spartan hegemony. It had taken Sparta a quarter of a century to bring down Athens. The same length of time had now proved sufficient for Aegospotami to be wiped out.

The year 378 marked the failure of Sparta's policy in Persia as well. Although individual Greeks had long advocated a rational crusade against the Persians, Artaxerxes was beginning to doubt the wisdom of supporting Sparta. There was little point in interfering in Greek affairs. The Greeks hated one another and could be relied upon to destroy themselves. Their threat to Persia was negligible.

The Second Athenian League; the restoration of democracy (378–371). The Athenian democracy, destroyed and then restored in 404, had never let the flame of its traditional imperialism die down completely. Helped by the satrap Pharnabazus, the Athenians had profited from Conon's victory at Cnidus (394), which destroyed Sparta's new naval supremacy, by recovering a number of islands, such as Lemnos, Imbros, and Scyros. Three years later, Andocides delivered his speech "On the Peace," in which he developed the theme of an Athens whose strength was based on the Piraeus, the walls, and the fleet, without alienating Sparta, and of prosperity through power. In 389 the theory was put into practice with Thrasybulus' campaigns in the north of the Aegean, which were hampered by the emptiness of the treasury but which aimed to re-establish an Athenian League and thus to bring money into the state's coffers. The King's Peace in 386 had lulled Sparta's anxieties by checking this policy of patient rebuilding. But in 378 the Theban alliance with Athens revived the latter's imperial ambitions. In the intervening period they had never missed a chance of keeping open relations and, where possible, of maintaining alliances with the Aegean islands. The islands, on their side, from 386 onward saw in Athens their own defense

against Persian domination. Athens had signed a pact with Chios in 384; it now formed an alliance with Byzantium, Lesbos, and Rhodes (378).

Constitution of the league

It seemed a suitable moment to find a constitutional formula that would replace these separate alliances and would encourage other new allies to join together into a single organization. Negotiations lasted from autumn of 378 through spring 377. This shows how much care went into them. Finally, the decree of Aristoteles gave the Second Athenian League its constitutional form.

The leaders of Athenian policy were basically well aware of the abuses to which Pericles' principles of imperialism had given rise, although it cannot be shown that they wished to guard against their repetition. Nevertheless, for the sake of form at least and perhaps to reassure the new members of the league, the constitution was based upon theories of federalism that were interesting by their novelty and that showed an impulse toward reform or at least a real intention of it.

The new federal organization respected the terms of the King's Peace as far as Persia was concerned but invited all peoples not under the King's rule to join the league. Their autonomy was guaranteed, putting them rigorously on an equal footing with Athens and giving Sparta no cause to interfere. No city had a foreign garrison or foreign governor. At the Synedrion, or Council, which met at Athens, each member had one vote only. On the financial side the "tribute," of bitter memory, gave place to a "contribution" to the federal treasury. These were the essential features of this first attempt in Greek history to form a representative international system with the cities sending delegates to a permanent assembly. But Athens, of course, kept its own Assembly, alongside the Synedrion, so that there existed two assemblies of equal standing in law, with that of Athens in fact exercising considerable influence. As it was, however, the organization was well-enough planned to last for 40 years, until 338, after Philip of Macedon had defeated the Greek forces at Chaeronea.

Borne up by new hopes, stirred as always by their spirit of initiative, the Athenians turned to domestic reforms affecting the Council, the law courts, the fleet, and above all—thanks to Callistratus of Aphidna—the financial system. Since there was to be no more tribute from the allies, the basis of taxation was revised, and a census of private wealth was taken. The citizens were divided into a hundred *symmories*, each of which was responsible through its own tax collectors for paying to the treasury $\frac{1}{100}$ of the tax levied by the state.

The lawmakers having done their job, it was time for the men of action to do theirs. Among the more prominent of them were the generals Iphicrates and Chabrias and a first-class admiral, Timotheus, son of Conon and follower of Isocrates. It was thanks to Athenians of this calibre that the new league got under way. The most extravagant hopes seemed justified, whereas Spartan disquiet grew proportionately. The Spartans had no success in operations against Thebes nor in an attempt to rebuild a fleet to regain the mastery of the Aegean. Chabrias and Timotheus, on the other hand, were winning victories everywhere, and each victory joined new cities to the league, which, young as it was, thrived to the point where it recalled the Delian League a century earlier. Nothing was missing save the Greek cities of Ionia, which the peace of 386 had abandoned to the King.

Sparta felt threatened and beneath this threat succeeded with the help of the King and of the elder Dionysius, tyrant of Syracuse, in concluding a new treaty with Athens (374): Athens recognized Sparta's supremacy in the Peloponnese, which was no great sacrifice, and Sparta, in return, recognized the Athenian League. Then Sparta changed its policy, broke the treaty, and again intervened against Athens in the west. It was merely a brush fire, however, for Timotheus, Callistratus, and Iphicrates, although bickering among themselves, soon brought Sparta to see reason.

Sparta was in full decline and no longer feared. Athens, however, had a new enemy on its hands and another in view on the horizon. In Thessaly, Jason of Pherae, thinker and man of action alike, succeeded in the difficult task of uniting the country and thus making it powerful. A still graver danger threatened in Boeotia, where at Thebes, slowly but surely, Epaminondas and Pelopidas had reorganized the army and indeed reorganized the whole country around the army, turning the old Boeotian League into a powerful and well-organized federal state. Wisdom suggested that peace in Greece should be set on a more solid basis.

Rise of Thessaly and Thebes

Sparta wanted peace, while Athens, prey to new financial troubles and worried by the growing disharmony among its best generals, was also interested. Athens, moreover, could count on the support of a league now 75 strong, though this was primarily moral support, since the members of the league did not have the financial resources to wage a war and were not prepared to sacrifice their own money once the danger from Sparta was gone.

It was thus the moment for peace. In 371 there met at Sparta a congress not merely of all the Greek states but one on a truly worldwide scale, since in addition to Dionysius of Syracuse the barbarians were also represented, notably Amyntas of Macedonia and Artaxerxes of Persia. Agreement was reached, and peace seemed assured for many years to come. But, at the last moment, one of Epaminondas' demands caused the Boeotians to be excluded from the treaty, to their great pride at this distinction. On the one side stood the world and on the other Thebes, not caring that it stood alone, because with its outstanding generals and its present strength it felt itself in a position to take over from Sparta and Athens the hegemonies they had long exercised.

Theban expansion and containment. With the stubborn patience of the peasant used to the inevitable cycle of the seasons, the Thebans had been waiting for their moment to come since 404. They had experienced the bitterness of the realization that in helping Sparta to crush Athens they had helped the most ungrateful of allies, who had then refused to share the harvest they had reaped. The King's Peace had shown them, moreover, that Sparta was capable of any treachery to consolidate its own supremacy in Greece. The peace had forced them to dissolve the Boeotian League; the seizure of the Cadmea by Phoebidas was the last straw, proving, as if further proof were needed, how little respect Sparta had for its staunchest allies. Thebes was ready for revenge as soon as the hour should strike.

Thebes could stay aloof of the peace of 371 without danger. In merely eight years Pelopidas and Epaminondas had made Thebes the most powerful city of Greece: working together, they had reorganized the army, had created the famous Sacred Band composed of 150 pairs of young men from the leading families, and had given an unusual flexibility to military tactics by inventing the charge in oblique order, by echelons, thus permitting the army to pierce the opposing line like a trireme with its ram. They also re-established the Boeotian League on a new basis and in defiance of the King's Peace.

At the beginning of July 371 the Spartan king Cleombrotus, relying on the reputation of his hoplites, unwisely invaded Boeotian territory. He supposed that with his army of 10,000 men he would have no difficulty in forcing Thebes to dissolve the league, which was intolerable to Sparta and which made a mock of treaties. But surprise carried the day. At Leuctra, west of Thebes, Epaminondas inflicted on the invading forces a blow from which Sparta never recovered. Leuctra was the first great defeat of the Spartan infantry.

Battle of Leuctra

All Greece was stupefied and flocked to join the victor. The Boeotian League was strengthened, the Athenian Confederation enlarged, the Peloponnesian League dissolved. The Spartan harmosts who still remained were driven out or killed. Soon Sparta felt itself threatened in the very heart of the Peloponnese, as the vice slowly tightened. The Argives rebelled, Elis recovered the territories Sparta had stolen from it, and Mantineia raised its head and welcomed the advances that were made by its old enemy Tegea. At Thebes's instigation a new league of Greek cities, the Arcadian League, was planned and indeed realized. A new capital was established, Mega-

lopolis, at the foot of Mt. Lycaeus, and a new federal constitution was introduced, followed by a police force. The first act of the new league was to call upon Epaminondas for help against Sparta.

Epaminondas came down into the Peloponnese and carried fire and slaughter to the very gates of Sparta. As Xenophon says, "The women of the city, who had never seen an enemy in arms, found the smoke of the fires lit by the enemy unbearable." Two Peloponnesian campaigns (a third was to follow) sufficed for Epaminondas to free Messenia. In 369 he again founded a capital, Messene, a new town with magnificent ramparts that made Sparta regret its pride in having no other defense than, as has been noted, the stalwart breasts and shields of its hoplites.

By climbing at the first blow into first place, Thebes upset the balance of Greece, which was always precarious. Its only difficulties were in the north, in Thessaly and Macedonia. Thessaly had revolted after the assassination of the brilliant Jason of Pherae, the ally of Boeotia, and Thebes intervened against his successor, who was allied to Athens. In Macedonia, Iphicrates seemed likely to establish Athenian supremacy. Thebes skillfully manoeuvred to secure an alliance with Macedonia, and the alliance was guaranteed by the despatch to Thebes of a hostage destined to a great future, the young prince Philip of Macedon, the future Philip II.

Everywhere else Thebes's successes were instantaneous and so widespread that they had two serious repercussions, in Asia as in Greece. Faced with the menace of Thebes, Greek envoys were sent to Susa in 367 to beg for the help of the King of Persia. Artaxerxes chose the stronger party, and by a rescript issued from Susa he opted for Thebes, against Athens and above all against Sparta, which was terrified of losing the advantages of the King's Peace. Athens, for its part, was forced to choose between Sparta and Thebes. Two years earlier, Iphicrates had helped Sparta, and Athens had granted the Spartans an alliance. In 367 the Athenians decided to stand by it, and backed it up by an understanding with the elder Dionysius of Syracuse, to assure its own security in the west.

General warfare in Greece

Despite these treaties and despite the King's intervention, the balance of power remained uncertain, and the Greeks feared a general conflagration. In 367 Thebes tried to summon a peace congress but failed; for Thebes itself would have been the arbitrator of peace. The enmity of the Greeks rekindled the war by sea and by land. By sea Athens sent a fleet to Asia under the command of Timotheus to put pressure on Persia, to obtain some alleviation of the rescript of Susa, and to support one of the satraps against the King. Epaminondas, on the other hand, built a fleet to challenge Athens' naval power and dared to launch a campaign right up to the Bosporus. Theban daring knew no bounds.

By land, war spread everywhere. The Thessalians were again restless, and Thebes had once again to intervene. The Theban victory at Cynoscephalae in 364 cost Pelopidas his life but secured peace in the north. In the south, however, the situation was becoming serious as the result of a local conflict between Elis and the Arcadians, which quickly assumed major proportions throughout the Greek world, since the sanctuary of Olympia lay in the territory of Elis. Every city took sides, and the Arcadian League itself split. This was playing Thebes's game.

In the spring of 362 Epaminondas decided on a fourth campaign in the Peloponnese, and on July 4 two Greek armies faced each other at Mantineia, north of Sparta. On the one side were contingents from Sparta, Athens, Elis, and the northern Arcadians, a total of 20,000 men. Epaminondas faced them with an equal number of Thebans, southern Arcadians, Messenians, and Argives. By skillful tactics the Theban general came within an ace of a great victory, but his death at the height of the battle left the result indecisive.

Only exhaustion stopped the war. Sparta was at the end of its tether, Athens had lost confidence in everything except the traditional supremacy of its triremes, and Thebes, with the deaths of Pelopidas and Epaminondas, had lost all chance of keeping its position as the leading state of Greece. The Boeotians could do no more than cast a regretful glance on their brilliant past as they went back to their fields. Untroubled peace confirmed a status quo of exhaustion among the Greek cities.

Peace and the balance of power in Greece (362–355). Plutarch relates that the last advice of his compatriot Epaminondas to the Thebans as he lay dying on the battlefield of Mantineia was to make peace. Peace was made but, as noted, for the simple reason that the Greeks had no strength left to make war and not because they had decided to abandon their quarrels and renounce war for good. Sparta, like Thebes in 371, kept aloof from the treaty of 362, because in its case the recurrent danger of helot revolt did not allow it to recognize the independence of Messenia. Nevertheless, while dreaming of its glorious past, Sparta did actually keep the peace.

Like Sparta and the Peloponnesian cities, Thebes turned in upon itself. Far from the battlefield, the soldier-peasant of the Boeotian League once more observed the succession of work and days. But Epaminondas and Pelopidas had left an example of how to combine force with flexibility, and this example was not lost on the young Greek–barbarian prince, Philip of Macedon, who had come to the great Thebans as a hostage and remained as their disciple and friend.

Until this young prince and, later, his son and successor Alexander had shown them the way, the Greeks in the first half of the 4th century BC conceived of no pattern of international relations except groupings of cities around one of their number that was for the moment powerful. The Athenian hegemony had ended in 404, the Spartan in 378, the Theban in 362. Moreover, none of these cities had been able to keep its pre-eminence for any considerable period except by an alliance with the next-strongest. The balance of power subsisting after the Battle of Mantineia rested solely on an equal degree of exhaustion all round. Moreover, Greece had, as it happened, nothing to fear for the moment from Persia, which was disturbed by the revolt of the western provinces of its overgrown empire.

Collapse of the Athenian League

Athens, however, no doubt because its league extending overseas gave it a fleet, seemed for the moment, at least in theory, the least feeble of the Greek cities. The empire was vital to Athens' existence, but its extension created problems, and before long the Athenian League caused Athens grave worry.

The chief anxieties came from the north, from the lands far beyond Thebes, lands that always played a role in Athenian history. In Thessaly, Alexander, the successor of Jason of Pherae, was beginning to challenge Athens even by sea; but the aristocratic Thessalian League was scarcely to be feared as long as the ancient family of the Aleuads counterbalanced the ambitions of the tyrants of Pherae. Moreover, Athens' fleet was still more powerful than Alexander's.

Nevertheless, this fleet had need of bases around the Aegean and of free access eastward to bring the grain of the Black Sea regions to the Piraeus. Athens, however, could not succeed in regaining control of Amphipolis, now a member of another league, that of Chalcidice, under the protection of the kings of Macedonia. Another enemy of Athens was Cotys, King of Thrace, who was waging war against them in the Hellespont. Only his assassination, in 359, delivered them from the annoyance of this petty barbarian chieftain. In the direction of Magna Graecia, there were also storm clouds on the horizon. Corcyra, torn apart by its eternal factional quarrels, had seen the oligarchs get back into power and in 361 had left the Athenian League.

The Athenians in their nervousness were all too ready to turn to the old remedy of blaming the generals for the deficiencies of popular policy. The black days of the trial of the victorious generals after Arginusae seemed to have come back. The best generals were accused, arrested, and sentenced to fines, exile, or death.

Plots against the national security were often facilitated by Thebans working under cover and helped by the existence at Athens of a pro-Theban party. In 357 Thebes

stirred up trouble in Euboea, the island that was Attica's natural shield and essential to the Athenian supply route. The revolt was at once suppressed, but it showed that things were sadly amiss in the league. Faced with the problem of tightening or relaxing the ties that united the allies, Athens tightened them, and they broke, resulting in the Social War, which lasted two years until 355.

In Autumn 357, Chios, knowing Athens' need for money, calmly refused to pay its contribution to the league. Urged on by the ruler of Caria, Mausolus (whose tomb was later among the Seven Wonders of the World), Chios allied itself with Rhodes to the south and Byzantium to the north. So disquieting was this threefold defection that Athens sent its best generals into the Aegean. At the Battle of Embata, near Chios, the fleet failed to carry the day, which led to a trial at Athens, wherein Iphicrates was acquitted but Timotheus condemned to exile, whereupon he died. The third general, Chares, stayed in Asia to stir up dissension among the satraps and trouble the new king of Persia, who was proving difficult. But the young Artaxerxes III Ochus feared nothing; he sent Athens an ultimatum: either recall Chares or face war.

Athens, no longer in a position to choose, gave in to the demand. Thus ended the last struggle between Greeks who were supposed to be allies before they were brought into line by Macedonia. Once more, it took the Great King's intervention to bring the Social War to an end. Athens was ruined, suffering hunger and thirst. It gave up the struggle with Mausolus, recognized the independence of Byzantium, Chios, Rhodes, and of the neighbouring island of Cos, in the Aegean; of Corcyra, in the Ionian Sea; and shortly afterward of Mytilene, on the coast of Asia. All Athens kept in its possession were some tiny Cyclades, a few harbours in the north, and the right to sail freely in the Aegean. Athens had lost another empire, 50 years after the first.

The end of Athenian power

Athens had lost not only its empire but its illusions and ambitions as well. The city lived from day to day, forced to restore its finances, rebuild its economy, replace its triremes by cargo vessels. Eubulus was probably the man it needed, in 355, to recover its balance and its good sense. He was an economist and financier, scrupulously honest, a man of peace. In their weariness the people accepted him as administrator in every area of government. He struggled against fraud, waste, peculation. He balanced the two treasuries, the one reserved for the army, the other for public entertainments, somewhat to the benefit of the latter, the *theōrikon*, which helped him to maintain his popularity and thus stay in power. The improvement in the finances permitted him to undertake an extensive and useful program of public works. Athens was, however, finished as a military power.

This, however, exposed the city to danger, and a new orator, Demosthenes, saw this and began to be disturbed. For, whereas leagues and confederations had everywhere failed, the Greek world saw the rise on all sides of princes, dynasts, tyrants, and kings, both great and small, intelligent and bold, such as the elder Dionysius in Sicily, Archytas of Tarentum, Jason and Alexander in Thessaly, Evagoras in Cyprus, Cotys in Thrace, and Mausolus in Caria, not to mention Artaxerxes III in Persia. Perhaps their strength was not firmly based, but they had dared raise their heads against Athens and had on occasion shaken it.

And now appeared on the scene Philip II, he who had meditated on the example of Pelopidas and Epaminondas, overcoming dynastic problems and seating himself firmly on the throne of Macedonia, just at the moment when Eubulus' policies closed the eyes of the Athenians to the danger he presented. The year 355, like 404, opened a new era in history. It marked the end of the Athenian Empire, along with the end of the Social War. It closed an era that had seen all dreams of Greek unity brought to nought. And with the new age of Macedonia it opened new horizons.

THE RISE OF MACEDONIA

Early Macedonia. United or otherwise, the Greek cities were a mere shadow, alongside which a new state had been gradually taking shape since the beginning of the 4th century—Macedonia, a backward country of hunters, drinkers, fighters, and peasants, whom the Greeks regarded as barbarians. The language of the country was basically Greek but so mixed with Illyrian and Thraco-Phrygian that it was incomprehensible to Athenians. Because of their violence and appeal to madness, Dionysiac rites were much celebrated. The form of government was that of a hereditary monarchy under the family of the Argeads, who enjoyed the dubious support of a great landowning aristocracy. With each generation, the kings of Macedonia tried to introduce more and more Greek civilization into their wild land.

During the Peloponnesian War, Athens had made great efforts to secure the alliance of King Perdiccas II, son of Alexander I. Macedonia could be a valuable ally. The region was rich in gold and shipbuilding timber and also lay upon the chief route for the transport of grain to several Greek cities, including Athens. The Athenians, however, might have done better to have Perdiccas as their enemy. He proved an inconstant friend and changed his coat several times before his death in 413. His son Archelaus proved, on the contrary, a great ruler, a Philhellene, and a friend of literature and the arts. He moved his capital to Pella and made it an intellectual centre to which he attracted Greek poets and musicians, such as the great tragedian Euripides, who conceived his tragedy, the *Bacchae*, at the court of this enlightened despot. Archelaus reorganized the government, centralized power, and fashioned his troops into a strong regular army. Under his reign, Macedonia ceased to be convulsed by its endemic political crises, and the historian Thucydides—a severe critic—praised him as a great builder and administrator.

Achievements of Archelaus

But, when Archelaus was assassinated in 399 by a favourite, Macedonia was plunged into 40 years of palace revolutions. Apart from its internal problems, Macedonia was also threatened from all sides: by the Paeonians to the north, the Illyrians to the west, and the Thracians to the east, while the Greek settlers in Chalcidice, to the southeast, could scarcely be considered allies; furthermore, the country lacked an outlet to the Aegean.

Amyntas III (reigned *c.* 393–370/369) had great difficulty in consolidating his throne and sought help first from Sparta, then from Athens, whose rights to Amphipolis he recognized in 371, one year before his death. His elder son and successor, Alexander II, was killed by his mother Eurydice's lover, Ptolemy, and the pair of them ruled as regents until Ptolemy, in turn, was assassinated by the second son of Amyntas III, the young Perdiccas III, then aged 20. This young king, a Philhellene and lover of philosophy, called on the exiled Athenian Callistratus to reorganize his treasury and followed the example of Archelaus in making Macedonia a strong power, until he was killed in 359 after a reign of six years in a campaign against the Illyrians, who were occupying part of his kingdom.

Problem of the succession

Macedonia was again in danger; but it can be supposed that these trials were tempering the character of the Argeads and forming a dynasty that was to produce Philip and Alexander. The problem of the frontiers was serious, but that of the succession was critical. In 359 several claimants sought the throne, supported by foreign armies. Perdiccas' son Amyntas was too young to be king, but the throne was claimed by his paternal uncles: Archelaus, Arrhidaeus, Menelaus, and Philip. Philip, the youngest, had spent three years as a hostage at Thebes, the friend of Epaminondas and Pelopidas. Philip was appointed regent for his nephew Amyntas and thus represented the legitimate power. Few could have guessed that this young Macedonian prince, aged 23, would, in less than 20 years, shape a Hellenic nation united under his own absolute authority.

The rise of Macedonia under Philip II. Unscrupulous, intelligent, determined, backed by no assembly, Philip succeeded where no purely Greek people had done so. In the past the cities of Greece had merely formed alliances or leagues under temporary hegemonies that masked or failed to mask the subjection of the weak.

Philip, by imposing a new domination on the Greeks, achieved for the first time a new unity whereby all were equally enslaved.

At 23 Philip was already experienced in politics and administration, having governed one of the provinces of Macedonia after his return from Boeotia. It took him only three years to oust his nephew from the throne. When in 356 the princess Olympias gave him his hoped-for heir, Alexander, he exchanged his regency for the title of king. This gave him the green light for his hopes and desires. He could devote himself to his many passions, of which the foremost was the greatness of his country.

He was an exceptional character, who held all the trump cards. Educated, civilized, Hellenized, indefatigable, he was also a *bon vivant* who loved wine and women and who knew how to use his charm on strangers. He had the flexibility of the diplomat, the steadfastness of the soldier, and all the qualities of leadership. Patiently and methodically, he tackled his problems one by one. He began by re-establishing order in his enormous kingdom. (Macedonia was at that time the largest and most populous state in the Greek world.) With power centralized in his own hands, he made Macedonia even larger, at the expense of its neighbours, and Hellenized the lands he conquered. Towns that resisted him were destroyed, and new ones were founded in their stead. He favoured urbanization, built roads, and encouraged agriculture. His peasants made excellent soldiers—trained, disciplined, and full of admiration for their commander in chief.

Philip's primary need was money. He balanced his budget in the simplest and most effective way by conquering Mt. Pangaeum, whose gold mines produced 1,000 talents a year. He then founded the nearby town of Philippi in place of Crenides. He gave his name to the golden staters that came from the mines, for he now began to coin money, and the handsome "philipps" became an instrument of his policy. He understood and practiced the art of buying consciences.

Reorganization of the army

His closest care was reserved for the army, in which area he was assisted by his faithful general Parmenio. Compulsory service was introduced, each region furnishing its quota of men. Three recruiting areas were established, each one producing a unit of light infantry, one of heavy infantry, and one of cavalry. The tactics, based on those of Epaminondas, combined flexibility of manoeuvre with the thrust of light and heavy troops. The keystone was the famous Macedonian phalanx, bristling with the long lance known as a *sarissa*, which presented to the enemy an impregnable wall of iron. Once the opponent was checked and held by the phalanx, the mobile troops had merely to attack in the rear. Siege tactics also progressed. Engineering and artillery now contributed to the speedy capture of the besieged town.

At the head of this superb army, Philip II took part of Epirus and extended Macedonia's southern frontiers at the expense of Thessaly. To guard his western flank against the Illyrians, he had recourse to a diplomatic marriage with Olympias, the daughter of an Illyrian king. Philip also realized that Macedonia needed an outlet to the sea. As regent he had recognized Athens' right to Amphipolis, though with mental reservations. Now, as king and sure of his army, he attacked the town without hesitation. It was vital to Athens and had already seen much Athenian bloodshed. Philip took it, promised the Athenians to give it back in exchange for Pydna, and finished by keeping both Amphipolis and Pydna. Paralyzed by the Social War, the Athenians could do nothing but accept a *fait accompli*. Nor did they even intervene in time to save Olynthus or Potidaea. Olynthus, capital of Chalcidice and of the Chalcidian League, threw itself into Philip's arms, and he bought the Olynthians' loyalty at no cost to himself by making them a present of the territory of Potidaea, which he had had destroyed.

While Philip, almost as soon as he was on the throne, became master of a unified, powerful, and populous Macedonia, the Athenians remained blinded by the pacifist policy of Eubulus. Only the orator Demosthenes, now entering on his political career, realized the danger. He believed that Athens should keep at peace with Persia in order to concentrate all its forces against Macedonia. Only Athens could stop Philip. The Athenians, however, would not face the danger that grew clearer day by day in the north, despite the importance of this region for their grain supply.

The Third Sacred War. Philip bided his time, waiting for a chance to intervene in Greece. He found his pretext in the so-called Third Sacred War, between the Phocians of Delphi and a coalition of Thebans, Thessalians, and Locrians. Ostensibly fought to defend the rights of an outraged god, Apollo of Delphi, the war in fact was fed by an old and deep-seated hatred between Thebes and Phocis. But the religious character of the war threatened to engulf all Greece, instead of uniting the Greeks against a common enemy.

In 356 the Phocians had laid sacrilegious hands on the gold and silver in the treasuries of the Panhellenic sanctuary of Delphi, which lay in the middle of their territory. With this they made war, raising and maintaining an army of 10,000 men, which was for some time the best in Greece. Under energetic commanders this army dared invade Thessaly and Boeotia, with more successes than failures.

This attracted the wolf into the sheepfold. In 354 Philip had taken the town of Methone, at the foot of Mt. Olympus, and in Athens Demosthenes had realized the gravity of the danger from Macedonia. Philip, moving down into Thessaly, lost a first battle against the Phocian army but soon had his revenge, and all Thessaly came under his sway. Thebes, the enemy of Phocis, had naturally taken his side, Athens and Sparta that of Phocis.

Philip and Demosthenes

The war ranged Phocis against Thebes at the front of the stage, with Macedonia versus Athens behind them. It also ranged policies and men against each other; the great duel between Demosthenes and Philip was to last almost 30 years. It was a dramatic duel, if ever there was one, but unequal, since Philip had the support of his army, his diplomacy, his undivided aims and control over his own decisions, whereas Demosthenes was armed only with his eloquence and was surrounded by other opponents, the worst of all among his own compatriots. At Athens an age of generals had given place to an age of orators, some of whom, such as Aeschines, had been bought by Philip, while others, such as Eubulus, supported by Isocrates, did not wish to see the danger and played on the people's apathy. At each step Philip made, Demosthenes tried in vain to shake the Athenian Assembly from its inertia by a speech compounded of violence and reason, but his eloquence was powerless against the strength of the King of Macedonia.

The latter's intentions were clear, at least to those who were willing to see, the day when he crossed Thessaly, which he had annexed, and appeared before Thermopylae (353). This bold stroke produced a reaction in Greece, and in a reminiscence of the glorious days of the Persian Wars Philip found himself faced by the Phocian army, reinforced by 10,000 Greeks from Sparta and above all from Athens. With the way forward barred, Philip saw that the fruit was not yet ripe. He withdrew, patiently awaiting a better opportunity.

Meanwhile, he strengthened his position in the north and east, pushing into the Chersonese and the Hellespontine regions, reaching as far as Byzantium, where he opened negotiations designed to undermine Athenian influence. At the same time, he obliged the King of the Odrysians in Thrace to sign a treaty of alliance. With his hands free on that front, he could turn on a new prey, his supposed allies, the Chalcidians. He in fact needed to get the Greeks of Chalcidice into his power, not only because its promontories offered threefold access to the sea but also because its geographic situation made it an intolerable barrier between Philip's own possessions, Amphipolis, Philippi, and Pangaeum, to the east, and Pydna and Methone, to the west.

Hoping for assistance, the Chalcidians concluded a treaty with Athens. Demosthenes delivered his first "Philippic" (351), followed by the three "Olynthiacs" when Philip revealed his intentions on Olynthus, the most powerful of the Chalcidian cities. To stop Athens from in-

terfering, Philip in machiavellian style formented a revolt in Euboea. Athens in fact sent three expeditions to Olynthus, but it still surrendered (348). Philip razed it to the ground for daring to resist, sold the inhabitants into slavery, destroyed 32 other Chalcidian cities, and incorporated the country into Macedonia. The Athenians saw their access to the north definitely cut off, and there was nothing to stop Macedonia replacing them as a maritime power.

Philip was then clever enough to make people believe that he would not abuse his victory. Like a good player, he sheathed the sword and turned upon Athens a friendly countenance. Many Athenians, following Aeschines, did not wish to undertake the necessary effort to keep independence and preferred to see Philip as a friend and come to terms with him. But all the rest as well, whatever their party or persuasion, equally desired an agreement with Macedonia, even Demosthenes himself, who realized how weak his country was at this time.

Negotiations were therefore opened, and the two governments played out a diplomatic comedy at Athens' expense. Athens sent an embassy to Pella, made up of Philocrates and of the two rivals in eloquence and opponents in policy, Aeschines and Demosthenes. The splendid reception that the King gave them won them over, except perhaps Demosthenes, who, however, lost his wits, either from illness or emotion, and was unable to speak. Philip's charm had done its work on the others. In return, a Macedonian delegation was sent to Athens to come to precise terms. These were debated by the Assembly, which in the confusion voted for the status quo between the two states. This meant recognizing the loss of Amphipolis and Chalcidice, while at the same time the Phocians were excluded from the treaty, which allowed Philip the latitude to continue or halt the Sacred War. But the peace was not yet signed, as the King's oath was still lacking.

Thus, a second embassy, made up of the same men as the first, set off for Pella. Demosthenes had realized that Athens must bring the affair to a swift conclusion, since it was in Philip's interest for it to drag on. He was in the process of completing his conquests in Thrace, and each day his share of the status quo grew larger. He also knew how to spend his money to cause delays. The Athenian ambassadors spent a month and a half running after him. It was only in July 346, a year after Plato's death, that Philip signed the peace known as the Peace of Philocrates. He signed in the name of the Macedonians and their allies, including the Thebans, with Athens alone as the other party. Although he was now Athens' ally, Philip could still complete the conquest of Phocis, thus becoming master of Thermopylae and of the sanctuary of Delphi, from which he expelled the Phocians permanently. His political supremacy was thus strengthened by a moral and religious authority in Greek eyes, since he now controlled the oracle of Apollo and presided over the Pythian Games of 346.

The Peace of Philocrates

Such was the result of the Sacred War. Philip's triumph, which he owed to his diplomacy and his army, ensured his complete sovereignty in northern and central Greece, and at Athens itself he increased the number of his supporters. Isocrates, the teacher of rhetoric, saw in him the man of destiny capable of uniting all the Greeks and leading them in a great crusade against the King of Persia, although the danger from Persia had ceased to exist for Athens, at least compared with that from Macedonia. Demosthenes was not deceived; but he realized that Athens, weakened as it was, had no alternative but peace, unreal though it might be.

Renewal of Athenian resistance. So Athens and Philip made peace. They even became allies. But the causes that had made the two states enemies and would make them enemies again had not been removed. Philip was still eaten up by his ambition to become supreme lord of the whole of Greece. In all Greece he found not one single city to oppose him; and Demosthenes was a man without support, save that of his own intransigent patriotism.

Isolated and obstinate, Demosthenes fought on every front. In the courts he launched a case against Aeschines over the embassy that had negotiated the Peace of Philocrates, accusing him of treason. But the citizen jury, afraid of arousing Philip's anger, acquitted Aeschines. In the Assembly Demosthenes once more denounced the Macedonian peril, courageously delivering the second "Philippic" (344). He was also active on the diplomatic front, to south and north. He sought to draw the cities of the Peloponnese into a union against Macedonia and judged it essential also to erect a living barrier between Philip and Athens. For this he had to effect a great reversal of alliances, forget the grievances against Thebes, forget that Thebes had profited from the Peace of Philocrates as ally (since 354) of Macedonia, and obtain at any cost an alliance with Thebes.

Such a reversal could not be achieved in one day, and bit by bit, while waiting, the meshes of the Macedonian net tightened. Philip negotiated with the King of Persia, stirred up fresh trouble in Euboea, and extended his threatening shadow toward the coast of Ionia and, bypassing Athens, up to the Isthmus of Corinth. Demosthenes, however, stood in the breach, persuading the cities to conclude defensive alliances with Athens. With characteristic flexibility, Philip did not insist but changed his front. Turning toward the Bosporus, he reduced Thrace to a Macedonian province (341). Next, he launched a campaign against the Thracian Chersonese, Athens' last possession in the northeast. He went in fact as far as to lay siege to Byzantium, which, thanks to Demosthenes, had rejoined the Athenian alliance. This intolerable audacity finally shook the Athenians out of their apathy, and they decided to send a fleet into the Hellespont. The Peace of Philocrates was truly at an end. Demosthenes now redoubled his activities: he delivered the last two "Philippics," brought into being a Hellenic league, improved the situation of the war chest, and restored the strength of the Athenian fleet, thanks to a reform of the trierarchy system. Byzantium was saved.

In 339 a fourth sacred war, called the Amphissean War, broke out in Greece as the result of a sacrilege committed by some men of Amphissa against the Delphic god. Apollo had to be avenged. Philip accepted the bargain. Moreover, the command of the army charged with punishing the guilty Amphisseans was conferred upon him. He secretly outflanked Thermopylae and by a surprise gained possession of Elatea, in northern Boeotia. Panic seized the Athenians: Philip was three days' march from the frontier of Attica; he had only to cross Boeotia, and his army would be in front of the walls of Athens.

The Amphissean War

In the general consternation, Demosthenes kept his head. Philip had sent to Thebes envoys loaded with promises. Demosthenes got himself sent there as well and proved to the worried Thebans that Philip's advance must be stopped. His eloquence won the day: Thebes chose Athens. Their common enemy could not tolerate such an alliance, and war broke out openly between Philip and the new allies. Fortune seemed at first to favour the Greeks during the winter campaign; the armies of Thebes and Athens won two minor victories, which had no follow-up. Instead, there ensued a fierce battle in April 338 on the plain of Chaeronea, in Boeotia, where the Macedonian phalanx routed the ill-led allied troops. It was the end of Greek liberty.

Philip's hegemony over Greece. Thebes was occupied by a Macedonian garrison. Athens got off more lightly: by what is known as the Peace of Demades, Philip, no doubt influenced by what he owed to Athenian civilization, paid homage to the conquered city and to the dignity of its resistance by granting easy conditions. He compelled Athens to sign an alliance with him and took the Chersonese; but neither the city nor Attica was occupied, the Athenians kept their fleet, and their ports remained free. Demosthenes' anti-Macedonian policy may be considered to have been justified by the leniency of the peace. If Athens had listened to Aeschines, it would have been humiliated without a fight. With Demosthenes it won the respect of a conqueror who was more Greek than barbarian. Alexander would not forget the lesson in moderation his father had given him.

After his victory Philip did not halt. Master of Athens

and of central Greece, he conquered Euboea and then descended upon the Peloponnese, where he imposed his will without a fight. Greece lay at his feet.

To succeed in his program, Philip attempted to establish a permanent organization of the Greek cities under Macedonian hegemony. He summoned to Corinth, the central city, delegates from all Greece. Apart from Sparta, all accepted the invitation to this extraordinary congress. There, in 337, there came into being a league truly "Panhellenic," since it included, apart from the Spartans, all those who henceforth called themselves officially Hellenes. Philip thus realized the ideal of Isocrates, though with one major reservation, namely that the Hellenes, though united, were united only under a foreign dominion, that of the King of Macedonia, now become their *hēgemōn* and *stratēgos* and wielding executive power. The league was represented by a council, the first example of a supranational institution or league of nations, but was entirely in Philip's hands, beneath the banner of Hellenism.

The states that adhered to the League of Corinth were declared autonomous. They paid no tribute, and their military obligations were reduced, consisting only of an obligation to provide a contingent to the federal army and fleet; it was forbidden to take up arms against Greeks, since all Greeks were henceforth linked in a general peace.

The new institutions were skillfully arranged, and Philip was diplomat enough to cover his iron hand with a velvet glove. He was also realistic enough to realize that a common enemy was needed to strengthen the ties among men so diverse and to free the Asiatic Greeks. In 336 he proposed at Corinth a general crusade against the barbarian. A Macedonian army of 10,000 men set out at once as an advance guard under the orders of Parmenio and Attalus. Penetrating Ionia, it seized Ephesus. The Achaemenid Empire seemed to be nearing its end.

Assassination of Philip

But Philip in his universal victories had counted without fate. At the beginning of the summer of 336, when he was 48, during the magnificent celebration that he had organized for the wedding of his daughter Cleopatra to the king of Epirus, he fell beneath an assassin's dagger. Philip was dead, but Hellenism, on the contrary, had a chance to survive. It was not extinguished in the country of Isocrates and Demosthenes, and it vibrated in the heart of the young Alexander, Philip's son and heir. Alexander was 20 years old. He had been 18 at the battle of Chaeronea and 13 when Philip had appointed Aristotle his tutor. Aristotle's pupil was to Hellenize the east. But, on the other side of Greece, Hellenism still shone, with its defects and its virtues, among the western Greeks.

THE WESTERN GREEKS

Conflict with Carthage. After the Peloponnesian War, the cities of mainland Greece continued jockeying for supremacy and the king of Persia never hesitated to intervene. It took Philip of Macedon to stop the process. The history of Greeks and barbarians in western Greece (*i.e.*, in Magna Graecia, southern Italy, and Sicily) reflected in its record of internecine warfare and barbarian intervention that of their mother cities in Greece itself. To understand the history of the 4th century, one must go back to 480, when the Carthaginians were defeated at Himera in Sicily. At the same time, the Persians were attacking the mainland Greeks. In 413 the Athenian disaster at Syracuse again produced a similar twofold barbarian intervention against the Greeks. While the King of Persia, with his satraps in Ionia, was working against both Athens and Sparta to ensure his own domination by the exhaustion of both rivals, the Carthaginians, who had been temporarily threatened by Alcibiades' plans, now on their side, breathed new life into their own Sicilian ambitions. This was a serious threat, since Sicilian unity had not survived the victory over Athens in 413.

Hermocrates

Hermocrates, the leader of the aristocratic party at Syracuse, had emerged by luck and the help of Sparta as the brilliant architect of victory. Scarcely had he turned his back to go to the Peloponnese—there to continue the struggle against Athens—when the ungrateful Syracusans exiled him and called in the democrat Diocles. On a wider scale, the same conflict also set one city of Sicily against another. Syracuse turned on Catana, Selinus on Segesta. And the Segestans called on Carthage for help against their fellow Greeks.

Carthage did not miss this chance of avenging the insult that the petty tyrant of the petty town of Gela had inflicted on it at Himera 70 years before. In 409 a Carthaginian army under Hannibal, son of Gisco, disembarked in western Sicily. It had no difficulty in beating the Selinians, whose town was promptly sacked. Himera, threatened in turn, was helped by Diocles of Syracuse but fell to superior numbers and was razed. Hermocrates, returning home, recaptured Selinus but died at Syracuse in 407. His supporters, including a young officer named Dionysius, were exiled. The following year the Carthaginians, aiming to end all resistance of the Sicilian Greeks, sent out a new expeditionary force under Hannibal and his grandnephew Himilco. They laid siege to Acragas, where Hannibal died, but Himilco's gold bought the treachery of Sicilian generals both at Syracuse and at Acragas itself, and Himilco took the town in December 406. At this moment, a saviour arose in the person of Dionysius, the young officer who had served his apprenticeship in arms with Hermocrates.

The rise of Syracuse under Dionysius. Seizing his opportunity, Dionysius, known as Dionysius the Elder, with the help of others of Hermocrates' supporters (such as the rich Philistus and Hipparinus, the father of Plato's young friend Dion), began by denouncing the treachery of the Syracusan generals. With nine new colleagues (whom he was soon to get rid of), he was elected general. He recalled the exiles, married one of Hermocrates' daughters, and organized a plebiscite. With the assistance of a group of supporters who shared his conviction that Syracuse needed a firm government based on the army, he surrounded himself with a bodyguard, by means of which in the summer of 405 he seized absolute power.

Struggle against Carthage

To retain power, he needed to remove the Carthaginian menace. Himilco had continued the conquest of Sicily, defeating Gela and Camarina. Dionysius did not rush to their help with any great haste, perhaps to leave open the possibility of an agreement with Carthage. In his absence, however, the Syracusans rebelled and killed his wife. When he returned, he put down the rebellion and succeeded in concluding an advantageous treaty with Himilco. Half of Sicily was abandoned to Carthage, but the independence of the cities in the eastern half was guaranteed and strengthened by the Syracusan tyranny, which Dionysius was to retain for 38 years, until his death, in 367.

Dionysius' first task was to establish his power on a firm basis. He reorganized the army by introducing mercenaries and fortified the royal palace on the island of Ortygia (which almost completely shuts in the Great Harbour) and fortified the Great Harbour itself. He erected defenses for the whole town by surrounding with enormous walls the extensive plateau of Epipolae, where the fortress of Euryalus became recognized as a masterpiece of military engineering. The philosopher Plato was to stay in this palace on several occasions from 388 or 387 onward.

He needed also to ensure the succession, since he proposed to turn his tyranny into a hereditary monarchy. With this in mind he turned bigamist, marrying on the same day Hipparinus' daughter Aristomache and Doris of Locri, who was destined to produce the future heir Dionysius II.

If the elder Dionysius was almost wholly devoid of scruples, nevertheless it should be said that, on the whole, he used his power with moderation, once he felt that it was assured, after a final Syracusan revolt in 403.

As master of Syracuse, his ambitions grew. He wanted to rule all Sicily and make it, in Plato's phrase, "one single city." In this he merely anticipated on a smaller scale Philip's plan to make mainland Greece a single state. To this end he reduced Catana, Naxos, Leontini, and Messana and became, under the title of "archon of Sicily," the head of the citizens of the Sicilian Greek cities and

also the protector of the original Sicilian inhabitants, insofar as the latter were not under Carthaginian rule. In this he anticipated Alexander as well, since he did not hesitate to deport entire communities in order to realize the fusion of peoples and races.

Projects in southern Italy

Since he still had plans for southern Italy, he proposed, without giving up the struggle against Carthage, to place some of the Italian cities under the control of Locri, home of his wife Doris; with the others, he would not deprive them of their liberty but would take them under his protection by bringing them together in a league presided over by Tarentum, under the aegis of his friend Archytas. He therefore allowed the latter, who combined the role of politician with those of Pythagorean philosopher and mathematician of genius, to give Tarentum a mixed constitution that made it the leading city of southern Italy. Then, after concluding pacts with the Messapians, Campanians, and Celts, he founded the towns of Ancona and Adria on the Adriatic. He was anticipating not only the two great kings of Macedonia but also the Romans, since he would doubtless have brought about the unity of Italy if throughout his reign he had not had to cope with several Punic wars, as the Romans were also obliged to do later.

The treaty of 405 had not in fact finally put an end to the danger from the Carthaginians. Dionysius began by calling on Carthage in 398 to evacuate the Greek cities of Sicily, and after a siege he captured Motya, which was defended by Himilco's fleet. Himilco, however, took his revenge the following year by destroying Messana and defeating Dionysius in front of Catana, which enabled him to besiege Syracuse. But Dionysius in 396 regained the initiative, destroyed Himilco's fleet, and allowed him to escape to Carthage, where he committed suicide, unable to bear his defeat. Dionysius then took Catana, rebuilt Messana, and occupied Tauromenium (present-day Taormina), founded by Himilco five years earlier.

A second Punic war (383–376) went less well for him. He defeated the Carthaginians at Cabala, where Mago was killed; but Mago's son, the younger Himilco, got his revenge at Cronium. The elder Dionysius was forced to ask for peace and to allow a third of Sicily to pass once more under barbarian rule. A third Punic war, toward the end of his reign, proved indecisive. The status quo was maintained, and Syracuse and Carthage shared the rule over Sicily.

Relations with mainland Greece

The elder Dionysius did not, however, have to do exclusively with Italy and North Africa. The mainland Greeks, like those of Sicily itself, kept their eyes always fixed on this notable tyrant who dared to intrude upon their affairs. He remained faithful to the Syracusan alliance with Sparta, closely followed Persia's intervention in Greece at the time of the King's Peace (386), and sent a representative to the congress at Sparta (371). Athens tried to seduce him from his Spartan alliance and to get the help of his mercenaries. It crowned one of his tragedies, *The Ransom of Hector*, which he had had presented at Athens during the Lenaean Festival of 367. Although Lysias had spoken harshly of him at the Olympic Games in 388, Isocrates in 369 wrote him an open letter to invite him, as he was later to invite Philip, to head a crusade against Persia. After his death, Xenophon followed with the treatise known as the *Hieron*, ending with an encomium of the enlightened despot. Even Plato made his condemnation of tyranny less harsh on his account, and it was his realization of the impossibility of achieving anything in Athenian politics that made him undertake in 388/387 his first journey to Syracuse, where he hoped to find the seat of the philosopher-king or king-philosopher. There he met the tyrant's brother-in-law, the young and brilliant Dion, aged 21, on whom he was to found so many hopes. His journey ended badly, but on returning to Athens he founded the Academy and continued to exchange letters both with Archytas of Tarentum and with Dionysius the Elder.

This remarkable tyrant may have had excessive literary pretensions and thought himself a greater poet than he was. But his political achievements were considerable. He was a great statesman who made Syracuse the richest and most populous city in the world. His example had some influence on Alexander, since Dionysius accomplished in the west what Alexander was to repeat on a larger scale 40 years later in the east. Like Alexander, Dionysius followed the principle of bringing together Greeks and barbarians.

The western Greeks to the beginning of the 3rd century BC. In Dionysius the Younger, or Dionysius II, the elder Dionysius did not have a worthy successor. The younger man was cultured but had no morality, nor had he any talent, either political or military. He welcomed Plato, who had accepted Dion's invitation to visit Syracuse a second time, still haunted by his ideal of the philosopher-king. But Dion, whose crime was to have tried to keep his nephew the tyrant from a life of debauchery, was exiled, and Plato, once more disappointed by Syracuse, had to leave a court that was wholly given over to intrigue, in order to take up again his teaching at the Academy. Dion followed the courses, while keeping in touch with the leaders of the opposition at Syracuse.

The visits of Plato

Hoping to reconcile Dionysius and Dion, Plato risked a third journey to Syracuse, with the idea of demonstrating to the tyrant the primacy of the soul over the body and of creating with him the republic of his dreams. Once more disappointed, he hurried back to Athens (361/360). Three years later Dion recruited mercenaries and assembled a fleet to obtain a return by force. He received a triumphal welcome at Syracuse, besieged Ortygia, won a naval victory over Dionysius, and after various ups and downs had himself made general with full powers. He thereupon declared that he was applying Plato's principles but in practice seemed to forget somewhat that he was his pupil. He got rid of a dangerous subordinate but fell victim to his Athenian friend Callippus, his host at Athens, and a year after taking power at Syracuse he was assassinated by his own soldiers (June 354).

Once again anarchy reigned in Sicily. Callippus, although an Athenian, made himself tyrant of Syracuse for 13 months. He had to fight against Dion's family and their supporters under Hipparinus, who took the town and regained power until his death, two years later. Hipparinus' brother Nysaeus reigned for five years (until 347), when the younger Dionysius returned and regained the tyranny for two years. In 345 the Syracusan exiles appealed for help to their mother city, Corinth, and secured the despatch of the Corinthian envoy Timoleon, who in 344 forced the younger Dionysius' abdication. The work of the elder Dionysius was destroyed, and Sicily, following a course only too typically Greek, returned to its discords.

Timoleon

Timoleon, however, surprised everyone. This Corinthian was a great man of the type celebrated by Plutarch, perhaps painted better than he was by history, probably enamoured of the ancient virtues; once master of Syracuse, he tried to re-establish peace in the city and to restore Sicilian unity. The initial success of his program reawakened Carthaginian fears, resulting in a new Punic war. Timoleon was lucky enough to win a great victory near Segesta that compelled the Carthaginians to ask for peace (341). He could then give himself entirely to his work for peace and civilization. He freed the cities of Sicily, brought in new blood with 60,000 settlers, distributed land, and established a wise and moderate regime in Syracuse. After seven years, considering that Sicily no longer needed him, he signed his own abdication, respected by all (337).

Timoleon had restored Sicilian unity, but his influence had not spread to southern Italy, where trouble continued. Despite this, Hellenism made steady progress, gradually winning over the barbarians by spreading among them moral and intellectual ideas. Hellenism had also reached Africa, where Cyrene shone by its prosperity, its schools, and its trade, as well as Gaul, where the Provençal centres of Saint-Blaise, Olbia, Antipolis, and Agde maintained a flourishing trade with Greece and Sicily.

Before losing its independence, Greek Sicily had one last fling. While the Greco-Macedonian monarchies were taking shape in the east after Alexander's death, the freebooter Agathocles temporarily halted the decline of the

island, which was doomed, with southern Italy, to annexation by the Romans. In 318/317 he seized power in Syracuse, had himself given the title of king, reopened with some success the Punic War, and was summoned by Tarentum to restore order in southern Italy. He was the man to save the Greek west and to establish a Greek "Kingdom of the Two Sicilies."

But Agathocles left it to the Romans to complete the job and to rekindle in the new capital of the world the fire of Hellenism that had burned brightly but fitfully at Syracuse.

GREEK CULTURE AND SOCIETY IN THE 4TH CENTURY BC

The changing style of life of the polis. Philip's success in uniting the Greek cities under Macedonian domination was the result, above all, of subtle diplomacy linked to force of arms. What he had laid low, however, was merely a giant already shaken by the struggles that had scarcely stopped tearing apart the Greek peoples, separately or in their leagues, from the Persian Wars to Philip's own time.

Since the victories never lasted, nor did the peace, the conquerors and the conquered suffered equally from the ills that war brought about: the conscription of farmers into the army, the devastation of the countryside, the flight from the land. The peasant and small landowner tended to disappear. Men returning from war were unwilling to take up the plow and instead hired themselves out as soldiers; after living dangerously but well in times of pillage, after seeing new, wider horizons, a man lost the taste for a narrow existence on land that demanded daily toil or felt himself stifled by being shut in within city walls. Mercenary soldiering was an inevitable evil if it was the means whereby alone men could live, but it contributed to the death of both country and city.

Decline of the city

The city in itself, moreover, was leading a precarious existence. Perpetual warfare slowly but surely brought economic stagnation, and the maritime leagues no longer brought in the easy tribute of the previous century. At Athens, the rich citizens were obliged to band together in order to meet the crushing charges that taxation imposed. The poorer classes expected the state to provide for them, and the level of public morality declined. The thousands of citizens who formed the juries in the lawcourts lost the habit of working, since they were paid for performing their duties as jurymen. The number of cases multiplied, while the number of informers living on blackmail and confiscations increased. Speakers had no difficulty in getting the people to vote the re-establishment of the Theoric Fund destined to pay for the people's pleasure, and this same people henceforth received a salary for being present in the Assembly. The 4th century provides the spectacle of the slow ruin of public spirit. Perhaps the life of the city-state was already moribund at the time of Philip's conquest.

The incessant wars between Greek cities and even within cities undermined by civil strife showed clearly that no regime, whether oligarchic, such as Sparta, or democratic, such as Athens, had been able to maintain equilibrium in Greece. After the death of Socrates, guilty, despite having done his duty on the battlefield, of rising above the city walls to think of man instead of defending the democracy, historians and philosophers had clearly defined the social and political problem. For the philosopher Antisthenes, founder of Cynicism, Greek and barbarian were equal, as were slave and free citizen. Among the intellectuals of the period and among those men of action who reflected upon the historical scene in which they were acting, there was born and took shape the idea that the separateness of the city-states had had its day.

Already in 392 the Sophist Gorgias took the occasion of the Olympic Games to proclaim the need to bring about a union of all Greeks, a position later taken up also by Lysias and Isocrates. The latter had at first thought the Greeks could be united under the leadership of Athens, but the history of his times gradually converted him to the idea that Greece would give up its divisions only under the direction of one single man.

The idea of Greek unification under a strong individual was also taken up by Plato, who thought for a time that he had found his philosopher-king, in Sicily, in the person of Dionysius the Elder, then of Dionysius the Younger, then of Dion. Isocrates placed his hope in Jason of Pherae, then turned his eyes on the son of the king of Sparta, Archidamus, and then again on Dionysius the Elder, who, tyrant though he was, succeeded in uniting the Sicilian Greeks in the face of the Carthaginian peril, a union that he made more lasting than any in Greece. And while Isocrates, the intellectual pure and simple, was idealizing another prince, Evagoras of Cyprus, Xenophon, with his experience of exile, travel, and war, was painting in his *Hieron* the portrait of the perfect monarch, whom he thought he had found in flesh and blood in the person of the king of Sparta, Agesilaus. He completed the exposition of his views by showing in the *Cyropaedia*, by the example of Cyrus the Elder, what the education of a prince should be and the benefits that his people can expect as a result. Isocrates finally came back to reality, a reality that endangered his country's liberty, when he decided that he had found in Philip of Macedon the man of destiny who could bring Panhellenism out of utopia into the realm of reality.

The desire for a philosopher-king

Philip had set his own statue beside those of the gods and thought of having the Greeks worship it in the sanctuary of Olympia. Alexander, persuaded of his divine parentage, instituted the cult of the ruler and demanded that he be made a god while he was still alive.

The arts and architecture. The same effort to escape from tradition, the same tendency to break the bounds of the city in order that the individual might forget his sufferings, showed in the arts in 4th-cenutry Greece. But the main principles held firm, and the combination of old and new still produced marvels of art, closer to man and his sufferings than in the past. Artists conveyed the uncertainty of a world in transition. Men sought to escape the anguish of the fleeting present by creating an ideal world. Their attention was concentrated less on the solid past, sometimes put in doubt today, than on daily life, even among the humble and the have-nots, and on the human personality and the passions that shake it. Thus developed the taste for analysis.

Developments in sculpture

The sculptors grew away from the city and forgot the existence of the gods. They felt the influence of painting, which renders shadows, reflections, and transparencies and translates the movements of the soul, and they set themselves to study man, no longer in what is universal in him but in his fears and weaknesses. Hence, they sought to reproduce the expression of the face and of the eyes down to the finest shades of meaning, as well as interpreting gestures and bodily attitudes of suffering or emotion. Scopas expressed the pathos of existence in heads looking backward, in mouths twisted or half open, in quivering muscles. Praxiteles, more sentimental, interpreted voluptuous revery through the sweetness of a smile of invitation or resignation; his subjects no longer have the strength to hold themselves up, their bodies drooping softly and supporting themselves on a column. He was interested in languid youths and dared to bare completely the female body, even of goddesses, who appeal to the pleasures of the senses.

A similar grace, deliberately indiscrete, inspired toreutic art and still more the delightful figurines of Tanagra, little girls or young women captured in the very moment of a light thought or passing fancy, now flirtatious and laughing, now sad and melancholic, wrapped in transparent draperies or sauntering beneath a parasol.

A new movement developed in the second half of the century. Lysippus turned from women to athletes and was also interested in monarchs. Philip lured him to Macedonia, and he accompanied Alexander to Asia; he was the first court artist. But there is nothing formal about his art. He has an eye for detail, seizes the act of movement, and shows man as he is, in his trade or in history.

In architecture, novelty abounded. Theatres multiplied, and stone tiers made their appearance at Syracuse and Epidaurus, arranged in conformity with laws whose secret is lost today, to produce unequalled acoustics. In the religious sanctuaries, round buildings developed, *tholoi*,

set above a crevasse giving access to the divinities of the underworld. Religious feeling remained strong in the Greek world wherever it was the expression of a community spirit. Cities and individuals gave considerable sums to reconstruct the great Doric temple of Apollo at Delphi, destroyed in an earthquake in 373. Later, thanks to Philip and Alexander, who were perhaps more interested in propaganda than in expressing genuine religious faith, Olympia was enriched with new buildings.

Ionian architecture

It was in Ionia, however, that architecture particularly flourished; it aimed both at ornament and at grandeur. Everywhere temples were built or rebuilt—at Sardis, Priene, Miletus, Xanthus. Indeed, on the Asiatic shore of the Aegean, already the home of the Ionic order, two of the Seven Wonders of the World were created: the temple of the goddess Artemis at Ephesus and the giant Mausoleum at Halicarnassus, the resting place of Mausolus of Caria, to whose apotheosis both Scopas and Praxiteles contributed.

Literature and oratory. In literature even more than in art, Athens remained the great capital of the Greek world. Writers who were not from Athens came there to learn its lessons. The Athenian political regime offered an example of what not to imitate, but in literature, as in art, the age of masterpieces was not yet over. But the masterpieces were no longer inspired by the same spirit as in the previous century, which had seen tragedy, comedy, and history blaze into glory.

Except in the theatre, poetry, in decline, gave way to prose. As was natural at a time when man was more worried about his future than about his destiny, the output of drama continued, especially where the theatre was under the protection of tyrants, but the tragedies of this period have not been preserved. It is known, however, that they were in the same line as those of Euripides, poet of anxiety and of passion. Aristophanes continued to write but produced a new style of comedy, either plunging into utopia to escape the harshness of the times or taking the direction that was to be that of New Comedy by depicting in realistic fashion the manners of everyday life, drawn particularly from the common people, and by subtly analyzing situations rather than individual characters.

In the lawcourts and the world of politics, the spoken word extended its domain, since it was necessary to work upon the citizens in their double capacity as judges and as masters of their country's destiny. The orators stayed in the city to defend their clients in court and influence the decisions of the people but left it to appeal to Greek opinion and put forward great ideas.

Isaeus remained the specialist in inheritance taxes. Lysias, as previously noted, was not simply a speech writer for the benefit of those who could not speak in public: a panegyric at Olympia gave him the chance of addressing all the Greeks. Isocrates went still further; having made a fortune as a lawyer, he became a professor of the highest form of eloquence, that for state occasions, insisted on a broad culture for his disciples, and left his school to urge Panhellenism, first upon the Greeks and then upon Philip of Macedon, when Greece became for him more important than Athens.

Demosthenes

On the opposite side, with a higher flame of eloquence, Demosthenes continued his incessant struggle against the rulers of Macedonia. By his public speeches and political addresses, he fought obstinately in defense of Athens, of its liberty and its brilliant past, with a settled ardour and a passion of logic that finally defeated the skilled talent of Aeschines. But he was to pay with his life for his courage in opposing the power of Macedonia.

Demosthenes' life and work are one with the history of his times. There remained, however, some professional historians. None had the force or greatness of Thucydides, although one is in danger of misjudging, on the basis of fragments or mere echoes, historians who are known mostly by name only, such as Theopompus, Ephorus, or that Ctesias who was also a doctor at the court of Susa. On the other hand, the immense output of Xenophon has survived complete. An Athenian who often, though not always, admired Sparta, he left an account of 50 years of Greek history, finishing the narrative of the Peloponnesian War that Thucydides left incomplete and continuing his researches down to the Battle of Mantineia, which reduced the cities of Greece to greater chaos than ever. Full of curiosity, more percipient than he is usually given credit for, though suffering by comparison with Thucydides as a historian and with Plato as a philosopher, he stirs and interests his readers by the diversity of his talents. His troubled life took him to the crossroads of the civilizations of Asia and Greece. He appeared on the battlefields of both Asia and Greece, then gave up a soldier's life to become a country dweller, horseman, and hunter. Faithful to the memory of his master Socrates, he created both biography with his life of Agesilaus and the historical novel with his *Cyropaedia*, which foretold the coming of Alexander.

Philosophy and political thought. The philosophers in their schools also, as was natural, invited men to reflection, whether it led subsequently to action or to meditation. All the major philosophical schools were influenced by Socrates, who had directed his powerful mind to search for the good and the beautiful—*i.e.*, the true—and who had drunk hemlock with calm courage to show the justness of his principles.

In Athens, Antisthenes, founder of the Cynic school, taught that neither happiness nor unhappiness existed in themselves but rather are subjective states. The Cynics believed the happiest creatures are the savage and the animal, since they suffer from none of the prejudices burdening man. They are lucky enough to know nothing of honours, country, religion. In short, to attain true happiness, one should "lead the life of a dog." Such was the blessed state that Diogenes was to attain in his barrel.

If the Cynics are the forerunners of the Stoics, the Epicureans derive from Aristippus of Cyrene, theoretician of hedonism. For this teacher and his disciples, everything in human life goes back to the senses. Happiness lies in pleasure; thus, the greatest happiness is the greatest pleasure. The wise man, however, is not insensible to the quality of the pleasure, and it is in his interest to remain master of his passions. Regarding the love he felt for a courtesan, Aristippus summed up in a striking phrase the teaching of the Cyrenaics: "I take; I am not taken."

Plato and the Academy

Like Xenophon, Antisthenes, and Aristippus, Plato was first molded by Socrates; their master perhaps taught them in the last analysis to be men more than citizens. That does not in any way mean that Socrates' teaching was harmful from the point of view of the state. On the contrary, Plato, before turning to philosophy, had wished to serve the city, and all his life he remained haunted by the problem of politics.

Socrates' death turned Plato away from active politics. After this, he combined the experience of travel with reflection in an attempt to discover the ideal regime of a philosopher on the throne. As has been noted, he thought he had found it in Sicily. Disappointed in this, he expounded the theory of it, first in *The Republic* and later in the *Laws*. Realizing that individualism was the death of the city, he sought to locate sovereignty in an all-powerful republic, authoritarian enough to abolish classes, and later in pure reason, creator of unchangeable laws. Plato would have liked to give concrete form to his abstract theory and suffered all his life from his inability to put his ideas into practice. Nor did he have the pleasure of seeing certain of his disciples write constitutions for cities after his death.

Since the regime and policies of Athens denied him a career as a statesman, he was a philosopher, having as his aim the formation of men for political life. On his first return from Sicily, in 387, he founded the Academy, a school of political science, where the listeners worked together under the direction of the master, aiming toward an ethical and scientific training whose ultimate aim was the reform of the state. Without ever repudiating Socrates, Plato went beyond him, broadened the field of reflection, and speculated on the realm of the absolute. Surrounded by his disciples, the most brilliant young Athenians of their generation, who live in his dialogues alongside Socrates, and in the company, too, of the Soph-

ists, whom he delighted to confound, Plato sought passionately after truth in all fields, from mathematics to metaphysics. His dialectical method came to him from Socrates. It allowed him, through questions often tinged with irony, to lead his partner in the conversation by stages to the discovery of the solution. Often the dialogue resolves itself into connected exposition, and, when only imagery can transmit the subtlety of the supreme reality, Plato, the antagonist of poetry as an agent of decadence in the city, himself has recourse to the poetry of myths.

Socrates convinced Plato of the reality of virtue, and Archytas of Tarentum and the Pythagoreans showed him that the universe was founded upon the ever-new beauty of numbers. He believed that the world was a mere illusion. Thus, it is desirable to free oneself from one's mortal body and, thanks to the soul that survives it and that, by virtue of the reincarnations, remembers its past lives, to climb from illusion to belief and from belief to the world of ideas, the only reality, since it is the image of God. Borne upon the wings of love, the soul finally achieves its happiness in the divine contemplation of the beautiful, which gives it no longer belief but actual knowledge of the good and the true.

Plato goes further and much higher than Socrates, and it has been said that all subsequent philosophies are merely footnotes of his thought. He retains, however, the secret he learned from Socrates of always providing the fascinating spectacle of a thought in the process of being born. At the same time realist and mystic, he is the father of the New Academy and of Neoplatonism. He opens the way for the Fathers of the Church and for Christian thought. The first in time of his famous disciples is, however, Aristotle.

Aristotle and the Lyceum

Aristotle had less charm than Plato but is more accessible. Where Plato opened up the lofty road of the idea, Aristotle opened up the earthly track of the fact. The two, however, make one whole—unique, incomparable, and profoundly Greek—in the sense that they settle forever the two complementary aspects of man's speculation.

Aristotle was born at Stagira, in the region of Chalcidice coveted by Philip of Macedon. His life, like that of Demosthenes, spans the middle of the century, from 384 to 322. For 20 years he was Plato's disciple, until the latter's death in 348/347. Five years later he accepted Philip's offer to take charge of the education of the young Alexander, then aged 13.

Aristotle, after living in Macedonia and touring Asia, founded a school known as the Lyceum at Athens in 335. There, scholarly discussion took place while walking, whence the name of "Peripatetic" for the school. In the afternoon the lectures were public. The morning was for work, in which Aristotle directed teams of his disciples. This is the only part of Aristotle's work that survived. For only the notes, fortunately copious, taken by Aristotle's pupils give us an idea, admittedly extensive, of the body of his work.

He left to Plato the realm of mathematical speculation but took in the rest of knowledge and established an encyclopaedic body of research on a combination of observation and syllogistic logic. In his eyes concrete objects alone exist, though they are constantly changing. He classified them, both flora and fauna, by genera and species, assembling an astounding mass of observations and conclusions. For him the life of species is determined by the action of a final cause, and motion is transmitted by a chain of events whose origin goes back to a god, perfect, unique, indifferent to the world, though not totally divorced from the world, since he passes on his energy to matter whose forms he shapes.

In ethics, Aristotle is distinguished from Plato by his realism, as the latter distinguishes himself from Socrates by his idealism. Aristotle considers that virtue lies in a just mean, a sort of equilibrium that is obtained, like the pleasure of watching a tragedy in the theatre, by a purging of the passions. But, before being a moral creature, man is a political animal, and the first of all skills is political skill. Here again, though influenced by Plato, Aristotle differs from his master in not constructing an ideal city. He studies actual existing constitutions and arranges and classifies them as he classified animals and plants, following the same method that allows him, through critical observations, to work out definite laws for the great literary genres.

It was natural that a mind of this quality, this universality, should have left its stamp on Alexander; it was natural, too, that so exceptional a pupil should have shown his gratitude by sending his master money for his research and rare specimens for his collections from the depths of Asia. The Greeks soon saw that the young disciple had replaced encyclopaedic research with conquests up to the limits of the unknown, and substituted for the spirit of inquiry a determination to impose balance and order on conquered humanity.

THE CONQUESTS OF ALEXANDER THE GREAT

Relations between the Greeks and Macedonians. When, at the beginning of summer 336, Philip II of Macedon fell beneath the assassin's knife, there was an explosion of joy in Greece: the national enemy was dead. As for his successor, Alexander, the Athenians could believe that he was too young—a mere 20 years old—to follow in his father's footsteps. They knew that he was brought up on Greek and trained by Aristotle. They could also know that, before taking part in the Battle of Chaeronea at Philip's side, he had studied passionately Homer, Pindar, Herodotus, and Euripides; they could suppose that his encyclopaedic knowledge might possibly direct him toward the things of the spirit.

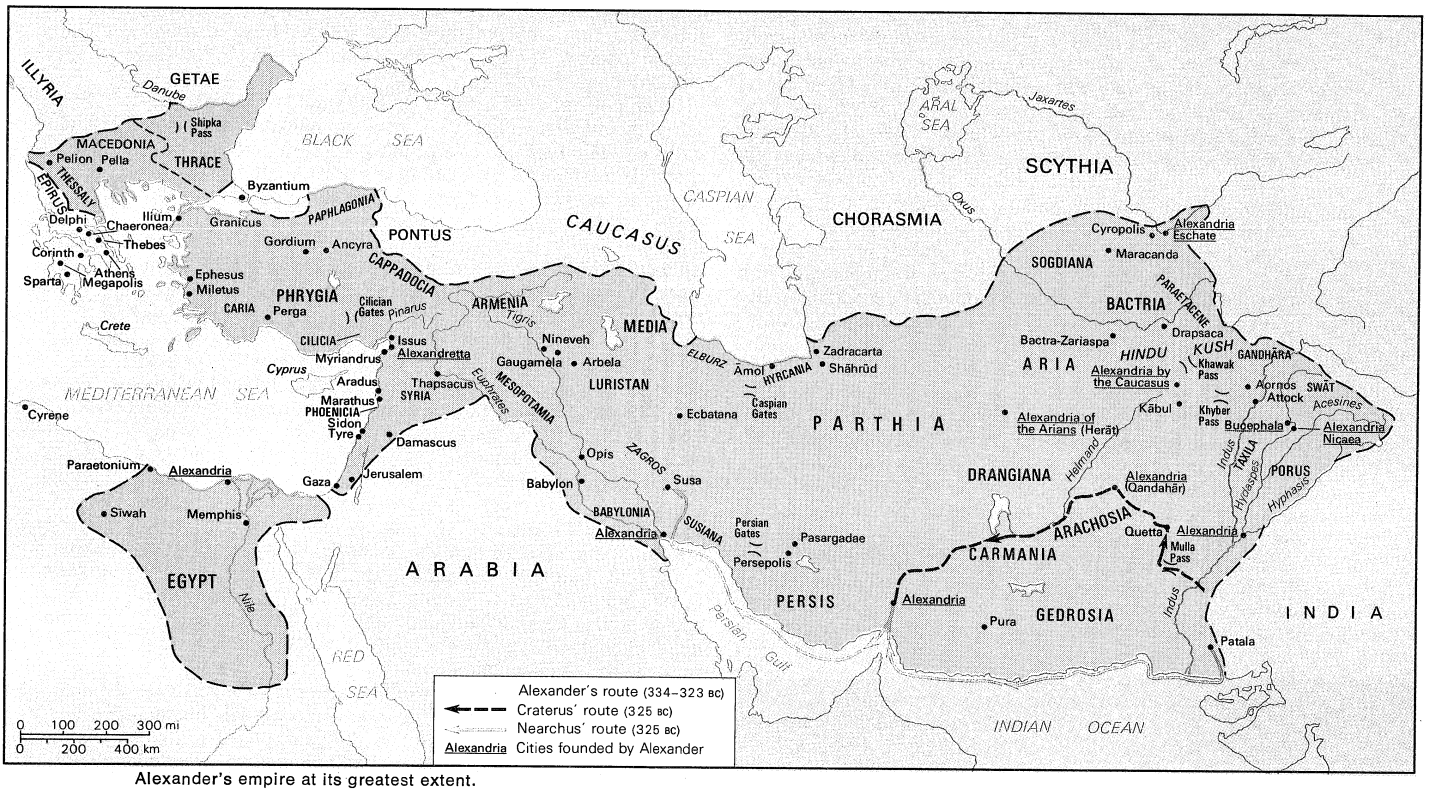
Alexander's background

Alexander had inherited from his father a taste for warfare, however, and from Homer and Herodotus he had drawn warlike lessons, as well as the example of a great struggle to be led against the barbarians of Asia. He had also listened to the advice of the orators; he knew of the crusades preached by Gorgias, Lysias, and Isocrates. He had not forgotten the march of the Ten Thousand across Asia nor the campaigns of Agesilaus; he knew that the Greeks had for too long allowed their quarrels to be settled by the barbarian king. As a Hellenized monarch, Alexander could believe he had every chance of bringing the Greeks finally into agreement beneath the dominion of someone not a barbarian. His father had shown him the way: Philip had created the League of Corinth in order to impose a general harmony under the aegis of Macedonia. Except for Sparta, all the Greek cities, members of the new league, had accepted in the spring of 337 the idea of a great war of revenge against Persia, by means of which Philip meant to cement their union.

Alexander took up Philip's program but turned it around: waging war in Asia was for Philip a means of ensuring his own sovereign authority over the Greeks; for Alexander, animated by an appetite for glory and for conquest, the war was itself the end. But, to attain such an end, he must necessarily have peace in his rear. His first act, despite his youth, was to intimidate the Greeks: he descended like a thunderbolt upon Thermopylae, haunted by the spirit of Leonidas, and won the succession to his father's seat on the Amphictyonic Council, whereupon the League of Corinth named him commander in chief of the Greeks.

A danger now became apparent on the other flank, in the north, in 335. There was trouble in Thrace, and beyond, among the Triballi and the Illyrians. Alexander, with his army, reached the northern bank of the Danube and re-established Macedonian order among the recalcitrant barbarians. Nevertheless, when a rumour spread that he had died in Illyria, the effect was seen of the embassies and of the gold that the King of Persia had sent into Greece to check the policy of Alexander's father and of Alexander himself. At Thebes the exiles returned and dared to kill the Macedonian governors. Thebes requested the support of the Arcadians, the Eleans, and, in particular, the Athenians, among whom Demosthenes was keeping alive the sacred flame of the struggle for liberty. They decided to send help secretly into Boeotia. But Alexander lost no time in showing that he was truly still alive.

From Illyria, he made a dizzy swoop into Greece: 12 days sufficed for him to appear before the walls of



Alexander's empire at its greatest extent.
Adapted from *Westermann Grosser Atlas zur Weltgeschichte*; Georg Westermann Verlag, Braunschweig

Destruction of Thebes

Thebes. The Thebans were dumbfounded, and the city fell. An example had to be made of them. The city was razed; its inhabitants, including the women and children, were slaughtered or sold into slavery. The punishment was so cruel that Alexander was struck with repentance; he made the pilgrimage to Delphi to beg Apollo's pardon and to cleanse himself in the god's presence from the stain of this bloody repression. He had commanded that naught be spared except the descendants of Pindar and the poet's own house. The terror was effective: the Greeks saw that it did not pay to take such a master lightly, even if he was flexible enough to measure out his vengeance sparingly, showing himself less severe toward Athens, a city for which he had always had a weak spot.

Now undisturbed either by barbarians or by Greeks, Alexander could turn toward Asia. He planned to make a bridge between the barbarian East and the civilization of the West, a project that presupposed replacing the Achaemenid Empire by a Macedonian monarchy. In the spring of 334 he took charge of a small but experienced army of 30,000 infantry and 5,000 cavalry and set off for the Hellespont. He had left behind his turbulent mother Olympias, who hated Antipater. But he could count on his faithful lieutenant to ensure the regency of Macedonia and the government of Greece.

The conquest of the Persian Empire (334–330). One might speculate about Alexander's aims as he crossed over from Europe into Asia. Did he intend merely to repeat the Persian Wars? to liberate from the Persian yoke the Greek cities of Ionia? to sow in the East the benefits of Hellenism? His ideas were large enough to take in all these goals, but it is certain that his sense of reality told him that he must crush Persia in order to put an end to the interventions of its arrogant diplomacy and hidden subsidies in the affairs of Europe. The hour had come for Macedonia to play the game up to the limits of the world. Alexander realized clearly what was necessary from a political point of view, but he also felt beating in his breast the heart of one of Homer's heroes; he wished to bring the *Iliad* to life.

Once he had crossed the Hellespont, Alexander made a sacrifice on the tomb of Achilles, whom he believed to be his ancestor, and then impetuously attacked the Persians. In 334, on the banks of the Granicus, after a crazy charge without a battle plan, he won a brilliant victory over the lieutenants of the Persian king. From the spoils of the barbarians he sent trophies to the Athenians that they might be consecrated in the Parthenon and adorn the Acropolis. He liberated the Greeks in the region of Troy from the King's yoke and exempted them from all the taxes imposed by his satraps. In a few months Ionia, land of the bards and fatherland of Homer, was conquered. At Sardis he restored to its inhabitants their ancient laws. At Ephesus he pacified the rival factions and gave the funds necessary to rebuild the temple of Artemis, one of the Seven Wonders of the World. Returning to northern Phrygia, he reached its capital, Gordium, on the road to Susa and there, according to legend, cut the famous Gordian knot that promised the empire of Asia to whoever should loose it. Alexander then moved south to cut Persia off from the Mediterranean, but in all the territory from the Hellespont to the Nile he wished to appear more liberator than conqueror and used force only where generosity failed.

Battle of Granicus

After experiencing serious difficulties in Cilicia in the autumn of 333, he penetrated into Syria and this time defeated the King himself in November at the Battle of Issus. Darius succeeded in escaping and reached Babylon, but he abandoned his dependents, his harem, and his family. Alexander respected his captives, and, when the Queen died in his camp as a prisoner, he ordered a funeral so magnificent that Darius considered leaving his throne to the Macedonian if he were finally defeated.

There was, however, no question of accepting a ransom or of negotiating. Alexander went his own way. Instead of making for Babylon, however, he completed the division of the empire into two halves, north and south, in order to break its naval power. Sidon opened its gates, and Tyre fell after seven months. Alexander escaped an assassin's knife and seized Gaza. Without a blow struck, Egypt fell into his hand and welcomed him as liberator. In the winter of 332/331, he offered sacrifice at Memphis, founded Alexandria, which was to be the cradle of Hellenistic culture, and advanced across North Africa toward Carthage.

Alexander next turned back toward the very heart of Persia. At Tyre he offered sacrifice to Heracles and also organized in passing a dramatic contest in the Athenian

manner. He received from Athens the homage of an embassy that landed from a state trireme, charged with obtaining the freedom of the mercenaries from Athens whom he had captured at Issus serving in the Persian army. Since Athens had just refused its support to the king of Sparta, Agis III, who was trying to subvert the Greeks against Macedonia, ruled by Antipater in the absence of its king, he granted the Athenian prisoners their liberty. After this curiously literary and diplomatic interlude, he took up arms again, crossed new deserts, and passed over the Tigris; and on October 1, 331, at Gaugamela, not far from Arbela, he crushed Darius in a third and decisive battle. At the same moment, in the Peloponnese, in front of Megalopolis, Antipater was annihilating the forces of Agis.

Battle of Gaugamela

Reassured as far as the Greeks were concerned and now conqueror on the Asiatic front, Alexander made a magnificent entry into Babylon. A whole month of festivities allowed the soldiers to relax and enjoy themselves. Victors and vanquished celebrated alike, with a feeling of relief that the Achaemenid Empire had come to an end. The complete success gave substance to the idea of one single new empire beneath a monarch who abolished frontiers to try to bring about a fusion of East and West.

Instead of following Darius, who had taken refuge in Ecbatana, Alexander advanced toward Susa, which opened its gates before him. To symbolize the end of the age-long hostility of Greeks and barbarians, Alexander adopted Persian costume and beliefs but at the cost of alienating his own men. He celebrated games, offered sacrifices, and freed the captive princesses, the mother and daughters of Darius, together with the young prince Ochus; he granted them Susa as their place of residence and had them taught Greek.

Advancing into Iran, he entered the holy cities of Persepolis and Pasargadae. He punished the resistance of the satrap Ariobarzanes by giving Persepolis over to pillage (apart from the royal palaces) and allowed his soldiers unprecedented loot. For his own part, he seized a fabulous treasure of 120,000 talents. At Pasargadae, he went to meditate before the tomb of Cyrus the Great and listened to the message of the dead conqueror. He had to adopt Eastern habits if he was to take the place of the Great King, even if it displeased part of his own army.

In the spring of 330 he organized at Persepolis an immense banquet, accompanied by music, songs, and dances. Abandon seized everyone. In a moment of exaltation, the son of Olympias, egged on by the courtesan Thais, mistress of Ptolemy, the future king of Egypt, set on fire the palace of the kings of Persia, which until that time had been spared. Darius, grown old, exhausted by his long flight, finished by falling into the hands of his satrap of Bactria, Bessus, who coveted the throne. He was put to death by the traitor; Alexander did not arrive in time to save him. He wept over the body of his enemy, had the last honours paid to him, and eight months later killed Bessus, whom he had taken prisoner. It was the end of the Achaemenids. There remained only one king of Persia, Alexander.

Death of Darius

The conquests in Bactria and the Indus Valley (330–323). Once Darius was dead, Alexander had the right to conclude that he had brought to a victorious end his mission as champion of the Panhellenic idea. He ceased to consider himself the delegate of the League of Corinth; he was no longer commander in chief of the Greeks but King of Kings, the successor of Cyrus the Great. He adopted the ceremonial of the Persian court, took a harem, and wore the tiara—a posture that led some to surmise that one Oriental despot had been exchanged for another.

Alexander went so far as to proclaim himself a god and demand from the Greeks abasement in his presence. In the army, anxiety gave place to anger. Plots were formed, and those suspected of complicity were executed, even if they were the King's old friends.

At a victory celebration in January 327 Alexander noticed Roxana, the daughter of his vanquished enemy; he eventually married her, thus realizing more than symbolic union of Greece and Asia.

There is no stopping on the high road of conquest. Mysterious India attracted Alexander because it held the traditions of his ancestral gods Dionysus and Heracles and because a pupil of Aristotle could there satisfy his curiosity as a geographer and his urge for exploration. What is more, the bearer of civilization would open up for the Greeks new commercial routes to the East, and his soldiers would restore their morale in guerilla warfare and high life alternately. The great adventure went on.

With an army of 120,000 infantry and 15,000 cavalry, he marched for 18 months until he reached the Hyphasis (Beās), an eastern tributary of the Indus. He suited his movements to the difficult country and the unknown climate; he adapted his tactics to the elephants of Porus, rajah of Lahore; he built a small fleet to sail on the Indus. In July 323 he finished off the resistance of Porus on the banks of another tributary of the Indus, the Hydaspes (Jhelum).

The final stage was to be the Ganges, at the very limits of the world, which scholars and legend said was bordered by the ocean. But his soldiers were upset by the tropical rain, his officers suffered from homesickness, and the army would go no farther. It was a bitter disappointment for the conqueror; he could not go on alone, and to the joy of the others he gave the order to return.

At the end of the autumn of 326 he launched a fleet, this time of 2,000 ships. Himself convalescing from a wound, he descended the Indus, a river that, instead of flowing into the Nile as was believed, has its own delta, as was discovered in July 325; thus, it flows into the ocean. To reach Susa, the army was divided into three sections. In the north, Craterus was to march by the land route across the country of Arachosia; in the south, by the sea route, Nearchus would sail with the fleet; Alexander chose for himself the central route, across the fearsome desert of Gedrosia, in the hope of maintaining contact with the fleet.

The northern army made its march without incident. The parallel march by the King across the desert and the voyage of Nearchus, braving a terrible sea with a river fleet along an unknown shore bristling with reefs, nearly ended in disaster. Human energy won the day. Army and fleet first made contact again, against all hope, in Carmania; their meeting was celebrated in the autumn of 325 by an astonishing orgy lasting seven days. They met for good six months later at the head of the Persian Gulf in the delta of the Euphrates. They had thus opened up the sea route linking East to West, which the great sailors of the future would follow.

The survivors of the expedition were at last reunited at Susa. Alexander decided to reward them and at the same time to crown his policy of joining the races. In February 324 he organized an astounding marriage celebration at which the sound of the trumpet orchestrated all the movements. Ten thousand mixed marriages took place at the same time between Greek soldiers and Asiatic women; all the King's lieutenants received chosen Eastern brides. He himself, who had lost a son of Roxana in India, celebrated a new marriage with the eldest daughter of Darius, the princess Barsine (Stateira).

Greco-Persian marriage at Susa

After this he could disband part of his army, those who were tired of the country; he replaced it by native troops to whom he taught the Greek language and Macedonian tactics. He then incorporated them into what remained of his old troops. He did not, however, forget Greece and wished to bring about a higher union: having, as he thought, obtained a reconciliation in Asia, he had to impose it on Greece. He signed two decrees at Susa and published them at the Olympic Games in September 324. By the first, he ordered the Greek cities to recall their exiles, and all accepted this general amnesty except Athens, where the anti-Macedonian spirit was still alive. By the second, he demanded to be accorded divine honours; the Greeks recognized him as the new Dionysus, brother of Apollo and son of Zeus. In 323, at Babylon, his sovereignty became that of a king and god; the Greek cities of the entire world sent their representatives to celebrate his apotheosis.

Alexander's ambition was insatiable. He envisaged new

explorations and studied fresh conquests, from the mouths of the true Nile to the Pillars of Hercules, Africa following upon Asia. But Nemesis lay waiting for him. His body had suffered too much from orgies, from wounds, and from the trials of war. Fever took him, increasing daily, and on June 13, 323, death struck the god who was not immortal.

The empire and achievements of Alexander. It is difficult to say what would have become of the empire of Alexander if it had been given him to reach the Ganges and finish his work. One's judgment must limit itself to what this exceptional man tried to do and to the trends that had already manifested themselves.

He did not live to see the known world united under his sovereignty, nor did he give the world a single capital. Greece stayed Greek, with its internal conflicts, beneath the rule of Antipater, Alexander's representative in Europe and governor of Macedonia. Asia remained Asian without having imposed upon it great innovations. The conqueror had the intelligence to preserve all that was workable and healthy. He was content to correct or refine the Achaemenid administration. He reorganized the satrapies but kept them in existence in Syria, in India, in Egypt: he gave each satrap a deputy, with a general as military commander. He set in order the payment of taxes, since the treasures of the kings of Persia were not inexhaustible, and issued a new coinage that covered the known world with his portrait. There was no area that did not bear the master's mark: justice functioned; circulation of money was encouraged; canals were dug and roads opened or repaired; town life and agriculture throve alike.

The master, king and god, might have chosen Babylon as capital, but he remained the pupil of Aristotle. He believed that it was the task of the West to enlighten the East, to stir up its masses, to energize its softness and delays. He protected artists, scholars, poets, and writers trained at Athens, whence came his admiration for the capital of the arts and of literature. Knowing by experience the vital strength of Greek language and civilization, he founded across the world 34 Alexandrias, future centres of Hellenism, on African or Asian soil. It is from Alexandria in Egypt and the Alexandrians that the world possesses all that Greek civilization from Homer onward had bequeathed.

Such indeed was the master plan of this young genius, and it succeeded after his death. As if he had realized that the immediate question of the succession would cause internal strife, he sought to ensure for the generations to come the survival of Hellenism. He went far beyond the utopian and superficial Hellenism preached by Isocrates, which Philip had almost brought about. If Alexander imposed his monarchy by divine right upon the world that he had made to yield or had conquered, it was not to leave side by side brothers who were still enemies or races separated by mutual hatred; it was in order to mold these divers elements into one single unit and to pour the best of Greek thought into an Eastern mold, where authority from on high would in no way derogate from the freedom of equals or inferiors.

When he married Roxana, when he gave his lieutenants Eastern princesses for wives, when he gave his soldiers Persian women, he was only setting an example or rather mildly demanding one. He meant to bring about a marriage between two civilizations, and he was certain that the fruit would have the best elements of both. Dead at the age of 33, after 12 years of action, Alexander had made history of a new sort, that of the conqueror who brings peace, who pardons the vanquished, who forces nations to accept his rule, only to help them to live and to unite them.

Ideas rule the world. Alexander passed on his ideas to the world that he had conquered. Before him, neither Sparta nor Thebes had been able to dominate Greece by arms; even Athens had not succeeded, either by force or by intelligence. With Alexander the rule of factions was over, the cities were broken, leagues were dissolved, nations cast into the melting pot, and races intermingled, and the time of great men itself was over.

While Aristotle was drawing up a catalog of the products of earth and man and was building up a systematization of the Greek world, his Macedonian pupil was saving the Greek heritage, was opening the Alexandrian Age, and was allowing Hellenism to diffuse throughout the world. (E.D.)

V. The Hellenistic Age (323–30 BC): Greece in the Roman period

THE ESTABLISHMENT OF THE HELLENISTIC KINGDOMS (323–276 BC)

Alexander's career gave rise to a new Greek world, in which the city-states played a role subordinate to territorial monarchies but in which the genius of the Greek people found fresh outlets in every field of culture and political experimentation. The three centuries from Alexander to the establishment of the Roman Empire under Augustus were among the most productive and influential in the whole of Greek history.

From Alexander's death to the death of Perdiccas, 323–321. Alexander's death (323) produced a grave political crisis over the succession; in a compromise solution the army vested power jointly in Philip's feeble-minded bastard, Arrhidaeus, who was an epileptic, and the unborn child of Alexander and Roxana, should it prove a boy (as it did). These two became Philip III and Alexander IV. Powers and provinces were apportioned. Antipater was confirmed in his control of Macedonia and the European possessions, including Greece. In Asia Perdiccas was made chiliarch (or grand vizier); Craterus was declared guardian of the kings. Ptolemy received Egypt (where Cleomenes of Naukratis had usurped power) and gained prestige by seizing Alexander's body, which was being conveyed to Macedonia, and diverting it to Alexandria. Antigonus Monophthalmus (the One-Eyed), satrap of Phrygia, received western Anatolia (Greater Phrygia, Lycia, and Pamphylia), and Lysimachus received Thrace (hitherto an appendage of Macedonia). Alexander's Greek secretary, Eumenes of Cardia, was sent to oust Ariarathes from Cappadocia and Paphlagonia. The satrap of Hellespontine Phrygia, Leonnatus, soon perished bringing reinforcements to Greece during the Lamian War. Of all these men Ptolemy, Antigonus, Eumenes, and Lysimachus were the most outstanding and played the greatest part in the struggles for power that quickly ensued.

Political crisis over succession

News of Alexander's death provoked a rising in Greece known as the Lamian War (from the siege of Lamia). The two main disaffected factions were, first, the Athenian democrats under Hyperides, who voiced the widespread resentment felt at Alexander's decree recalling exiles, and, second, a body of mercenaries back from Asia who had provoked disturbances at the mercenary market of Taenarum in Laconia and, under the Athenian Leosthenes, now brought help to Athens. Aetolia and Thessaly joined the movement, soon followed by other Greek cities. Their forces besieged Antipater in Lamia; but Leosthenes fell in battle, reinforcements reached Antipater, and the Athenian fleet was crippled in a defeat at Amorgos. In 322, under pressure from both Antipater and Craterus, Athens negotiated; in the peace it lost Oropus and had to pay an indemnity and accept a Macedonian garrison in Piraeus. Both Hyperides and Demosthenes perished in the reprisals, and an oligarchic government was established under Phocion and Demades. Greece was now bound more tightly than ever to Macedonia, without even the pretense of alliance that the League of Corinth had offered. Aetolia alone escaped because Antipater and Craterus were recalled to Asia.

In Asia, in Craterus' absence, Perdiccas had usurped guardianship of the kings and was making a bid to unite the empire under his control. Antigonus of Phrygia hated Perdiccas; and personal issues exacerbated the conflict. Though married to Antipater's daughter Nicaea, Perdiccas was tempted by the offer of the queen mother Olympias to give him her daughter Cleopatra, Alexander's sister. Perdiccas' equivocations helped provoke a coalition of hostile satraps: Antipater, Craterus, Antigonus, Lysimachus, and Ptolemy all united against him.

Coalition against Perdiccas

But war was averted when Perdiccas, invading Egypt, was assassinated (321). Ptolemy, who had had the usurper Cleomenes murdered, was now firmly in control there and strengthened by the addition of Cyrenaica to the west; he was wise enough to decline an invitation to take Perdiccas' place. Perdiccas' death brought the first period of manoeuvre to an end. It also isolated Eumenes, who was closely linked to Perdiccas; but a drive against him proved a fiasco, though it brought about Craterus' death.

The career of Antigonus Monophthalmus, 321–301. The two decades following Perdiccas' death were dominated by Antigonus' attempt to succeed him as the unifier of Alexander's empire.

The elimination of Eumenes, 321–316. A meeting of Perdiccas' opponents at Triparadisus in north Syria (321) declared Antipater guardian of the kings in his stead. Babylonia was assigned to Seleucus, one of Perdiccas' murderers. Antigonus was entrusted with the task of hounding down Eumenes, whom the allies had condemned to death; having appointed Antigonus general of Asia, Antipater returned to Europe with the kings.

During the next two years (321–319), in the drive against Eumenes, Antigonus rode roughshod over his colleagues in Asia Minor and was virtually master of the whole area. Eumenes was penned up in Nora in Cappadocia, when Antigonus learned of Antipater's death (319). Antipater had arbitrarily appointed as regent to succeed him Polyperchon, an old officer of Philip II, instead of his own son Cassander. Cassander, an able and ambitous man, promptly crossed into Asia, there to organize a coalition of Lysimachus, Antigonus, and Ptolemy against Polyperchon. Seduced by these new hopes Antigonus rashly offered to give Eumenes his satrapy back in return for his collaboration—an offer Eumenes readily accepted without any intention of keeping his word. Meanwhile, Ptolemy, on hearing of Antipater's death, had launched an attack on Syria–Phoenicia, an act less important for what it achieved than it was as an indication that he had now made both the kingdom and the strategy of the pharaohs his own.

In Europe Polyperchon, opposed by the other governors, attempted to improve Macedonian relations with the Greeks. Athens was promised the return of Oropus and Samos, and the events of the Lamian War were expunged in a formal proclamation; the oligarchies set up by Antipater were to be abolished and exiles restored. These gestures were halfhearted and short-lived. At Athens a brief spell of democracy (during which Phocion and others were executed) ended in negotiations with Cassander's troops in Piraeus as soon as it became clear that Polyperchon was losing ground.

An agreement made in 317 established the Peripatetic philosopher Demetrius of Phaleron as master of Athens, where he remained in power for the next ten years. Under his guidance Athens enjoyed peaceful and efficient government, in which some of the precepts of Aristotelian political thought were put into practice; the presence of Cassander's garrison stifled opposition. Polyperchon himself, meanwhile, had been expelled from Macedonia; having lost his fleet in a defeat by Antigonus near the Bosporus and with Cassander gaining ground in Greece, Polyperchon fell back on the Peloponnese, where he still commanded some support.

Power struggle in Macedonia

Meanwhile, events in Macedonia had ended all pretense of loyalty to the family of Alexander the Great. The quarrel between Polyperchon and the other dynasts had been reflected in the fortunes of the kings. Eurydice, the wife of Philip III, had joined Cassander and proclaimed him regent; and Polyperchon, who still held Alexander IV, now recalled Olympias from Epirus to strengthen his claims. Olympias had both Eurydice and Philip III killed; but she was herself imprisoned by Cassander shortly afterward, tried and condemned by the Macedonian army assembly, and executed. To improve his own position relative to Polyperchon, Cassander married Thessalonice, a half sister of Alexander the Great (and sometime later named the city he founded on the site of Therma after her).

Meanwhile, Polyperchon had taken the obvious step of allying himself to Eumenes and as regent had assigned to him the generalship of Asia. From 318 onward Eumenes fought several impressive and successful campaigns against his united enemies—in Asia Minor, Phoenicia (where he recovered some of Ptolemy's conquests), Babylon (which he took), and Persia. But in 316 in Persia he was betrayed by his own troops, handed over to Antigonus, tried, and executed. Over several years Eumenes had employed skill and resilience in the name of the kings (though whether indeed with the selfless loyalty alleged by the ancient sources may be questioned). His death left Antigonus free to act against the satrapy of Iran, where anarchy was rife, and to pursue on his own behalf the unity of the empire. In so doing Antigonus became the rival and hated enemy of the other generals; thus Eumenes' death marked a crucial step in his struggle to win the succession of Alexander.

From 316 to the peace of 311. Antigonus gained control of the whole area from Asia Minor to Iran, and in 315 he turned southward to expel Seleucus from Babylonia. As a refugee in Ptolemy's court, Seleucus urged action on the other generals, and an ultimatum was delivered on Antigonus in north Syria by Ptolemy, Cassander, and Lysimachus. Because Antigonus' action against Eumenes had been on behalf of them all, they called upon him to restore Babylonia to Seleucus, to abandon Syria to Ptolemy, to hand over Hellespontine Phrygia to Lysimachus, and to surrender Cappadocia and Lycia to Cassander; further, they demanded that Eumenes' treasure be shared with the rest. Rejecting these demands, Antigonus continued his annexations; soon he held all southern Syria except Tyre, and he turned thence to seize Bithynia and Caria. Like Eumenes before him, he also allied himself with Polyperchon, whom he recognized as general of the Peloponnese.

Antigonus' manifesto of 315

In 315 at Tyre, Antigonus published a manifesto recording the condemnation of Cassander by his army assembly for the "murder" of Olympias and the seizure of Alexander IV and his mother; he also published his own appointment as guardian of the surviving king, thus stressing his role as defender of the legitimate line. Further, the proclamation declared that the Greek cities should be free, autonomous, and ungarrisoned—a demand to which Monophthalmus was to remain consistently faithful; he followed it up (probably in 315–314) with the creation of a confederation of the Aegean islanders (Nesiotes), engineered a revolt of Delos from Athens (it was to remain independent until 166), and sent agents to Greece to intrigue against Cassander.

Leaving his son Demetrius to guard Phoenicia against Ptolemy, Antigonus planned to attack first Lysimachus in Thrace, then Cassander in Macedonia. But Ptolemy, after some delay caused by revolts in Cyrene and Cyprus (where he had now gained a footing), yielded to the pressure of Seleucus and, in 312, attacked and defeated Demetrius at Gaza, thus enabling Seleucus to recover Babylonia and drawing Antigonus southward. Seleucus' change of fortune, however, suggested that negotiations might profit both Antigonus and Ptolemy (who had returned to Egypt), and after some preliminaries an uneasy peace was agreed (311).

This peace between Antigonus and Cassander and Lysimachus, on the one hand, and Ptolemy, on the other, represented a compromise based on the status quo. Cassander was to remain general of Europe during Alexander IV's minority; and Antigonus, general of all Asia. Lysimachus' control of Thrace and Ptolemy's of Egypt were confirmed. Seleucus, no party to the peace, was quickly excluded from all mention (he remained at war with Antigonus for two more years); nor did the peace concern Polyperchon. A clause underwriting the liberty of the Greek cities was destined to furnish each signatory with a convenient *casus belli* against any other who could be shown to have ignored it. In about 310 Cassander removed the limiting clause on his own powers by assassinating Alexander IV and his mother. For Antigonus the peace merely afforded a breathing space before his next attempt to absorb the other four territories.

From the peace of 311 to the death of Antigonus, 311–301. The next ten years reveal a complicated story because the general alignment against Antigonus did not prevent his rivals from intriguing against each other or from making temporary accommodations with him. For both Antigonus and Ptolemy naval power was assuming increased importance; Antigonus had his eyes on Macedonia, and Ptolemy's control of Cyprus and Syria–Phoenicia turned on the possession of a navy. In 310 Ptolemy made his brother Menelaus general of Cyprus, and it was perhaps in 306 or even earlier that the Egyptian ruler entered into a long-lasting alliance with the maritime commercial city of Rhodes. Meanwhile, Seleucus recovered the upper satrapies, inflicted a defeat on Antigonus, and made a peace requiring Antigonus to surrender his claim to Iran but leaving him free to fight against the Mauryan king Candragupta in India.

About the same time, Cassander made a rapprochement with Polyperchon (who abandoned his support of Heracles, an alleged bastard of Alexander). That accord led Antigonus and Ptolemy to patch up an agreement, which may have been linked with Ptolemy's assassination of Polemaios, Antigonus' agent in Greece. The background is obscure, however, and Ptolemy's reason for this turnabout can only be surmised. The mainland Greek cities, faced with a coalition of Cassander and Polyperchon, appealed to Ptolemy, who invaded the Peloponnese in 308; but failing there, he came to terms with Cassander (though Ptolemaic garrisons remained installed in Corinth and other Greek cities). The following year (307), while Cassander was busy in Epirus, Antigonus' son Demetrius Poliorcetes sailed to Athens, seized it, and expelled Demetrius of Phaleron, who was superseded by a democracy. When Demetrius Poliorcetes seized Cyprus, however, the friendship with Ptolemy terminated. In 305 Demetrius began his famous siege of Rhodes, which lasted a year and ended in a compromise peace. Rhodes accepted the Antigonid alliance provided it entailed no disloyalty to Ptolemy; the outcome of the siege was celebrated by the Rhodian erection of the Colossus in the harbour at Rhodes and by the title of Poliorcetes (the Besieger), thenceforth borne by Demetrius.

Conquests of Demetrius Poliorcetes

The year 306 was marked by a spate of kingship claims. After Demetrius' seizure of Cyprus, Antigonus assumed the title of king for both himself and his son—a clear implication that they were kings of Alexander's empire. A year later Ptolemy also took the title—no doubt implying "king of Egypt" (where he was already pharaoh to the native Egyptians). Cassander, Lysimachus, and Seleucus quickly followed suit. In proclaiming a Seleucid era from October 312 (April 311 in the Babylonian calendar), Seleucus had already perhaps taken a step in this direction; but his regnal years date only from 305–304.

Demetrius was now master of Athens; and after gaining the Isthmus region he resuscitated the Greek confederacy, with himself and Antigonus as presidents. Its constitution, known from an inscription found at Epidaurus, owed much to Adeimantus of Lampsacus. Like the league of Philip II, it was intended primarily as a tool of Macedonian domination. (Shortly afterward Polyperchon disappears from the records and was probably dead.)

Antigonus' enemies now prepared seriously to oppose him. A new coalition of Cassander, Lysimachus, Ptolemy, and Seleucus (who had ceded India to Candragupta in exchange for a troop of invaluable war elephants) made ready to confront Monophthalmus, who recalled Demetrius from Athens. Father and son met the combined forces of Lysimachus and Seleucus (with his elephants) at Ipsus in Phrygia (301). Antigonus lost the battle, perishing in the melee; Demetrius escaped.

The Battle of Ipsus and death of Antigonus

The victors shared the spoils. Lysimachus took Asia Minor to the Taurus (except parts of Lycia and Pisidia, which became Ptolemaic, and Cilicia, which was governed for a time by Cassander's brother Pleistarchus). Seleucus declared his claim to Syria, but he did not care to press it against his benefactor Ptolemy, who held all south of Aradus and Damascus. Cassander remained satisfied with his European possessions. The battle at Ipsus marked the end of all pretense that the empire would be united; despite Lysimachus' realm astride the straits, Europe and Asia were thenceforth to go different ways.

The failure of Demetrius Poliorcetes, 301–286. Though defeated at Ipsus, Demetrius still possessed the islands, his fleet, and several cities in Greece and Asia; but he lost Athens, which had been offended by his outrageous behaviour and had swung over to Cassander as a nominally free city under the government of Lachares. Demetrius' first chance came when Ptolemy made a double marriage alliance with Lysimachus; Seleucus, in consequence, turned to Demetrius (rather than Cassander) to make a counteralliance (at Rhosus in Syria), by which Seleucus married Demetrius' daughter Stratonice and Demetrius received Cilicia. It proved short-lived, however. Attempting to take advantage of Cassander's death in 297 (or possibly 298), Demetrius started in 296 to regain a foothold in Greece and retook Athens, which he held from 294 until 288; but he lost his remaining Asian possessions—Cyprus to Ptolemy, Cilicia to Seleucus, and his Ionian cities to Lysimachus. In 294, however, he successfully invaded Macedonia, killed one of Cassander's young sons, expelled another, and was proclaimed king of Macedonia by the army. Soon afterward he celebrated his new dignity by founding Demetrias in the Gulf of Pagasae (modern Gulf of Vólos, in Greece).

Demetrius' volatile temperament was inconsistent, however, with any firm policy. In 292–291 he invaded Asia, only to be forced to return to Europe by risings in Boeotia and Aetolia stimulated by Pyrrhus, the young king of Epirus whom Ptolemy had restored to his kingdom in 298–297 to maintain pressure on the Macedonian frontier (see below *Epirus and Sicily, until their absorption by Rome*). Demetrius suppressed the Boeotian rising and in 291–290 recovered Corcyra, which Pyrrhus had received as dowry from Agathocles of Syracuse, whose daughter Lanassa he married in 303. Demetrius' conflict with Pyrrhus went on for two more years, ending with a peace in 289. From then onward Demetrius' position deteriorated. By about 287 he had lost all the Aegean islands to Ptolemy; worse still, in 288 Lysimachus and Pyrrhus had combined to fall upon Macedonia and expel Demetrius, while Ptolemy's fleet liberated Athens. Thenceforth Demetrius was a wanderer without a base. He attacked first Lysimachus, then Seleucus, who took him prisoner in 285 and put him under house arrest. Demetrius, who had become an anachronism in a world of solidifying frontiers, died from drink in 283.

The expulsion of Demetrius

The end of the struggle for succession, 286–276. Antigonus' death at Ipsus had left Lysimachus the most serious challenger to the throne of Macedonia. He now held territories on both sides of the straits, where already in 309–308 he had built the stronghold of Lysimacheia. In 288–287 he had annexed northern Macedonia (having expelled Demetrius with Pyrrhus' help); and in 285, having broken with Pyrrhus, he seized the rest of Macedonia and Thessaly—though Demetrius' son, Antigonus Gonatas, still held Demetrias, Chalcis, and Corinth, later to be known as the "fetters of Greece" and certainly the bastions of Macedonian rule. To these for a time was added Piraeus.

Lysimachus was brought down at the height of his power, thanks partly to a domestic drama. In the interest of her children, his third wife, Arsinoe, Ptolemy's daughter, persuaded Lysimachus to execute his son Agathocles as a traitor. Thereupon Agathocles' widow, Lysandra, and her brother Ptolemy Ceraunus (Ptolemy's children by an earlier marriage) stirred up Seleucus to make a bid for Macedonia. Seleucus, perhaps also encouraged by Lysimachus' governor in Pergamum, who now went over to him, invaded Asia Minor in 282; Lysimachus was defeated at Corupedium and perished.

Seleucus, the last of Alexander's generation, seized Lysimachus' possessions in Asia, crossed into Europe, and was promptly assassinated by Ptolemy Ceraunus. Seleucus' death did not threaten his realm, where his son Antiochus I already reigned as co-monarch. The deaths of Lysimachus and Seleucus thus offered the greatest opportunity to Antigonus Gonatas, who could

not, however, stop Ceraunus from being accepted as Macedonia's king. Ceraunus' reign, however, was violent and brief. To strengthen his claim to the throne, Ptolemy Ceraunus married Lysimachus' widow, Arsinoe (his own half sister); but when he murdered two of her sons, Arsinoe fled to Egypt and married her own full brother, Ptolemy II (ruled 285–246), who had succeeded their father. Soon afterward (probably in 280) Macedonia was invaded by a band of Gauls (Celts) migrating across Europe and quick to penetrate frontiers weakened by Lysimachus' defeat. Ceraunus perished at their hands, probably in January 279; a Gaulish kingdom of Tylis was established in Thrace, and one group reached Delphi, where the Aetolians brought about their destruction. Later other bands crossed into Anatolia (see below *Bithynia, Pontus, Cappadocia, and Galatia*).

Ceraunus' death was followed by short weak reigns in Macedonia and (apparently) by a period without kings at all. Gonatas, who had been weakened by his failure to check Ceraunus, faced revolts in Greece, where Sparta rose under Areus and where Argos and Megalopolis expelled his garrisons. But after a short campaign in Asia and the striking of what was to be a lasting alliance with Antiochus I (Gonatas married Phila, Antiochus' half sister), Gonatas returned to Europe to defeat the Gauls in Thrace at Lysimacheia. This success, widely publicized in the propaganda of the dynasty, enabled him to make himself master of Macedonia and Thessaly by 276; a king since 283, he had at last found a kingdom. Thus, in just less than 50 years from the death of its founder, Alexander's empire had settled down into the three territorial states that along with Pergamum (see below *Pergamum*) were to constitute the great powers of the Hellenistic world until the Romans overthrew them: the Ptolemies in Egypt, the Seleucids in the composite state based on Syria and Mesopotamia, and the Antigonids in Macedonia. Those three states were the survivors of the dynastic struggles of the successors; and each dynasty had its own special problems.

The three great powers of the Hellenistic world

THE HELLENISTIC MONARCHIES IN THE 3RD CENTURY BC

Ptolemaic Egypt. Many Greeks had entered Egypt in the 4th century as mercenaries and traders; and, after Alexander founded Alexandria, their numbers were swollen by Greeks and by half-Hellenized Thracians and Anatolians who poured in as mercenaries.

Hellenization. Settled in the few Greek cities—Alexandria, Naukratis, and Ptolemais—or on plots throughout the countryside, where as cleruchs they constituted a reserve army and there organized themselves in their own communities (*politeumata*), the Greeks built gymnasiums and wrestling schools, formed clubs, and read Greek books. This Greek minority formed the base of Ptolemaic rule. Alexandria was an exception: besides Greeks it contained Jews (only superficially Hellenized) and Egyptians; and, although it housed the palace and central administration, it was officially separate from Egypt and never a Greek city in the traditional sense. In the countryside, moreover, though encouraged by the early Ptolemies, Hellenism was a wasting asset; as the influx of immigrants declined, native pressures grew. Intermarriage produced mixed families, and the cultural traditions of uprooted Greek immigrants gradually gave way to the centuries-old social patterns of the native Egyptians. From about 200 BC onward Egyptian traditions began to prevail in all spheres. Nevertheless, so long as Ptolemaic rule endured, there remained a strong incentive to acquire the hallmark of Greek culture. Thus in Egypt as elsewhere, Hellene came to be a cultural term, valued especially because the Greeks remained the privileged class on whom the administration rested.

The character of Ptolemaic government was linked closely with foreign policy. The security of Egypt was assured by outposts in the eastern Mediterranean and Aegean and by an expensive naval supremacy. Southern Syria, seized in 301, remained Ptolemaic until 200. Cyprus, too, after some dispute with Antigonus I and Demetrius I, became permanently Egyptian. For most of the 3rd century, the Ptolemies controlled the Cyclades (and Samos to 201), much of the Asian coast from the Calycadnus to Ephesus, and much of the Hellespontine and Thracian coasts with Lesbos and Samothrace, together with outposts in the Peloponnese and Crete.

Though perhaps conceived defensively, the acquisition of the Ptolemaic Empire led to conflicts with both Antigonids and Seleucids (see below *Ptolemic policy in the Aegean and Asia Minor*); nevertheless, it brought economic gain to the early Ptolemies, who pursued a mercantilist policy, perhaps primarily in support of their costly imperial program. Essential metals, timber, and manpower were lacking in Egypt, together with many of the immigrants' basic needs. Thus the Ptolemies were led to develop a highly centralized bureaucracy, which was the most individual feature of their kingdom.

From 200 onward the bureaucracy became less efficient, and a revival of Egyptian influence occurred in many fields—in the army after 217, in the adoption of Egyptian customs, and in the predominance of the Egyptian calendar. Nevertheless, Egypt was the last of the Hellenistic states to be absorbed by Rome (see below *Roman annexation of Egypt*).

The Seleucid kingdom. The Seleucid dominions covered three distinct areas—Asia Minor (Anatolia), Mesopotamia (with Babylonia, Assyria, northern Syria, and, after Antiochus III's victory at Panium in 200, Coele Syria [Palestine]), and Iran. Their straggling and heterogeneous character created rival claims to attention, which made for weakness. Asia Minor contained a variety of states and peoples. The Seleucids controlled Hellespontine Phrygia, Phrygia proper, Lydia, inland Caria, eastern Cilicia, and for a time southern Cappadocia; several minor dynasts acknowledged them as overlords, but only the royal road to Sardis held the area together. Between the ancient trading and manufacturing districts of Mesopotamia and the Syrian and Phoenician coastal cities, which had long been exposed to Mediterranean influences, were the nomads, shepherds, or cultivators, including the Jews. In this predominantly Semitic area, with its worship of Bel, Marduk, Yahweh, and many Baalim, the Seleucids encouraged the revival of Babylonia, where the old religion and cuneiform writings underwent a renaissance. Farther eastward a mixed population of nomads, shepherds, farmers, and, in the cities of Media and Persia, traders shared an Iranian culture resistant to Hellenism. Failing there, the Seleucids gradually fell back upon the centre, where Antioch was to outlast Sardis and Babylon as the capital; but the Seleucid attempt to unite this disparate empire was an experiment not wholly fruitless.

Three distinct Seleucid regions

Hellenization: the Greek cities. The amorphous Seleucid realm was unified only by personal allegiance to the king and by the Hellenization he promoted. Alexander's foundations were supplemented both by full cities—often with dynastic names, like Antioch-on-the-Orontes, Seleucia-in-Pieria, or Laodicea-on-Sea—and by military settlements (which later sometimes achieved the status of cities), the inhabitants of which manned the phalanx. Many of the cities enjoyed nominal independence; the degree to which this was achieved in reality reflected the goodwill or weakness of the king. Antiochus III claimed full sovereignty over the Asian Greek cities; yet in 196 Miletus and Magnesia could engage in warfare. Many towns had a royal governor, and freedom often meant only freedom from tribute.

The cities and settlements Hellenized the surrounding areas, sometimes accepting non-Greeks as citizens or in corporate bodies with defined rights. Greco-Syrian law also promoted Hellenism: thus, in the 1st century, leases between men with Persian names in Kurdistan are in Greek, and, except in Babylonia, Greek was the common language, though it was rarely the native tongue east of Phrygia. Native religions were preferred to Greek; and Greek culture grew more tenuous, although its strength in the early Hellenistic period is impressive.

Government and foreign policy. The Seleucids had a double task—to reconcile the Greek settlers with the Iranians, who had previously been masters in the land, and to defend the eastern frontier against a substantial

power in Chorasmia (modern Turkistan). By 303 Seleucus I (ruled 312–281) had lost Gandhāra, eastern Arachosia, and Gedrosia to the Mauryan Candragupta; and, although Antiochus I, as co-ruler, had special responsibilities in the east, after the battle of Ipsus (301) western pressures constantly diverted the Seleucids from the east. Later dynastic conflicts and problems raised by Egypt and Pergamum prevented a consistent eastern policy; this led to Iranian irredentism and Greek disaffection.

Seleucid clashes with Egypt

The earlier Seleucid clashes with Egypt are badly documented. On Seleucus I's death (281), Ptolemy II fomented a revolt near Apamea but made peace in 279; Antiochus I (ruled 281–261) was then occupied against the Galatians introduced by Nicomedes of Bithynia, until in 275–274 he defeated them and settled them in the later Galatia. The First Syrian War (274–271) brought Ptolemy II some success, and Antiochus I had to accept the independence of Pergamum. The Second Syrian War (*c.* 260–255 or 253) brought Ionia and many lost territories to Antiochus II (ruled 261–246); it ended in his marriage to Ptolemy II's daughter Berenice and led to a Seleucid revival. But on Antiochus II's death, Ptolemy III (ruled 246–221) championed his sister Berenice, Antiochus' widow, and her young son against Antiochus' former wife, Laodice, and her son, Seleucus II (ruled 246–225). In the Third Syrian (or Laodicean) War, Ptolemy III crossed the Euphrates but without lasting effect. About this time the people later known as Parthians seized Parthyene, and Bactria seceded. Seleucus II was distracted by civil war with his brother Hierax, but he made a compromise peace (236) and soon invaded the eastern provinces, with small result. His eldest son, Seleucus III (ruled 225–223), was fully occupied against Attalus I of Pergamum. He was assassinated in 223. His brother Antiochus III (ruled 223–187) launched the Fourth Syrian War (219–216) against Ptolemy IV. Antiochus lost the Battle of Raphia (217), but an Egyptian rising that followed permanently weakened the Ptolemaic kingdom. Although between 212 and 205 Antiochus III marched as far eastward as Paropamisadae (see below *Roman entrance into Hellenistic affairs*), by the end of the 3rd century the Seleucids were more directly involved in the west than at any time since Seleucus I.

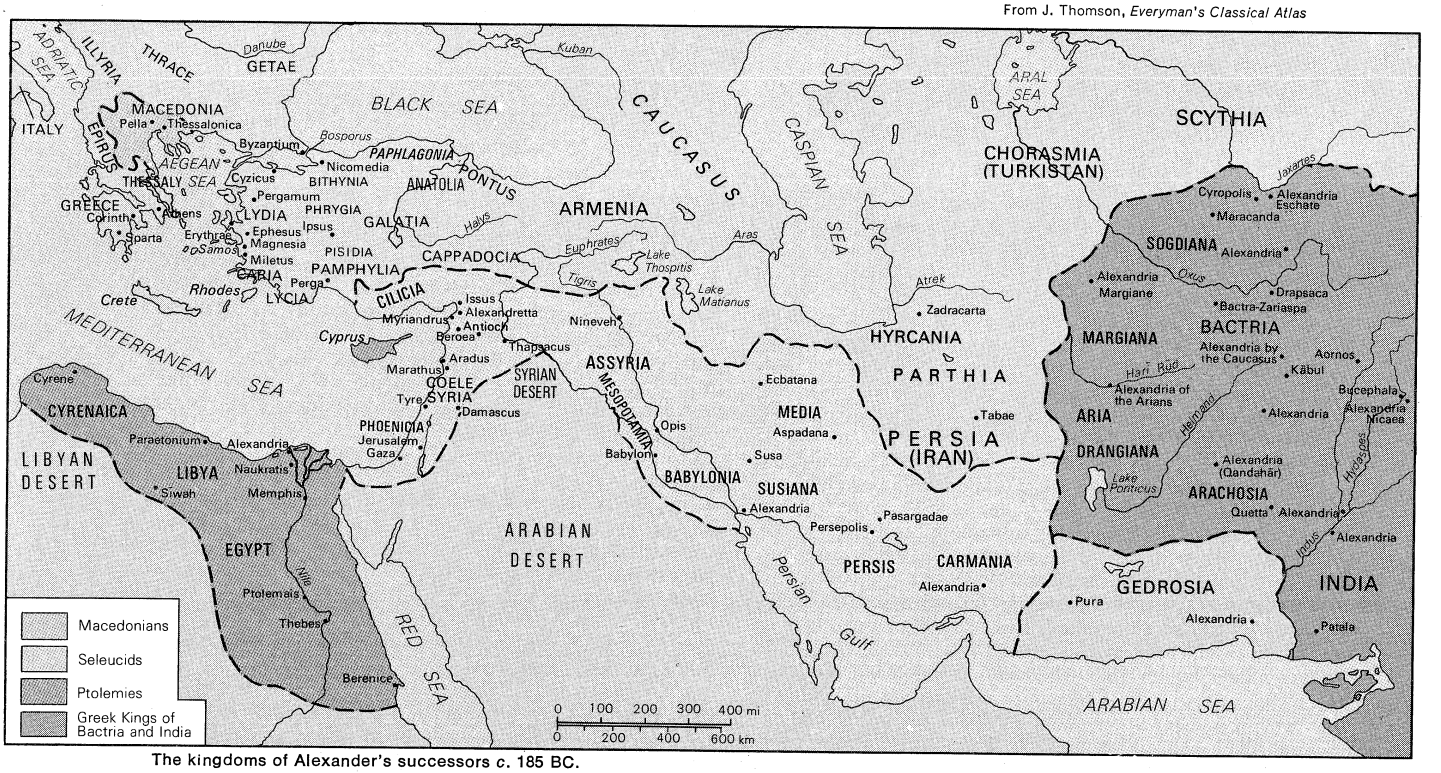
The Greeks in Bactria and India. About 239 BC Diodotus, satrap of Bactria, took advantage of Seleucid weakness to set up an independent Greco-Macedonian kingdom in Bactria, Margiana, and Sogdiana, supported by Iranians who feared a nomad invasion. His son was overthrown by Euthydemus, whom Antiochus III recognized as king after a siege of Bactra (208–206); and Euthydemus' son Demetrius annexed Aria, Arachosia, and perhaps Gedrosia and Carmania. Sometime after 187 another Demetrius of Bactria reached Gandhāra (Punjab); and in about 171 King Eucratides united all the Greek areas around Bactria. Eucratides' son Heliocles was defeated by the nomadic Yüeh-Chih around 135; but there were Greek princes on the Indus and farther eastward much later, the most famous in Greek and Indian tradition being Menander (others are known only from coins). Though inadequately documented, the story of the Greeks in the Far East exemplifies the vitality of the Hellenic tradition.

Pergamum, the smaller Anatolian kingdoms, and Rhodes. *Pergamum.* Pergamum, a stronghold on the Caïcus River in northwestern Asia Minor, was entrusted by Lysimachus to Philetaerus, who maintained good relations with the Greek world. He bequeathed Pergamum to his nephew Eumenes, who defeated Antiochus I near Sardis and declared himself independent. Eumenes' nephew, Attalus I, defeated Antiochus Hierax's Gaulish allies (*c.* 238) and took the title of *sotēr* ("saviour") and, shortly afterward, that of king. In 133 the last Attalid, Attalus III, left the kingdom to Rome; at its height, through Roman favour, Pergamum had come to embrace much of Asia Minor.

The Attalids of Pergamum

Like the Ptolemies, the Attalids zealously pursued personal gain. Their lands were rich in pitch and timber and they manufactured parchment from hides imported from the Black Sea area. They founded many cities, especially in the newer territories, and even more military colonies for their mercenaries; the cleruchs (soldiers who settled on land allotments) paid a 10 percent tax on produce, but their other subjects paid a fixed amount of corn and a heavy burden of other taxes. Pergamum itself became a splendid city with outstanding amenities and public-health provisions. The Attalids patronized learning, and their library was second only to that of Alexandria. Agricultural experiments were fostered, and industry (resting more than elsewhere upon slave labour) was developed. The Attalids posed as democratic rulers, especially in their capital; but they were widely detested by the Greeks, who regarded them as instruments of Rome.

From J. Thomson, *Everyman's Classical Atlas*



The kingdoms of Alexander's successors *c.* 185 BC.

Bithynia, Pontus, Cappadocia, and Galatia. Bithynia, astride the lower Sangarius, had eluded conquest by Alexander. In 298 or 297 its prince, Zipoetes, declared himself king. His successor Nicomedes I joined the Greek cities against Antiochus I but later brought the Gauls (Galatians) over into Asia Minor as mercenaries; in about 260 he founded Nicomedia. The later kings cultivated Greek friendship by generous gifts and founded many Greek cities, mainly on sites already inhabited. The population, largely Thracian, was substantially Hellenized. Of the later kings, Prusias I supported Rome against Antiochus III, and Prusias II played an ambiguous role in the Third Macedonian War; after a career as a noteworthy flatterer of Rome, the latter fell victim to a dynastic imbroglio (149).

Pontus, on the southern shore of the Black Sea, east of the Halys, was seized by Mithradates, the nephew of a tyrant of Cios, in 302; he made himself king and cultivated friendly relations with the Seleucids. To the south, Cappadocia, stretching from Lycaonia to the Euphrates and as far as the Taurus Range, had also remained independent under its satrap, Ariarathes; about 255 Ariarathes III became king and cemented good relations with Seleucus II. Both Pontus and Cappadocia remained feudal and mainly Iranian in outlook and custom.

The Galatians, intruders in the Greek world, were a branch of the same Celtic peoples who had invaded Italy in the early 4th century. Following their invasion of Macedonia and their attack on Delphi, a group including the Tectosages, Trocmi, and Tolistobogii (Tolistoages) was brought into Asia Minor by Nicomedes I of Bithynia in 277 and attacked Cyzicus, Erythrae, and Didyma. Settled by Antiochus I around the upper Sangarius and middle Halys, they were feared and hated by the Greeks. (Attalus I's victories over them were later commemorated in the frieze of the great altar of Zeus at Pergamum.) The Galatians kept their language, customs, and tribal organization down to the Christian era.

Rhodes. Its skilled navy and commercial strength enabled the island city of Rhodes to play an independent vigorous political role, especially in defense of the freedom of the seas. In 220 it fought Byzantium to keep the straits open, and it later made war on the Cretan pirates. In 305–304 the Rhodians successfully withstood the siege of Demetrius Poliorcetes. About 250 they defeated Ptolemy II off Ephesus in circumstances that remain obscure; generally, however, Rhodes was closely aligned with Egypt and inherited its control of the Aegean. Donations from other states to help repair the damage caused by a serious earthquake of *c.* 227 indicated Rhodes's high prestige throughout the Greek world. The richer families held power in a limited democracy and used their wealth to stave off social troubles; the 1,000,000 drachmas raised annually from import and export duties in 170 are an indication of Rhodes's wealth. After 167 Rhodes fell foul of its friends in Rome and suffered greatly in consequence. The Rhodian schools of philosophy and rhetoric were famous, and the city possessed many works of art; its maritime law may be partly incorporated in the Byzantine code known as the Rhodian Sea Law.

Macedonia: government and foreign policy. The Antigonids retained some of the traits of "heroic monarchy" evident under the Argeads. The king was the state; but the army assembly elected a regent and tried cases of treason. The monarchy was neither absolute nor personal; the king was first among his peers, neither a conqueror nor (to the Macedonians) a god. Macedonia was a poor land: the land tax raised only 200 talents annually. It was a country of peasant landowners (though the kings granted conquered land to settlers and veterans); mines and forests were royal possessions. The capital, Pella, became a Hellenized city and nominally autonomous. The upper classes spoke Greek and worshipped Greek gods. Among the few new cities were Thessalonica and Cassandrea. Antigonid power, solidified by Antigonus II Gonatas (ruled 276–239), rested on the citizen army; Macedonia avoided the problems of mercenary armies familiar to the Ptolemies and Seleucids.

The Antigonids' "heroic monarchy"

Not until the death of Pyrrhus of Epirus (272)—who had almost conquered Macedonia in 274—did Gonatas feel secure. For 50 years Macedonia tried to prevent any strong power from arising in Greece. To that end Gonatas maintained garrisons in Corinth, Demetrias, Chalcis, and Eretria; friendly governments in Argos, Megalopolis, Elis, and Megara; and a tacit agreement with Aetolia. A coalition raised by Athens and Sparta and backed by Ptolemy II took four years to suppress (266–262) in the Chremonidean War, named after Chremonides, who introduced the war measure at Athens; Gonatas' defeat of the Egyptian fleet off Cos may have occurred in 262. Ten years later, Alexander, governor of Corinth, revolted; and, although Gonatas recovered the city on Alexander's death, he soon lost it again to the Achaeans. Both he and his successors—Demetrius II, Antigonus III, Philip V, and Perseus—were much concerned with repelling the northern barbarians. But in Greece the policy of those rulers turned on their relations with the Achaean and Aetolian confederations, which were to involve Macedonia with the growing power of Rome (see below *The Greek leagues and the Antigonids in the 3rd century BC*).

The Greek states. The Hellenistic Age saw a number of changes affecting the Greeks, both on the mainland and overseas.

Social and political changes in the polis. The particularism of the *polis* underwent modification. In the mainland cities the machinery of government still functioned; but many were controlled by Macedonia, and relations with Macedonia became the main political issue alongside the rival interests of rich and poor. Elsewhere much of the content had gone out of political life. There was considerable material prosperity; the economic centre had shifted to Asia (especially to Rhodes), but Corinth, Delos, Pagasae, and Ambracia all flourished. In many ways life had improved. Attempts were made to humanize the practice of war. Many cities gained neutral status as holy places; and inscriptions record grants of immunity from the widespread custom of reprisals. Arbitration was common, and cities frequently invited outside commissions to settle lawsuits.

Freeing of slaves was frequent, some slaves buying their own freedom. Greater wealth led to new and more elaborate festivals and to an increase in social clubs; in some areas, like Boeotia, excessive social life became an abuse. Though depopulation was not serious before the 2nd century, some family limitation was practiced, probably through infanticide or abortion. In the 2nd century, however, the Roman wars brought a deterioration to most Greek cities and utter ruin to some.

The Achaean and Aetolian leagues. In the 3rd century federalism increased, thus providing compensation for the weakness of the separate cities. The Boeotian and Arcadian leagues were both important, but the most effective leagues were those in Achaea and Aetolia.

The Aetolian cantons were federally organized by the early 4th century, and after 300 the league included most of central Greece. Outlying areas enjoyed a loose form of citizen exchange (*isopoliteia*), and both Delphi and the Delphic amphictyony (association of neighbouring states) were under Aetolian control. The league had a representative council and a primary assembly meeting in spring and (at Thermum) in autumn. A cavalry leader, secretary, and other magistrates assisted the general, who was military leader; but a board of 30 *apoklētoi*—a council committee under the general—directed policy. The land was poor; this encouraged mercenary service abroad and piratical expeditions, especially as the accessible seaboard grew. In the First Macedonian War the Aetolians fought with Rome against Macedonia.

An Achaean league existed in the 5th century, but it had lapsed. In 280 Dyme, Pharae, Patras, and Tritaea reformed it; and by 272 the league of ten cities, including Aegium with the federal centre of Zeus Homarios, was complete. The accession of Dorian Sicyon under Aratus in 251 introduced an anti-Macedonian policy, which lasted until the rise of Cleomenes of Sparta (see below *The Greek leagues and the Antigonids in the 3rd century*

BC). Achaea, like Aetolia, had a primary assembly and a representative council. In 255 its two generals gave place to one, who might not serve in consecutive years; there were subsidiary magistrates, including a cavalry commander and secretary, and a ten-member board meeting under the general. The organization of the assemblies is still controversial; but apparently during the 3rd century the assembly met four times a year at Aegium, together with the council and magistrates, for routine business and elections. These meetings (*synodoi*) were restricted in the 2nd century to the council and magistrates. Special meetings (*synklētoi*) could always be summoned for special business, and after 200 they became mandatory to decide peace or alliance or to hear communications from the Roman Senate. There were federal laws and courts (as well as those of the cities) and, in the 2nd century, federal coins.

The Achaean League eventually embraced the whole Peloponnese. Its failure to strike a satisfactory compromise in its relations with Rome and an overweighting in favour of the propertied classes, who were unable or unwilling to make concessions to the increasingly radical poor within the cities, led to the league's downfall; but this does not detract from what it achieved during the 3rd and early 2nd centuries.

From W. Shepherd, *Historical Atlas*



The Aetolian and Achaean leagues.

Athens and the other Greek states. Such cities as Athens, Thebes, and Sparta found it hard to accept a diminution in scale and opportunity in the Hellenistic Age. For 60 years after Alexander's death, Athens aspired to an independent policy despite its defeat in the Lamian War (322). It had ten years of conservative administration under Demetrius of Phaleron, but after its "liberation" (307) it indulged in wild adulation of Antigonus I and Demetrius I. The Athenian and Spartan revolt against Antigonus II (the Chremonidean War) was crushed in 262; subsequently Macedonia controlled and garrisoned Athens until 229, when the governor Diogenes sold the city its freedom. Thenceforth it pursued a modest role under a conservative democracy, enjoying Ptolemaic protection. Under both the Antigonids and the restored democracy, Athens enjoyed high repute because of its past and even more as the home of the philosophical schools and the New Comedy.

Aspirations of the large cities

Thebes, destroyed by Alexander, was restored in 316 by Cassander; but its position on the fringe of Aetolian- and Macedonian-held areas restricted its political life. A supple policy, mainly based on friendship with Macedonia, kept it independent; but the Aetolians broke its military power in 248, and internally it suffered from inadequate administration, social conflict, and extravagant expenditure, foreshadowing its early-2nd-century clashes.

Sparta, consistently anti-Macedonian, warded off both Demetrius I and Pyrrhus. King Areus joined in the Chremonidean War but fell at Corinth; and under Agis IV (ruled 244–241) Sparta was briefly allied with Achaea against Aetolia. Agis was the first to attempt to deal with social ills found elsewhere but most evident at Sparta, where wealth was in the hands of the few and where a citizen body that depended on the possession of land had dwindled to a few hundred. Agis cancelled debts but was murdered before he could divide up the land as he proposed. His program was taken over and put into effect by Cleomenes III (ruled 235–222), whose imperial ambitions led the Achaean Aratus to call in Antigonus III. After Cleomenes was defeated at Sellasia and fled to Egypt, Sparta became a constant source of friction with Achaea. Revolutionary policies were revived under Machanidas and Nabis but later they degenerated into faction; and Sparta declined into a sterile imitation of its Lycurgan past, becoming a tourist centre under the Roman Empire.

Epirus and Sicily, until their absorption by Rome. By the middle of the 3rd century BC, the western Greeks were drawn irrevocably into the current of Roman history. In 317, after much faction, Agathocles, a man from Thermae, made himself tyrant at Syracuse. Engaged in a Carthaginian war, which was going poorly, he slipped out of Syracuse to land in Africa; there Ophellas of Cyrene joined him but was soon murdered by Agathocles. The war was inconclusive, but Agathocles on returning to Syracuse made a peace that established the Halycus River as the boundary with Carthage in Sicily. In 305 Agathocles took the title of king; he later campaigned in Italy and gained Corcyra, which, however, he gave to Pyrrhus of Epirus as dowry with his daughter Lanassa. Dissension in his family over succession caused Agathocles on his death in 289 to restore freedom to Syracuse. His reign had wasted Sicily by inconsequential policies and savage massacres.

The western Greeks

Pyrrhus, the Molossian king of Epirus (ruled *c.* 307–272), came nearer than anyone to bringing southern Italy and Sicily within the Hellenistic sphere. A relative of Alexander, he was perhaps the stormiest figure of his generation. Expelled from Epirus in 302, he was restored by Ptolemy I and Agathocles and set about giving a Hellenistic stamp to the monarchy of Epirus. His aim was Macedonia; and he seized part of it, with Thessaly, in 288 but was driven out by Lysimachus in 285. In 280 Tarentum invoked Pyrrhus' help against Rome; but, after victories at Heraclea (280) and Asculum (279), he crossed to Sicily, where anarchy had followed Agathocles' death. He almost drove the Carthaginians from the island; friction with the Greeks, however, led to his return to Italy and, after a drawn battle at Beneventum, to Greece. Failing in attacks on Antigonus II and Sparta, he perished in a night melee in Argos (272). Pyrrhus' policy was opportunist, and his real aims, if any, remain unfathomable; his return to Greece once more, however, severed the threads linking East and West.

At Syracuse Hieron II (ruled *c.* 270–215) seized power and took the royal title after defeating the Mamertini, Campanian mercenaries of Agathocles, who had seized Messana (Messina). The Mamertini invoked Roman aid. Hieron at first allied himself with Carthage; but in 263, after a defeat, he made a peace that initiated a period of some 50 years in which he ruled as a successful client prince of Rome. Thenceforth southern Italy and Sicily remained part of the Roman world.

RELATIONS AMONG THE HELLENISTIC STATES, 275 TO 27 BC

Ptolemaic policy in the Aegean and Asia Minor. Together with Coele Syria, the Ptolemaic overseas possessions were valuable for their wealth and formed an empire for the defense of Egypt; under Ptolemy IV (ruled 221–*c.* 204) and Ptolemy V (ruled *c.* 204–180), the possessions were mostly lost to Syria, or they either asserted or were granted independence. By 195 Egypt held only Thera and a few Cretan towns, having also lost Coele Syria, the main object of the long series of 3rd- and 2nd-century wars between Syria and Egypt.

Of these wars the first four (see above *The Seleucid*

kingdom: Government and foreign policy) left Coele Syria Egyptian; but the weakness of Ptolemy IV after the Battle of Raphia (217) and Antiochus III's experience and prestige made Syrian retaliation inevitable. By 203 Antiochus was interfering in Caria; his secret agreement with Philip V of Macedonia in 203–202 to dismember the Ptolemaic possessions after the accession of the boy king Ptolemy V probably assigned Caria to Philip. After advancing to Gaza and retreating to the sources of the Jordan, Antiochus won a victory at Panium (200), thereby gaining Coele Syria, which thenceforth was Selucid and under a general of Coele Syria and Phoenicia. In 195, having tried unsuccessfully to annex Cyprus, Antiochus made peace, marrying his daughter Cleopatra to Ptolemy V. Peace lasted 25 years, during which (in 186) Ptolemy V at last suppressed rebels who had held out in Upper Egypt since early in the previous reign. Antiochus, following on a defeat by Rome (see below *Roman entrance into Hellenistic affairs*), lost his western provinces; he was succeeded by his son Seleucus IV in 187. Seleucus was assassinated in 175 and succeeded by his brother Antiochus IV Epiphanes, who ruled until 163.

The Syrian–Egyptian wars

Antiochus had seized the throne with Pergamene help and did much to restore the kingdom. In 170 the regents of Ptolemy VI Philometor (ruled 180–145), the younger son of Ptolemy V and Cleopatra, declared war on Antiochus, hoping to recover Coele Syria. Both sides appealed to Rome; but the Senate, by then at war with Perseus of Macedonia (see below *Roman conquest of Macedonia and Greece and the acquisition of Pergamum*), delayed a decision. In 169 Antiochus invaded Egypt, besieged Alexandria, then returned home, leaving Ptolemy VI in control; Ptolemy, however, joined his sister Cleopatra II and his brother Ptolemy VIII against Antiochus and once again appealed to Rome.

The next year Antiochus seized Cyprus and again invaded Egypt. By then, however, Rome had defeated Perseus, and it ordered Antiochus to evacuate both Egypt and Cyprus. Antiochus yielded. The incident marked a decisive stage in Syrian decline; Egypt, too, was shaken, and thenceforth it could pursue an independent policy only through Roman indifference.

Syria was further weakened by a revolt of the Maccabees against the Hellenizing Judaean aristocracy, a movement that eventually ended in the establishment of a Jewish kingdom. In 150 Ptolemy VI married his daughter to the Syrian pretender, Alexander Balas; and another son of Seleucus IV, Demetrius I (ruled 162–150), was defeated and killed by Alexander Balas. Three years later Demetrius' son, Demetrius II (ruled 145–139; 129–125), landed in Cilicia to challenge Balas, whereupon Ptolemy transferred support to him in the hope of receiving Coele Syria; Demetrius defeated Balas in the Battle of Oenoparas (145), and Coele Syria remained Seleucid. In 132 Ptolemy VIII successfully raised a usurper against Demetrius II, who was tortured to death in 126 or 125; later, however, Ptolemy supported Demetrius' son, Antiochus VIII Grypus.

The years 123–83 witnessed the final dissolution of Seleucid Syria, rent with dynastic strife and prey to Jews and Arabs within and to Parthians without. Tigranes of Armenia seized the remnants of the Seleucid kingdom in 83, which by then was ripe for Roman annexation. Ptolemaic Egypt survived semi-independent for another half century.

Dissolution of Seleucid Syria

The Greek leagues and the Antigonids in the 3rd century BC. Soon after bringing Sicyon into the Achaean League, Aratus in a peacetime coup seized the Acrocorinth. Antigonus II Gonatas made no attempt to recover this position in southern Greece, but he leaned more heavily on the friendly tyrants in the Peloponnesian cities. In 241 Achaea made peace with Macedonia (and also with the Aetolians, who had recently expanded into central Greece and had been raiding the Peloponnese); nevertheless, Aratus launched a series of attacks on Argos and Athens.

Upon succeeding his father Antigonus, Demetrius II (ruled 239–229) made an alliance with Epirus, marrying the princess Phthia (Philip V's mother). This development provoked an Achaeo-Aetolian alliance, and, in 239–238, war broke out between Macedonia and the leagues. The Aetolians coveted Epirote Acarnania, and Aratus aimed at bringing all the Peloponnese under Achaea. Aratus seized Cleonae and Heraea, but he failed at Argos, where in 235 Aristomachus succeeded his brother Aristippus as tyrant. That same year, however, Lydiades resigned his tyranny at Megalopolis, which joined the Achaean League, as did Orchomenus and Mantineia. Demetrius, meanwhile, was busy fighting the Aetolians; in 237–236 he entered Boeotia and took Megara, thus driving a wedge between the leagues. In 229 Demetrius died while fighting in the north. His son, Philip (later Philip V), was only nine.

Macedonia was imperilled. The Aetolians seized Phthiotis, Thessaliotis, and Hestiaeotis; but Antigonus Doson, a nephew of Gonatas, was elected regent by the army and restored the position. In three years Antigonus Doson re-established the northern frontiers, recovered most of the Aetolian conquests, and seized part of Phocis. Doson was recognized as king (Antigonus III), but he loyally held the position in trust for Philip. In 227 he made a naval expedition to Caria, an obscure episode. Meanwhile, the monarchy in Epirus had been swept away, and Illyrian attacks caused the state to disintegrate. Ambracia and Amphilochia went with Aetolia, Athamania became independent, and Epirus itself eventually joined the Illyrians. Farther south, Demetrius' death led to the freeing of Athens (229–228), and Aristomachus surrendered his tyranny to join Achaea (228). Under Argive and Megalopolitan influence, Achaea became more hostile to Sparta, where the revolutionary Cleomenes III had succeeded in 235. Shortly before, perhaps as compensation for its greater burden in the war, Achaea had ceded Tegea, Mantineia, Caphyae, and Orchomenus to the Aetolians; in 229 they were given to Cleomenes.

Cleomenes' invasion of Megalopolitan territory initiated the Cleomenean War with Achaea (229–222), which quickly developed in his favour. In 227 he deposed the Spartan ephors, raised the citizen body to 4,000 by radical land redistribution, and introduced Macedonian tactics. Aratus, seeing the league dissolve and fearing social revolution, appealed to the Macedonian Antigonus Doson, agreeing in 225–224 to surrender Corinth in return for Macedonian help. With the Macedonian army confronting Cleomenes at the Isthmus, Argos revolted from Cleomenes and forced him to withdraw into Laconia. In 223 Cleomenes' Arcadian gains fell away; and, though in winter (223–222) he seized and destroyed Megalopolis, his paymaster Ptolemy (who had abandoned Achaea a little before) deserted him. Defeated at Sellasia (222), Cleomenes fled to Egypt, where he perished in a rising shortly afterward. Antigonus occupied Sparta, but a Dardanian invasion soon drew him back to the north; after a victory there, he died of consumption in 221.

In 224 Antigonus had created a new organization to control Greece—the Hellenic Alliance, a loose body of leagues (not cities), including Achaea, Epirus, Phocis, Boeotia, Acarnania, Opuntian Locris, Euboea, Thessaly, and (nominally) Macedonia. The Macedonian king was president; a council made all decisions, which each state had to ratify. This alliance threatened Aetolia with encirclement, and in 220 the Achaeans clashed with Aetolians operating in Messenia. The ensuing Social War was essentially a conflict between Achaea and Aetolia, with Antigonus' successor Philip V (ruled 221–179) and the alliance playing reluctant parts. Elis and Sparta joined Aetolia, and various Cretan cities aligned themselves on the two sides for domestic reasons. An impressive list of demands was presented to Aetolia. In the three years of fighting Philip revealed tactical skill and personal energy —crushing internal opposition among his ministers and sacking Thermum. But when peace was concluded at Naupactus (217), both sides kept what they held.

The Hellenic Alliance

Already Philip's eyes were elsewhere. In 221 the Illyrian chieftains, Demetrius of Pharos and Scerdilaidas, had broken the terms imposed by Rome in 228 by sailing southward on a piratical foray. In 219 a speedy policing

expedition (the Second Illyrian War) had crushed Demetrius, who, as an exile at Philip's court, now urged him to pursue Macedonian (and Illyrian) interests in the Adriatic.

Roman entrance into Hellenistic affairs. Through the Hellenic Alliance, most of Greece was drawn into a new conflict between Macedonia and Rome. At first Philip enjoyed mixed fortunes in Illyria: in 215 he approached Hannibal of Carthage, and a treaty was made in very general terms to secure Philip's Illyrian aims. In 216–212 Philip suffered a naval setback but reached the Adriatic coast by land; in Messenia his inept policy alienated Achaea. In 211 the Romans engineered an Aetolian alliance, under which the Aetolians were to have conquered territory while the plunder went to the Romans or was shared. Soon Elis, Sparta, and Messenia came in on the Aetolian side, along with Attalus of Pergamum (to whom the Aetolians sold Aegina, taken from Achaea). In 209 and 208 Philip was characteristically active defending his allies, and Attalus retired to Asia to meet the attack of Prusias of Bithynia. Neutral attempts at mediation—by Egypt, Rhodes, Chios, and Athens—were fruitless. In 207 and 206 Roman inactivity allowed Philip to invade Aetolia and force it to a separate peace; meanwhile, the Achaean army, reorganized by Philopoemen, had destroyed the Spartan tyrant Machanidas at Mantineia (207). The Peace of Phoenice, made with Rome in 205, represented a limited success for Philip, who then turned his attention to the Aegean.

Simultaneously Antiochus III was looking in the same direction. After Raphia (217) he had spent three years crushing the revolt of his cousin Achaeus and recovering some of his territory in Asia Minor. He then led his forces eastward, covering Media, Parthia, Bactria, Paropamisadae, and Arachosia and ending at Gerrha in Arabia. Antiochus thereby won wealth and the title of great and reasserted Seleucid authority in Media. His treaties with rulers farther eastward, however, merely disguised their virtual independence. In 204–203 he took Amyzon in Caria; but Ptolemy IV's death and the accession of a boy, Ptolemy V Epiphanes, changed the whole situation.

In winter 203–202 Philip, who had already been treacherously attacking Rhodes and raising money by piracy, compacted with Antiochus to divide the overseas possessions of Ptolemy; and in 200 Antiochus recovered Coele Syria (see above *Ptolemaic policy in the Aegean and Asia Minor*). In 202 Philip campaigned in Thrace, seizing the Aetolian possessions of Lysimacheia, Chalcedon, and Cios; in 201 he sailed a new fleet across the Aegean, attacking Rhodes and Pergamum and campaigning in Caria. Attalus and Rhodes appealed to Rome, and in 200 a Roman embassy declared war on Philip as he was besieging Abydos. Roman motives have been variously assessed: Philhellenism, fear of Philip, fear of Philip and Antiochus together (in the light of exaggerated accounts of their pact), and pure imperialism have all been suggested. Probably motives were mixed; but irrational alarm undoubtedly figured prominently among them.

Roman declaration of war against Philip

Though isolated, Philip resisted for three years (200–197). In 198 the Roman Titus Quinctius Flamininus defeated him on the Aous in Illyria, forcing him back into Macedonia. Aetolia and Achaea (with some opposition from Argos) joined the Romans. There were negotiations in winter 198–197, but, once assured that his command would be renewed, Flamininus made sure they collapsed. In June 197 Philip was decisively beaten at Cynoscephalae in Thessaly, the Aetolians contributing greatly to the victory.

The Roman demands had been gradually stepped up until they included the general requirement that all Greeks be free; the proclamation of this policy at the Isthmian Games of 196 aroused tremendous excitement. Philip was restricted to Macedonia; much of Thessaly was organized in federations; Aetolia received some territories but emerged discontented. The Greeks were freed; but the Romans expected their willing collaboration (failure to appreciate this was later to cause disaster). Nabis, the tyrant of Sparta, was obliged to free Argos (made over to him by Philip) but kept his throne. In 194 the legions left Greece.

Antiochus, meanwhile, had reached the Hellespont (the Rhodians having withdrawn opposition upon the news of Cynoscephalae). Smyrna and Lampsacus appealed to Rome, and Antiochus was ordered to withdraw. But when at a conference at Lysimacheia (summer 196) the Romans bade him refrain from attacking Greek cities or Ptolemy's possessions, Antiochus parried this with a pertinent reference to the Greeks in Italy and announced his forthcoming alliance with Ptolemy (see above *Ptolemaic policy in the Aegean and Asia Minor*). In 195 Hannibal fled to Antiochus' court; and in 193, through a Syrian embassy at Rome, Antiochus was ordered to stay out of Europe or see the Romans intervene in Asia. Probably neither side sought war. But in 192 the Aetolians tried to raise a coalition against Rome, and, though Philip wisely declined and Nabis was soon assassinated, Antiochus felt he must seize the opportunity; in autumn 192 he crossed, inadequately prepared, to Greece.

Elis, Boeotia, and Amynander of Athamania joined Antiochus; but Achaea and Philip both fought alongside Rome. In 191 Antiochus was defeated at Thermopylae and returned to Asia. The Romans granted the Aetolians a truce, which was renewed in 190 when Lucius Scipio as consul, with his brother Africanus as his legatus, crossed to Greece and marched northward through Macedonia to the Hellespont. Three naval victories—at Corycus (191) and at Side and Myonnesus (190)—gave Rome's allies, including Rhodes and Pergamum, command of the seas. Antiochus evacuated Thrace; and in winter 190–189 his army was defeated by the smaller but more disciplined Roman force at Magnesia-by-Sipylus (Magnesia ad Sipylum).

In 189 the Aetolians were given terms—the first recorded example of a treaty requiring the other party to "preserve the majesty of Rome." In Asia the new consul, Manlius Vulso, turned a nominally punitive expedition against the Galatians into a large-scale plundering of Asia Minor. The Peace of Apamea (188) fixed Antiochus' boundary at the Taurus Range and limited his fleet and armaments. Excluded from Asia Minor (except eastern Cilicia), his realm was thenceforth firmly centred in Syria. The program of liberation was necessarily compromised. Cities that had been free or had abandoned Antiochus before Magnesia remained free; the rest, along with large parts of Asia Minor (excluding Lycia, Caria, and the Maeander Valley, which went to Rhodes), were assigned to Eumenes II of Pergamum, the main beneficiary of the peace.

Roman conquest of Macedonia and Greece and the acquisition of Pergamum. Despite Philip's discontent with the outcome of the peace, his reorganization of Macedonia between 188 and his death in 179 (he built up the royal revenues by taxes, the encouragement of trade, and the exploitation of mines) was not necessarily with a view toward war on Rome, as some sources suggest. Constant complaints to Rome from the Greeks, especially in Pergamum and Thessaly, resulted in decisions invariably to Philip's disadvantage. He was obliged to relinquish Aenus and Maroneia in Thrace, which he had seized. From 184 onward, he campaigned and intrigued in Thrace, but his plan to raise the Bastarnians against the Dardanians fell through upon his death in 179. On succeeding, his son Perseus (ruled 179–168) revised Philip's policy and cultivated friendly relations with the Greek states and the amphictyonic council and made marriage alliances with the Seleucids and with Prusias of Bithynia. The Seleucid alliance was upset, however, in 176–175, when Perseus' enemy Eumenes sponsored Antiochus IV as Seleucus IV's successor (see above *Ptolemaic policy in the Aegean and Asia Minor*). In 172 Eumenes visited Rome and accused Perseus of hostility to Rome; the Romans resolved to eliminate Perseus, perhaps fearing a repetition of the situation of 201 because Antiochus IV was beginning to loom large in the Middle East. The Third Macedonian War (172–168) began in the winter of 172–171; in 171 a Roman force crossed over to Greece. For three years the Romans made little headway.

End of the Antigonid monarchy

Generally sympathy was with Perseus, and Epirus and the Illyrians under Genthius joined him; both Rhodes and Eumenes himself appear to have been compromised. In June 168, however, Lucius Aemilius Paullus defeated Perseus at Pydna; and the Antigonid monarchy came to an end.

Macedonia was divided into four separate tribute-paying (if nominally independent) republics with capitals at Amphipolis, Thessalonica, Pella, and Heraclea Lyncestis. Athens was rewarded for its loyalty by the gift of Lemnos, Imbros, Scyros, Delos, and Haliartus. But Epirus was savagely crushed, its population enslaved as an example to others.

Greece was disturbed and uneasy after this war. In Achaea from 189 onward there had been constant trouble with Sparta (brought into the league in 188) and Messenia (where a revolt was crushed in 184). Roman indecision and occasional malevolence heightened the general feeling of unease. After Perseus' defeat, pro-Roman politicians like Callicrates gained power, and 1,000 leading citizens were deported to Rome and kept there until 151–150; they included the historian Polybius. Despite this, internal problems continued to arise, and at each step the contenders appealed to Rome.

In Macedonia the new republics were popular only with the richer classes. In 149 a pretender, Andriscus, took the name of Philip and carried Macedonia before him. He destroyed one legion (148) but was defeated and sent to Rome by Metellus. Macedonia was then made a province, and its proconsul was given responsibility also for Illyria. About the same time general political and social discontent precipitated revolt in Achaea. Further trouble arose over Sparta; and Diaeus, a popular Achaean leader, adopted a truculent attitude toward Rome. In 147 Lucius Aurelius Orestes was sent out to Greece and demanded that Sparta and also Corinth, Argos, Orchomenus, and Heracleia-in-Trachis be made independent of the league. Further envoys found the Achaeans stubborn and bent on war with Sparta. In 146 a general conflict flared up, including Boeotia, Phocis, and Locris; but Metellus defeated the Achaean force in central Greece, and Lucius Mummius sacked Corinth as an act of deliberate punitive terrorism. The Achaean League was temporarily dissolved and, together with those states that had supported it, was put under the control of the proconsul of Macedonia. Rome had now firmly established itself in the Balkans, and Greek political freedom was at an end.

In Asia the war with Perseus had left Eumenes II under a cloud. The Galatians and Prusias II of Bithynia had exploited this fact by attacking him; but he was succeeded by Attalus II (ruled 159–138), his brother, who recovered the goodwill of Rome. Prusias was forced to desist from war in 156, and Pergamum grew in prosperity. Attalus III (ruled 138–133), a son of Eumenes II, succeded his uncle; and on his premature death for unknown reasons left his kingdom (apart from Pergamum itself and its territory) to the Roman people. Attalus III's death precipitated a civil war because Aristonicus, a bastard son of Eumenes II, claimed the throne. Pergamum and most of the cities supported Rome; Aristonicus, with some following among the Greco-Macedonian population, was mainly supported by the rustic and slave elements. He cancelled debts, freed slaves, and set up a centre called Heliopolis in Mysia, the significance of which has been much debated: he may have been trying to establish some kind of mystical utopia; at least it served as a means to rally supporters.

The disappearance of Pergamum

In 130 the rebels were defeated and Aristonicus was taken prisoner. In 129 Manius Aquilius and a commission settled the fate of the Pergamene territories. Many fringe lands in the interior were handed over to such other kingdoms as Pontus and Cappadocia. The Thracian Chersonese and Aegina were placed under the proconsul of Macedonia. But the heart of the kingdom—Mysia and the Troad, Lydia, southwest Phrygia, and perhaps parts of Caria—became the Roman province of Asia. The disappearance of Pergamum at the very moment that the Seleucid kingdom had just lost Mesopotamia to the Parthians under Mithradates I (who had taken Demetrius II prisoner in 140–139) marked a further stage in the disintegration of the Hellenistic world and the transformation of the Greeks from allies to Roman provincials.

Rome and Mithradates VI. The kingdom of Pontus had gained some prominence early in the 2nd century, when Pharnaces had attempted to expand but had been forced to surrender most of his gains in a peace underwritten by Rome (182–179). Mithradates V Euergetes (ruled *c.* 150–*c.* 120) was given part of Phrygia for helping the Romans against Aristonicus, but upon Mithradates' assassination the Romans revoked the gift—a source of bitterness to his successor, Mithradates VI Eupator (died 63 BC), who became effectively king of Pontus in 112. After gaining the realm of Bosporus, north of the Euxine, and expanding his territories eastward to Colchis and Trapezus, Mithradates VI agreed with Nicomedes III of Bithynia to attack Paphlagonia and Cappadocia. Rome intervened, compelled the evacuation of both countries, and placed Ariobarzanes on the Cappadocian throne (95). After being expelled and restored by Sulla (92), Ariobarzanes soon became a refugee at Rome along with Nicomedes IV of Bithynia, whom Mithradates then replaced with a puppet. A Roman commission restored both. Later, when the Romans provoked an invasion of Pontus, Mithradates decided on war (88). Asia was seething with hatred toward Rome because of its oppressive system of tax collection, both in the cities and the countryside. Mithradates was therefore welcomed everywhere. Three Roman armies were overthrown, and from Ephesus Mithradates organized a massacre of 80,000 Italians. His fleet crossed the Aegean to Athens, which had joined him; in Asia, only Rhodes resisted him.

The Roman general Sulla recovered Athens in 87–86, and, after two Roman victories at Chaeronea and Orchomenus, the war in Europe was over (86). A second Roman army, sent out by the supporters of Sulla's rival, Marius, was successful in Bithynia; and in 85 Sulla made an agreement of doubtful validity with the Pontic general Archelaus—the Peace of Dardanus, by which Mithradates surrendered all his gains since 88. Nicomedes and Ariobarzanes were restored, and collaborating cities in Asia were subjected; Rhodes was rewarded with additional territory. Effectively, this was the end of any freedom for the Greek cities of Asia.

Between 84 and 74 Mithradates was subject to some provocation from Roman officers; after Sulla's death the Romans broke off negotiations with him. They were increasingly concerned, however, with the problem of piracy in the Mediterranean; its centre in Cilicia Trachea was excluded from the province of Cilicia, which was set up in 80. Hostilities with Mithradates broke out again in 73, when he invaded Bithynia. Between 73 and 70 Lucius Lucullus defeated him, took over Pontus, and carried out a humane reorganization of Asia (thereby antagonizing financial circles at Rome). In 69–68 Lucullus advanced into Armenia, where Tigranes had given asylum to Mithradates; but his army refused to go farther, and by 67 Mithradates was back in Pontus, and Lucullus was without a command.

The arrival of Pompey

Intrigues at Rome had brought about the appointment of Gnaeus Pompeius (Pompey), first, to act against the pirates—which he did in 67—and then to take over the Mithradatic War with powers far greater than those of Lucullus. In 66 Pompey marched from Cilicia, where he had ended the piratical war, over Taurus to occupy Cappadocia; and by the end of the summer he held Pontus. Mithradates fled to Colchis and later to the Crimea, where he committed suicide in 63. Pompey meanwhile settled Asia Minor, now consisting of four Roman provinces—Asia, Cilicia, Bithynia, and Pontus—and a number of client princes, notably in Cappadocia, Galatia, and Paphlagonia. In 66–65 Tigranes submitted, and Armenia became a client state. After campaigning in the Caucasus, Pompey in 64 entered Syria, which was made a province; evidently the decision had already been taken at Rome. A last Seleucid, Antiochus XIII, briefly restored by Lucullus, was deposed and murdered by the Arab shaykh Sampsiceramus of Emesa.

In Judaea, where the high priesthood had developed royal characteristics and civil war over succession had been in progress since 76, Pompey gave his support to John Hyrcanus II and expelled his younger and more vigorous brother, Aristobulus; Judaea, however, became tributary, and Hyrcanus was styled ethnarch. Pompey's settlement in this region was on similar lines to that in Anatolia. Commagene took its place beside Armenia, Galatia, and Cappadocia as part of the cordon of client kingdoms between Roman territory and Parthia. Tarcondimotus of Kastabala in Amanus covered the communications between the provinces of Cilicia and Syria; and Judaea was a buffer to the south, with the Ituraean and Nabataean kingdoms flanking it.

This settlement—the first to be carried out by a Roman general without a senatorial commission—brought Pompey great prestige and a vast patronage, to which he was later to turn in his war with Caesar. It also marked the real end of the Hellenistic order. The contrast was thenceforth between a Mediterranean world, under Rome, and the Iranian East including Mesopotamia. The one exception was the still nominally independent kingdom of the Ptolemies in Egypt.

Roman annexation of Egypt. Upon the death of Ptolemy VIII (Euergetes II) in 116, the kingdom was divided between his three sons—Ptolemy IX Philometor (Soter II), who took Egypt; Ptolemy X Alexander, who took Cyprus; and a bastard, Ptolemy Apion, who took Cyrene. In 96 Apion died, leaving Cyrene to Rome (but it was made into a quaestorian province only in 74); and in 88 Ptolemy IX murdered Ptolemy X and reunited Egypt and Cyprus. During the Mithradatic War he pursued a cautious neutral policy, in view of the fact that Mithradates had gained possession of his two sons, who were in Cos when the Pontic king seized it. On the death of Ptolemy IX in 81, Sulla sent the son of Ptolemy X (who had escaped to Rome) to succeed him; as Ptolemy XI his reign was short (19 days) and important mainly for his will authorizing the Romans to annex Egypt and Cyprus when they wished.

Ptolemy IX's sons, however, soon appeared at Alexandria. The elder, Ptolemy XII—usually known as Auletes, the Fluteplayer—spent 20 years and vast sums of money in bribes to get his position sanctioned at Rome. Thenceforth the main importance of Egypt was as a factor in the internal political rivalries at Rome. In 64–63 Ptolemy sent an embassy to Pompey at Damascus; and at Rome Cicero prevented the passing of a land law that would have included the annexation of Egypt (in the interest of Crassus and the commercial class). At last in 59 Caesar was won over, and Ptolemy XII was recognized as king; Cyprus, however, was detached from Egypt and made into an independent province. In 58, a rising in Egypt caused the expulsion of Ptolemy, who was restored forcibly (for a large sum of money) by Aulus Gabinius, the governor of Syria, in 55. To achieve the restoration, Egypt was mulcted, its people were harassed, and many were murdered.

The succession of Cleopatra VII

On his death (51) Ptolemy Auletes left his kingdom to his children Ptolemy XIII (aged ten) and the famous Cleopatra VII (aged 17), as consorts. The ministers Pothinus and Achillas set the brother against his sister; and during her flight Pompey arrived in Egypt as a fugitive after his defeat at Pharsalus and was assassinated. Upon Caesar's arrival, Cleopatra returned to Alexandria and gained his support. In the ensuing civil war Caesar was hard pressed but eventually victorious, and Ptolemy XIII perished in 47. Cleopatra's sister Arsinoe later appeared in Caesar's triumph. Cleopatra, who was married to a second brother, Ptolemy XIV, went with him to Rome, where she remained until after Caesar's murder in 44. Some time later she replaced Ptolemy XIV as coruler with Ptolemy XV (Caesarion), allegedly her son by Caesar.

Within Egypt Cleopatra relied largely on the native element. She could speak Egyptian, an unusual accomplishment for a Ptolemy. It was also essential to have a strong friend at Rome. In 41 she was summoned to Tarsus by Mark Antony (Marcus Antonius), now in charge in the East; but the ensuing liaison was broken in 40 when Antony was reconciled to Octavian (later the emperor Augustus), whose sister Octavia he married.

Between 40 and 37 Roman armies in the East were fighting the Parthians, who had swept over Syria and Anatolia; in 38 the Romans drove them back to the Euphrates, and in 37 Herod, the son of the Idumaean king Antipater, expelled them from Judaea and established himself on the Judaean throne.

Antony resumed his liaison with Cleopatra, sending Octavia back to Italy; and in 37–36 at the so-called Donations of Alexandria, a last short-lived organization of the Hellenistic East took place. Asia and Bithynia remained provinces; but most of Asia Minor was assigned to the three augmented kingdoms of Pontus, Galatia, and Cappadocia (Cilicia Campestris was added to Syria, and the Lycian cities remained independent). Farther southward Nabataea and Judaea were client kingdoms, while Egypt was given southern Syria, Cyprus, and part of Cilicia Trachea. In 34 Cleopatra was designated queen of kings; Caesarion, king of kings; and the three children Cleopatra had borne to Antony were assigned satellite kingdoms—Armenia and lands beyond the Euphrates to Alexander Helios; the lands west of the Euphrates from the Hellespont to Phoenicia to Ptolemy Philadelphus; and Libya and Cyrenaica to Cleopatra Selene. These fanciful plans were all destroyed in 31, however, when Octavian defeated the Antonian and Egyptian fleets at Actium. By 30 Octavian was in Alexandria and the Ptolemaic monarchy dissolved. Egypt was thenceforth a province of Rome, enjoying very special relations with Octavian and his successors in the principate. The Hellenistic world had become part of the Roman Empire.

HELLENISTIC POLITICAL, SOCIAL, ECONOMIC, AND CULTURAL INSTITUTIONS

Hellenistic monarchy and royal administration. Hellenistic monarchy united features taken from Macedonian kingship, in which the god-descended king claimed the personal allegiance of his peers, and from the absolute monarchies of Persia and Egypt. It built up a set of fairly homogeneous institutions, which varied slightly from state to state. The outward marks of Hellenistic monarchy common to all the kingdoms included dynastic rule, often with co-regency to ensure the succession; a divine pedigree; the use of the royal diadem, purple robes, and seal; the king's portrait on the coinage; and reckoning by years of the reigns. Dynastic intermarriage was normal and polygamy not unknown, but only in Egypt was brother–sister marriage usual. The wives of the kings frequently exercised great and independent power.

In the new realms (excluding Macedonia) the king was the state and his powers were absolute. His claim to this position rested on conquest of the land and his capacity to rule. But his position was reinforced by the adoption of ruler cult, an institution with mixed origins. To the native population of Egypt the pharaoh was always a god; but the official cults were something different and more Greek. In 324 Alexander had demanded worship from the Greek cities. The main step, however, was taken by Ptolemy II, who in 283 consecrated the dead Ptolemy I as the "saviour god," adding his wife to the cult on her death. In 270 he consecrated his own late wife Arsinoe II as the "brother-loving goddess," and he later set up a joint cult of himself and Arsinoe as the *theoi adelphoi* ("brother and sister gods"). His successors added themselves to this growing cult embracing all the Ptolemies; Alexander was set at the head, and Ptolemy IV incorporated Ptolemy I and his wife (the saviour gods) in the sequence. Royal cult was introduced into the Seleucid kingdom, probably by Antiochus I; and all the other kingdoms quickly followed suit. The Greek cities also instituted cults, separate from those of the states in which they were situated.

Mixed origins of the ruler cult

The precise meaning of such cults is unclear. God-kings were the object of vows but not of prayer. Yet the view that deification was designed to give the king a footing in the Greek cities is not supported by evidence. There was

a strong political element, a means of uniting often very disparate areas and people in a common loyalty; the Greeks had always been familiar with gods who had mortal parents (like Apollo or Dionysus) or who had actually been mortal (like Asclepius or Heracles), and they were ready to equate achievement and service to mankind with godhead. The institution was later to be taken over by the Roman emperors.

The courts and administration of the various kingdoms shared certain features. The king had a council of friends (later elaborately subdivided into various grades) with advisory duties; these and the "kinsmen" were of Achaemenid origin. As the fount of law, the king possessed a secretariat to draft the decrees and letters in which law was formulated; and there was a secretary who wrote the official journal. From these functions a chancery developed with its own protocol and language sufficiently distinctive to influence the style of contemporary writers.

Other institutions common to most of the kingdoms were the guard (*agēma*), the corps of royal pages, and the bodyguards (*sōmatophylakes*), all inherited from Alexander. Normally there would be a finance minister (*dioekētēs*)—like Apollonius under Ptolemy II—and sometimes (perhaps only for an emergency) there might be a minister of affairs. The various districts within the boundaries of the state were usually governed by generals (*stratēgoi*) with military powers, responsible directly to the king (though the Antigonids used them rarely in Greece and not at all in Macedonia and Thessaly); in the Seleucid kingdom, down to the reign of Antiochus III, satraps were found, except in Asia Minor. There were admirals (navarchs) to control naval forces and, often, city governors (*epistatai*).

The army-assembly tradition

The traditional army assembly continued to be a factor in Macedonia, and some of its rights—the acclaiming of a new king, the appointing of a regent when the king was a minor, the approbation of the king's will and testament, and the judging of those charged with high treason—were sporadically exercised under the immediate successors of Alexander. Occasional political acts of the Seleucid and Ptolemaic armies in Antioch or Alexandria may be based on this Macedonian tradition, though often they seem to have acted more like a praetorian guard. Those armies ceased to contain any genuine Macedonian elements, except, of course, in Macedonia itself. In general, the administrative machinery was over centralized, and, in consequence, far too much had to be done (or was left undone) by the king in person. Especially in Ptolemaic Egypt, the king faced an overwhelming volume of paperwork as petitions and references travelled up and decisions travelled down the chain of control.

The old and new Greek cities. Though really independent *poleis* were rare, the Greek city played a fundamental part in all the Hellenistic monarchies except perhaps Egypt. It was pre-eminently through the cities that Hellenism was maintained. King and city were interdependent, but their relationship was ill defined and rested ultimately on a basis of power. Although the machinery of an independent state was usually maintained through its council, assembly, and magistrates, royal domination was frequently assured by a king's party, by uniting several towns or villages into a single community to create virtually new cities out of old, and by the presence of royal governors and garrisons. The king's attitude toward the cities was based mainly on expediency, and the slogan of "liberation" was cynically exploited.

Ultimately only a city such as Rhodes, capable of military resistance, could be truly independent; for the rest, freedom from taxation was the nearest approach to liberty within their grasp. Many of the cities of Greece proper and the Aegean maintained a theoretical independence, which in practice was diminished by the presence of Macedonian garrisons at key points and, in some periods, by local tyrants in Macedonian pay. The development of leagues and other federal bodies compensated for this weakness (see above *The Greek leagues and the Antigonids in the 3rd century BC*) unless, as frequently, such bodies were themselves the instrument of a king (*e.g.*, the Islanders' League or the Hellenic Alliance of Antigonus Doson). Because Macedonian domination was always resented, the "liberation" of Greece by Flamininus in 197 was greeted with joy.

Under such restrictions men turned from politics to social and private life; clubs became more important than the state, and the main magistrates came to be those concerned with the food supply and the gymnasiums. The reduction in the number of small wars was a positive gain, an advantage that was increased by the kings' role in assigning foreign judges or making grants of asylum to shrines or even cities.

New Asian cities

The new cities in the kingdoms of Asia differed from the old city-states. Though often Macedonian military settlements in origin, they became centres of Hellenism, even if the racial basis of citizenship was less rigorously enforced. They were usually built on the gridiron pattern, and their population was divided in the old Greek way into tribes. As already mentioned, these cities played an important role in Seleucid Asia; in Egypt, they were rare and scarcely essential to the Ptolemaic pattern of government (apart from Alexandria, which was hardly a normal Greek city). The Attalids emulated the Ptolemies in maintaining a strict control of the cities within their dominions. But nowhere did the Greek cities have any say in policy; their function was to radiate Hellenism and to provide manpower reservoirs from which the kings could draw their administrators.

Hellenizing of non-Greek lands. Alexander's campaigns had initiated a period of Greek colonization comparable to the early migrations and to the movement of the 8th to 6th centuries; during the late 4th century thousands found new homes in Asia or Egypt, the more easily because the new political order facilitated movement. With them they took their political institutions, their language, and their way of life. Throughout the new monarchies Greek was the language of the army and the bureaucracy. In its Hellenistic form (the Koine) Greek became a world language, spoken from Spain to Afghanistan. Other Hellenizing institutions were Attic law, which was widely adopted in forms modified to suit local traditions, and especially Greek education. To have passed through the gymnasium—usually a private institution, though occasionally aided by royal endowment—was to have entered the gateway to Hellenism and material success. Out of the old military training (the *ephēbeia*) had also grown a new physical and intellectual regime designed to inculcate Greek values and produce Greek gentlemen and bureaucrats capable of serving in the administration.

For a century the Greeks, constantly reinforced from the homeland, remained exclusive in their culture. Then gradually upper class Asians became accepted and, when properly Hellenized, would live like Greeks and take Greek names. The values of the gymnasium were reinforced by the performances in the Greek theatre, which was to be found as far away as Babylon by the mid-2nd century. The guilds of travelling actors, on the fringe of upper class life—the *technitai* of Dionysus—found a new field of activity in the monarchies and travelled widely throughout the expanded Greek world. Greek, too, were the many private social clubs on which so much of men's interests now centred and that served as an avenue into society for foreigners arriving in a new city.

Class structure in Greece and eastward

For a few the new kingdoms brought great wealth. The higher officials in the army, the navy, or the court were well-off, indeed. In the cities of mainland Greece and the Aegean islands, wealthy landowners, shipowners, and owners of workshops employing slave labour were still important; but farther eastward the upper layers of society were filled rather by the professionals employed by the kings to run the state. Professionalism was strengthened in many fields, such as warfare, by the availability of technical treatises. As a result, it took the place of the amateurish organization of the city-state, as befitted an age that fostered the systematic recording of knowledge and experience. Both the upper class and those immediately beneath were Greek or Hellenized natives who made a substantial livelihood as settlers, merchants, craftsmen, and soldiers, or practiced the profession of

architect, engineer, doctor, or teacher. These people formed the active middle class in the new Greek cities, and social and economic life in the new lands rested largely upon that class.

At the bottom were the labouring classes—usually made up of the native population, whose status varied from place to place; they could include freemen, serfs, and slaves. They were mainly non-Greek (though mixed populations were found, especially in Syria and Egypt), and their status depended on the tradition of the various lands. In general, however, they were poor and oppressed, though there was no racial or nationalistic policy directed against them. Their hostility and, in Egypt, their frequent resort to flight were reactions to the hardship inherent in their situation. Slavery was not especially significant, except in Greece proper; there were domestic and temple slaves. Serfdom was far more widespread in both Seleucid Asia and Ptolemaic Egypt. In a society where all were the king's subjects, this lack of personal freedom was not, however, greatly important.

Economic life. Until the campaigns of Sulla in 87–86, the upper classes in Greece enjoyed considerable economic prosperity; not all areas, however, shared equally in this. Athens fell behind its earlier standards of wealth until it recovered in the late 2nd century; and Sparta was rent with social struggles. Northern Greece prospered, especially Aetolia and Ambracia and the port of Pagasae under Macedonian control; and evidence points to expanded trade and a high standard of living in the cities of Asia Minor, where riches were far greater than anything recorded from Greece proper. This prosperity is reflected in the amount of tribute raised by Rhodes from its mainland possessions, in the size of gifts or dowries or subscriptions to clubs in a variety of cities, and not least in the amount of plunder eventually seized by the Romans. Evidence from Corinth and Messenia shows mainland Greece also sharing in this prosperity, though to a lesser degree. A progressive fall in interest rates in the early 3rd century may be a partial indication that money was plentiful. But while their upper classes were rich, the cities themselves were often poor and living from hand to mouth. Most needs were met by raising loans (which were often tardily repaid), and public works frequently languished. A property tax was rare; cities preferred to raise funds by indirect taxes on imports and exports, pasturing fees, harbour dues, and the like.

Conditions were less good for the little man and the artisan. For the free worker available figures suggest hard times in the 3rd century, as money depreciated in value, and wages failed to keep abreast of rising prices for food and housing. The gap that thus grew between rich and poor created unhealthy tensions; and many cities tried to ease conditions for the indigent by distributing free grain. Although strikes were virtually unknown, revolutionary risings occurred in several cities. The call for debt cancellation and land redistribution was sometimes reinforced with demands for the liberation of slaves (to serve as troops in the revolution). Such a pattern is most clearly evident at Sparta under Agis and Cleomenes and later Nabis, as well as elsewhere, including Achaea around 150–146.

Conditions within the Hellenistic kingdoms

Within the Hellenistic kingdoms conditions were different. Those regimes were organized as largely closed economies with a high degree of governmental control. Especially the Ptolemies, themselves active in commerce, operated a strictly centralized economy outside the operation of the market; thus, ultimately, against the heavy burden imposed by the bureaucracy, the only remedy open to the peasantry was flight. The other monarchies, if to a lesser extent, also employed a similar system. In Asia much of the wealth released by Alexander's campaigns together with other capital went to promote trade farther eastward until the initial impulse died down after about 250. In Egypt in the 2nd century, inflation was widespread and contributed to a growing inefficiency in the bureaucratic machinery. The centralized controls within the monarchies allowed agriculture to benefit from some improved techniques, more intensive exploitation of the land (which was sometimes added to by draining, as in the Fayyūm in Egypt), and the introduction of new plants and crops together with a few new domesticated animals. Large estates, devoted to the enrichment of a minority, also afforded opportunities for experiment and more scientific farming. Though a few large slave-manned establishments were found in such cities as Pergamum and Alexandria, industry remained (as in the Greek cities) on a small scale; as in agriculture, the main advantages went to the government through heavy taxation or through the licensing of monopolies for the production of such commodities as salt, oil, papyrus, and grain. Trade was burdened with heavy customs dues, and foreign trade tended to be largely in luxury goods and slaves. Production for the market was rare.

The declining economic life of the Hellenistic world suffered heavily from the ravaging that accompanied the Roman conquest. From 150 onward, the growth of piracy and slave dealing further disrupted economic life, and recovery had to await the Augustan peace and the incorporation of Egypt, Asia Minor, and Syria into the Roman Empire.

Hellenistic culture. The Hellenistic world was culturally one: extending from the western Mediterranean to Central Asia, it was linked by the use of the Koine form of Greek and by new facilities for travel. A lively atmosphere of curiosity prevailed, and a large public, even though not highly educated, turned eagerly to popularized knowledge and to a wide range of reading, often of indifferent merit. Ideas circulated, and among the cultural minority controversy stimulated research. There was cross-fertilization between the various philosophical schools; and in the arts masters created their own following, often associated with particular cities. The barriers of the *polis* had broken down and with them the old division between Greek and barbarian. In the ferment of such centres as Antioch, Pergamum, and Alexandria, the East made its contribution no less than did the old Greek cities such as Athens.

A great role was played by royal patronage. The library and museum at Alexandria fostered literature and scholarship: the systematization of all kinds of knowledge, already pursued enthusiastically within the school of Aristotle, was there developed into a science. Modern texts still owe much, both for good and bad, to the Alexandrine philologists, whose prototype is perhaps the famous Didymus (*c.* 83–10 BC), author of 3,500 books. A vast literature grew up in all the main areas of the Hellenistic world; almost all of it is lost, but some of the New Comedy plays of Menander (*c.* 342–*c.* 292 BC) are among the works that have been recovered on papyri from the sands of Egypt. Writers showed a growing interest in personality, reflected in such works as Theophrastus' *Characters* (*c.* 300 BC) and a host of biographies, which culminated in the *Lives* of Plutarch in the 1st century AD. The end of political independence led to a diminution of political oratory but not rhetorical studies. The sensational and dramatic aspects of the historiography of the 3rd century came under heavy criticism in the next century by Polybius, the historian of Rome's rise to world power. New genres appeared, such as literary and cultural history and art history; chronology was scientifically studied, and Polemon of Ilium (2nd century BC) made use of inscriptions. The widening of frontiers brought with it a growing interest in geography, both descriptive and scientific; here the works of Eratosthenes (*c.* 276–*c.* 194 BC), Pytheas (*c.* 300 BC), and Polybius (*c.* 200–118 BC) were to lead on through Poseidonius (*c.* 135–*c.* 51 BC) to the work in Augustan times of Strabo. Vast works of compilation and encyclopaedias were balanced at the other end of the scale by romantic novels and scurrilous works of scandal.

Literature and scholarship

In the visual arts there was a similar variety. A general restlessness encouraged change and the exploration of new concepts. Sculpture lost its formal composure and became tortured or realistically down-to-earth, portraying the individual rather than some idealized type. Romantic themes occur, and there is an interest in children. Famous schools grew up at Rhodes that were lively and

technically very competent; of Chares' "Colossus," unfortunately, no clear account survives. The school of Pergamum, deriving largely from Scopas (4th century BC), is famous for such violent and emotional pieces as "The Dying Gaul" and above all the great frieze on the altar of Zeus depicting the battles of the gods and titans, the symbol of a world in struggle. Of painting only the faintest idea can be gained from such sources as the gravestones of Pagasae, the mummy cases of the Fayyūm, and a remarkable mosaic representing Alexander and Darius, probably at Gaugamela. Hellenistic painting was clearly an important art form, which was to be influential (along with sculpture) in fertilizing the arts of the East during imperial times.

Oriental traits in religion

Reactions to the new world after Alexander are reflected especially in the growth of Oriental traits in religion. The Olympians, associated closely with the old city-states, had ceased to mean anything to ordinary people; and attempts to revive their cults, especially in the later 2nd century, proved ineffective. The Greeks, as they travelled and settled, sometimes brought their own gods along but often adopted those of the native people. Worship became more personal and linked with club life rather than the city. The combining of beliefs was widespread: there was felt to be one god, no matter how many names he bore.

The multitude of gods in Syria had a wide following, especially Atargatis (who was associated with Artemis); many were personalized cult objects, stones, or meteorites. Above all, men worshipped the great Anatolian nature goddess in her many guises. Of the older gods Dionysus was the most popular; and he was sometimes identified with Sarapis, the great creation of Alexandria, intended to link Greeks and Egyptians, who, however, rejected it. Isis, Sarapis, and Anubis formed a triad, worshipped outside and inside Egypt; Isis especially had perhaps the most widespread and far-reaching cult in the Hellenistic world and was later to influence Christianity through the figure of the Virgin Mary. But mystery religions that offered personal salvation flourished everywhere, as did magic, the shortcut to the achievement of one's immediate and often trivial desires.

Philosophy appealed to a different level of the population. Athens remained a great centre and gave hospitality not only to its own Academy (the school of Plato) but also to the Peripatetics (Aristotle's school) and later to Epicurus (341–270 BC) of Lampsacus and Zeno (*c.* 335–263 BC) of Citium, founders of the two most important Hellenistic schools—the Epicureans and the Stoics. After the work of Theophrastus and Strato, the Peripatos had lost its momentum and counted for little by 250; and the Academy under Arcesilaus took over much of the teaching of the Skeptics, who advocated the suspension of judgment, so that it became largely identified with the criticism of Stoicism.

Both Epicurus and Zeno were concerned with personal conduct rather than with abstract truth—with the individual rather than the nature of the outside world. They showed small interest in politics or indeed in the learning and science of their age. Both set out to equip men to deal with the blows dealt by fortune in an age of violent change—Epicurus by a doctrine of quietism and happiness, resting on virtue in a universe where the random movement of atoms was the ultimate causative force and the gods were remote and aloof; and Zeno by preaching the overriding force of universal law, the power of destiny, that man's greatest goal was to become the Stoic "wise man," and to live "according to nature." Of these, Stoicism was to live on, modified by Panaetius (*c.* 180–109 BC), into the Roman Empire.

Scientific development

At the other end of the spectrum, the age was one of scientific development and of progress in the arts that impinge on the sciences. Systematic work in botany, zoology, geography, ethnography, and hydrography was pursued in the Peripatos. The description of animals was furthered by the availability of zoos in such cities as Alexandria; and Theophrastus wrote a *History of Plants;* both zoology and botany were handicapped, however, by the lack of any scientific classification. Long controversy raged between supporters of the heliocentric and geocentric theories of the universe: the incorrect but more convincingly argued Earth-centred theory of Hipparchus (fl. between 146 and 127 BC) of Nicaea carried the day against the Sun-centred theory of Aristarchus (*c.* 270 BC) of Samos, and it was to persist until the Renaissance. Much headway was made toward creating a world map, despite the inability to solve the problem of detecting longitude. Eratosthenes, however, succeeded brilliantly in estimating the size of the Earth's circumference. Mathematics was developed by Eratosthenes and by Archimedes (*c.* 287–212 BC), who worked out the value of π and who also made a reputation (which he despised) as a practical mechanic, with such inventions as the screw for raising water. On the whole, although progress was made in siege engines, water clocks, eventually a water mill, a water organ, and the diopter (an improved form of water level), the Hellenistic world tended to agree with Archimedes in rating these activities below pure speculation. Some fields were thereby seriously handicapped—*e.g.*, physics and chemistry, which languished for want of optical glass. But progress was made in medicine, with increased knowledge of the nerves and arteries (though no one managed to detect and describe the circulation of the blood).

In one field, however, progress was linked up closely with the practical application of knowledge. The new cities gave wide scope for architects and engineers. The gridiron pattern originated by Hippodamus of Miletus (5th century BC) was everywhere reproduced (with appropriate concessions to the contours of the natural site); and the temples, marketplaces, and wide streets of Priene, Pergamum, or Miletus can still be reconstructed. Fortifications were devised to counter the increasingly effective artillery of the age; and many cities installed impressive water and sewerage systems, thus providing agreeable living conditions for large populations.

Commerce, travel, and exploration. Alexander's campaigns had widened the horizon of the Greek world, and his successors carried exploration further. Ptolemy II, in search of ivory and elephants, opened up the Red Sea along the Trogodyte coast and explored the Arabian coast (about 280); and the later Ptolemies pushed the boundaries of Egypt up the Nile to beyond Wadi Halfa and explored the African coast down to Somaliland. The Seleucids were active in the region of the Caspian, and farther eastward the Bactrian kings colonized Fergana and made contact with Chinese Turkistan. In the West, Pytheas of Massilia, probably a contemporary of Alexander, circumnavigated Britain and reached Norway or Jutland, but he met with widespread disbelief, among others from Polybius, who himself made a voyage in the Atlantic (146).

Trade by land and sea with India

Such journeys were undertaken partly in the interest of trade, and trade fostered knowledge of the outer regions. The Seleucids were interested in the route to India—by sea up the Persian Gulf to Seleucia, where it was joined by caravans coming overland from the Punjab through Taxila, Bactra, Hecatompylos, and Ecbatana; thence several ways led down to the coast, that by Antioch going forward along the old Royal Road through Tarsus to Ephesus. The Ptolemies encouraged the development of the route around southern Arabia, especially after Ptolemy V had lost Coele Syria; and toward the end of the 2nd century the collapse of the power of the Sabaeans in the Yemen opened up the way beyond. Eudoxus of Cyzicus was probably the first European to sail direct to India and back along the coast (*c.* 130 BC). Farther northward the Nabataeans at Petra built up a strong power by controlling the desert routes and sea trade in the Gulf of Aqaba.

Within the Hellenistic world itself, trading was active. Grain from Egypt, Cyrenaica, and the Crimea (and later Numidia) came to the larger cities of Greece and Ionia, largely through depots at Rhodes and Delos. North Syrian and Ionian (and island) wine went everywhere. Textiles were produced in most areas that reared sheep, and silk from wild silkworms was a valued luxury; by the first century Chinese silk may have been reaching

Egypt. There was an adequate supply of metals—iron in many parts, gold from new Nubian mines, silver from Asia Minor and Spain, quicksilver from Cappadocia. Many places had their specialties—Attic oil and honey, Bithynian cheese, salt fish from Byzantium, fruits and nuts from Pontus, and dried fruit from the Levant. Papyrus from Egypt vied with parchment from Pergamum. Glass, pitch, marble, bitumen, timber, granite, purple dye, and ivory were all well-attested articles of trade. But above all, it was the slave trade, centring in Delos from about 160 onward, that brought riches and misery as it grew along with Roman imperial expansion and the prevalence of piracy. (Delos, however, never recovered from the massacre carried out by Mithradates VI in 87.) Especially prized were the gems, spices, and frankincense from the East—from India and Arabia as well as from Palestine. For these Rhodes and Alexandria were important distribution centres.

Commerce contributed to the ferment of ideas as well as to the exchange of commodities and played a large part in the knitting of the Hellenistic world into a cultural whole. In general, this was a time when travel was not unusual. Among the many who travelled about the world were soldiers, especially mercenaries, professional men (teachers, philosophers, doctors, architects, and engineers) in search of patronage, actors, official envoys from kings and cities, some members of the royal courts and bureaucracies, and, not least, uprooted men (exiles, brigands, and pirates). But the majority, as always, stayed at home.

MACEDONIA AND GREECE UNDER ROMAN RULE (TO C. AD 395)

Social distress and slave uprisings

After 146 BC the decline of Greece continued. Social distress was accentuated by economic exploitation at the hands of the rich, both native and Italian. Athens was an exception, so long as it held the thriving port of Delos, but there were slave uprisings in Attica in 134–133 and 104–103, and in 115 Dyme was rent with class conflict. Sulla's campaign against Athens during the Mithradatic War brought widespread ruin, and during the 1st century BC the western coast of Asia Minor became the real centre of a Greek culture that continued to flourish despite political disasters. The civil wars between Caesar and Pompey and later those against Caesar's murderers and between Antony and Octavian bore hard on Greece: the battles of Pharsalus (48), Philippi (42), and Actium (31) were all fought on or around Greek soil. But under the early empire, Nicopolis, Patras, and Corinth (refounded as a veteran colony) became flourishing centres of new life. Athens was in decline; but Sparta, under its tyrant Eurycles during the Roman civil wars, revived some semblance of its old Lycurgan regime, if mainly for the tourist trade. Delphi was impoverished, and the amphictyonic council counted for little until its revival under the emperor Hadrian (ruled AD 117–138). From 27 BC onward, southern Epirus, the Ionian islands, and the Cyclades were joined with Greece proper to form the senatorial province of Achaea under a proconsul of praetorian rank resident at Corinth and without a garrison.

On November 28, AD 67, Nero aroused wild enthusiasm by liberating Greece and freeing it from all tribute. The Peloponnese was renamed the Neronese, and Nero himself was worshipped as Zeus Eleutherios (Liberator). Nero's successors, the Flavian emperors, however, quickly reversed this gesture and restored taxation to the province. Prosperity revived to some extent under the emperors Trajan (ruled AD 98–117) and Hadrian, for although the upper class in Macedonia was somewhat mulcted to produce supplies for Trajan's wars in Dacia and Parthia, Greece generally benefitted from these conflicts and also from Hadrian's favour and enthusiasm for building. Athens received a whole new suburb, new buildings, and Panhellenic games were instituted; a new Athenian era was dated from the first of Hadrian's three visits in AD 124. The high-water mark of the revival of Athenian prosperity came under the Antonine emperors.

In the second half of the 2nd century AD the shadows lengthened. A plague brought back from the East by Lucius Verus' armies decimated Greece, and the barbarian Costoboci burnt Eleusis. Under Commodus and the Severi in the late 2nd and early 3rd centuries, tension between rich and poor grew greater and the institution of the *dekaprōtoi*, responsible for extracting taxes from an increasingly impoverished upper class, brought a final end to all pretense of free city institutions. In 251 the death of the emperor Decius opened the gates to barbarian invasion; in 267 the Heruli penetrated the Peloponnese and ravaged Corinth, Sparta, and Argos, until, after taking Athens, they were defeated by the historian Dexippus and a force of 2,000 men. Despite these disasters Athens saw some sort of revival coinciding with the teaching of Plotinus (*c.* 205–269/270) and the reign of Gallienus (ruled 260–268). Diocletian's reign (284–305) strengthened the central bureaucracy. Achaea and Macedonia, which now formed part of the diocese of the Moesiae, both decayed further. When the last Olympic festival was held in 393, few Greeks any longer took part. The whole peninsula had long been a backwater of small villages and large estates. (F.W.W.)

BIBLIOGRAPHY

Greek civilization from c. 1200 to 500 BC: ANTONY ANDREWES, *The Greeks* (1967), presents the historian's view; C.M. BOWRA, *The Greek Experience* (1957), approaches the Greeks through their literature. N.G.L. HAMMOND, *A History of Greece to 322 B.C.*, 2nd ed. (1967), gives a full account and is mainly devoted to political history. See also *The Cambridge Ancient History*, vol. 3–7 (1925–28), and some relevant chapters in the new edition of vol. 2, which is appearing in fascicles (1970–). MICHAEL GRANT (ed.), *The Birth of Western Civilisation: Greece and Rome* (1964), contains essays by different scholars on Greek history, art, and thought (fully illustrated). (*Geography of the Greek homeland*): MAX CARY, *Geographic Background of Greek and Roman History* (1949), is a useful treatment; and J.L. MYRES, *Geographical History in Greek Lands* (1953), contains essays including studies of geography and colonization. (*Sources for the study of Greek civilization*): MICHAEL GRANT, *The Ancient Historians* (1969), is of much value. P.L. MACKENDRICK, *The Greek Stones Speak* (1962), contains a survey by periods and sites of archaeological discoveries in Greece. The principal ancient historical sources, all available in translation, are HERODOTUS (Archaic period), THUCYDIDES (Peloponnesian War), DIODORUS (5th–4th century), ARRIAN (Alexander), XENOPHON (4th century and the East), PAUSANIAS (description of Greece), and PLUTARCH (biographies). (*The Early Archaic period*): V.R.d'A. DESBOROUGH, *The Last Mycenaeans and Their Successors* (1964) and *Protogeometric Pottery* (1952), contain detailed studies of the full range of archaeological evidence. J.N. COLDSTREAM, *Greek Geometric Pottery* (1968), includes a survey of the archaeological evidence for the history of the period. A.M. SNODGRASS, *The Dark Age of Greece* (1971), provides a full archaeological synthesis of the evidence. C.G. STARR, *The Origins of Greek Civilization, 1100–650 BC* (1961), gives a historian's view of the period. M.I. FINLEY, *The World of Odysseus*, rev. ed. (1962), projects Homer's world into the Dark Ages. (*The Greek colonizing movement*): JOHN BOARDMAN, *The Greeks Overseas* (1974), provides a survey of the archaeology of colonizing and trading posts. T.J. DUNBABIN, *The Greeks and Their Eastern Neighbours* (1957), contains lectures on relations and trade with the East. M.M. AUSTIN, *Greece and Egypt in the Archaic Age* (1970), is a short but detailed study of mercenaries and trade in Egypt. T.J. DUNBABIN, *The Western Greeks* (1948), gives a full survey of all sources for south Italy and Sicily. (*Archaic Greek culture*): G.M.A. RICHTER, *Archaic Greek Art Against Its Historical Background* (1949), is a survey by regions and periods; and JOHN BOARDMAN, *Pre-Classical: Style and Civilization* (1967), a brief study of art and society in the Archaic period. A.R. BURN, *The Lyric Age of Greece* (1960), is a good study of all sources with accounts of literature, art, and thought, as well as political and social history. See also W.K.C. GUTHRIE, *A History of Greek Philosophy*, vol. 1, *The Earlier Presocratics and the Pythagoreans* (1962); G.E.R. LLOYD, *Early Greek Science: Thales to Aristotle* (1970); and KARL SCHEFOLD, *Frühgriechische Sagenbilder* (1964; Eng. trans., *Myth and Legend in Early Greek Art*, 1966).

Development of the Greek poleis from c. 750 to c. 500 BC: HERODOTUS' *History* is the most valuable source for the period. ARISTOTLE's *Constitution of the Athenians* is useful, especially on Solon and Cleisthenes. PLUTARCH's lives of Lycurgus, Theseus, and Solon preserve some valuable facts. Fragments of the poems of Tyrtaeus and Solon with translations

are accessible in J.M. EDMONDS, *Elegy and Iambus*, vol. 1 ("Loeb Classical Library," 1931). Twentieth-century scholarship begins with K.J. BELOCH, *Griechische Geschichte*, 2nd ed., 4 vol. in 8 (1912–27); the first volume deals with the period before the Persian Wars. J.B. BURY, *A History of Greece*, 3rd ed. rev. by RUSSELL MEIGGS (1951), has still to be surpassed as a one-volume survey. ANTONY ANDREWES, *The Greeks* (1967), offers a good account, more descriptive than narrative. W.G. FORREST, *The Emergence of Greek Democracy, 800–400 B.C.* (1966), provides a brief and thoughtful introduction. *The Cambridge Ancient History*, vol. 3–4 (1929–39), though old, is still useful, especially on interstate relations. (*Tyranny and hoplite tactics*): ANTONY ANDREWES, *The Greek Tyrants* (1956), collects together plentiful information. H.T. WADE-GERY, "The Growth of the Dorian States," ch. 22 of *The Cambridge Ancient History*, vol. 3, presents admirably one theory of tyranny. A.M. SNODGRASS, "The Hoplite Reform and History," *Journal of Hellenic Studies*, 85:110–122 (1965), supersedes previous work on the beginnings of hoplite tactics. The same author's *Arms and Armour of the Greeks* (1967) gives a more discursive account of hoplite tactics. (*Sparta*): H.T. WADE-GERY, "The Spartan Rhetra in Plutarch, *Lycurgus VI*," *Classical Quarterly*, 37:62–72 (1943) and 38:1–9, 115–126 (1944), reprinted in *Essays in Greek History* (1958), is a good starting point for the study of early Sparta. ANTONY ANDREWES, *Probouleusis, Sparta's Contribution to the Technique of Government* (1954), offers a theory of conflict between an aristocracy and a wider class. G.L. HUXLEY, *Early Sparta* (1962), is provocative. W.G. FORREST, *A History of Sparta 950–192 BC* (1968), is a brief survey. ANTONY ANDREWES, "The Government of Classical Sparta," in E. BADIAN (ed.), *Ancient Society and Institutions: Studies Presented to Victor Ehrenberg on His 75th Birthday* (1966), makes important points. (*Athens*): CHARLES HIGNETT, *A History of the Athenian Constitution to the End of the Fifth Century B.C.* (1952), offers a view on most issues. H.T. WADE-GERY, *Essays in Greek History* (1958), includes several articles on constitutional questions. B.L. BAILEY, "The Export of Attic Black-Figure Ware," *Journal of Hellenic Studies*, 60:60–70 (1940), has fundamental importance for the economic history of Archaic Athens. ALFRED FRENCH, *The Growth of the Athenian Economy* (1964), is a good treatment of a difficult subject. Two valuable studies of the sources for Athenian constitutional history are: FELIX JACOBY, *Atthis, the Local Chronicles of Ancient Athens* (1949); and JAMES H. DAY and MORTIMER CHAMBERS, *Aristotle's History of Athenian Democracy* (1962). Two specialized studies have brought considerable advance in understanding Dracon and Solon: EBERHARD RUSCHENBUSCH, "Φονος zum Recht Drakons und seiner Bedeutung für das Werden des athenischen Staates," *Historia*, 9: 129–154 (1960) and *Σολωνος νομοι: Die Fragmente des solonischen Gesetzeswerkes* (1966).

The Greek world from c. 500 to 404 BC: The main sources for the period are HERODOTUS (trans. by G. RAWLINSON, introd. by F.R.B. GODOLPHIN, 1942) and THUCYDIDES (trans. by R. CRAWLEY, introd. by J.H. FINLEY, 1950). Herodotus records in detail the Persian invasion of Greece (Books VII–IX), with a long introduction on the events that led up to it. He had to rely largely on oral tradition and prejudiced sources and is unreliable on strategy and tactics, but his narrative is vivid, colourful, and very readable. Thucydides, writing a generation later, is a more mature and critical historian. His theme is the Peloponnesian War, with a brief introduction on the period between the two great wars. His main interests are the nature of power and the reaction of people to events, but he attached great importance to accuracy in the detail of his narrative. These literary sources are supplemented by inscriptions recording decrees, dedications, financial accounts, and casualty lists, especially from Athens; texts of the most significant, with commentary by RUSSELL MEIGGS and DAVID LEWIS, are in *A Selection of Greek Historical Inscriptions to the End of the Fifth Century B.C.* (1969). PLUTARCH, writing some 500 years later, in his lives of Aristides, Themistocles, Cimon, Pericles, Nicias, Alcibiades, and Lysander (*Plutarch's Lives*, trans. by JOHN DRYDEN), preserves important material from earlier sources that are lost, though he is primarily interested in the characters of his heroes and is reluctant to dismiss a good story. There are two good recent studies of Xerxes' invasion: A.R. BURN, *Persia and the Greeks* (1962); and CHARLES HIGNETT, *Xerxes' Invasion of Greece* (1963). DONALD KAGAN, *The Outbreak of the Peloponnesian War* (1969), discusses the policies that led up to the war, which are differently interpreted by G.E.M. DE ST. CROIX, *The Origins of the Peloponnesian War* (1972). RUSSELL MEIGGS, *The Athenian Empire* (1972), analyzes the evidence for the growth and character of the Athenian Empire. Of general histories of Greece, GEORGE GROTE, *History of Greece*, new ed. (1888, reprinted 1971), is still worth reading; N.G.L. HAMMOND, *A History of Greece to 322 B.C.*, 2nd ed. (1967), is detailed and up-to-date, but compressed; VICTOR EHRENBERG, *From Solon to Socrates* (1967), is also comprehensive and easier to read. At shorter length, J.B. BURY, *A History of Greece to the Death of Alexander the Great*, 4th ed. (1974), has been revised to take account of new evidence. For the economic background HUMFREY MICHELL, *The Economics of Ancient Greece*, 2nd ed. (1957), is a useful review of the evidence; ALFRED FRENCH, *The Growth of the Athenian Economy* (1964), studies Athens in more detail. CHARLES HIGNETT, *A History of the Athenian Constitution* (1952), is a good critical review, often unorthodox, of Athenian politics and constitutional changes. W.G. FORREST, *The Emergence of Greek Democracy, 800–400 B.C.* (1966), is the most refreshing new look at Athenian politics and can be usefully supplemented by the examination of the mechanics of Athenian democracy in A.H.M. JONES, *Athenian Democracy* (1966), and the detailed analysis in P.J. RHODES, *The Athenian Boule* (1972). The best general account of Greek living conditions is ROBERT FLACELIERE, *La Vie quotidienne en Grèce au siècle de Périclès* (1959; Eng. trans., *Daily Life in Greece at the Time of Pericles*, 1960); the Athenian scene is gaily illustrated in the comedies of Aristophanes, which combine an excellent sense of theatre with strong political satire, fantasy, and boisterous bawdry.

The Greek world from c. 404 to 323 BC: A select bibliography of the history of the Greek world in the 4th century must necessarily concentrate on general works, which themselves contain more detailed bibliographies. The basic works for the history of 4th-century Greece in French, English, and German, are: GUSTAVE GLOTZ and ROBERT COHEN, *Histoire grecque*, vol. 3, *La Grèce au IVe siècle* (1936) and vol. 4 (pt. 1), with the assistance of PIERRE ROUSSEL, *Alexandre et le démembrement de son empire*, 2nd ed. (1945); *The Cambridge Ancient History*, vol. 4, 3rd ed., *Macedon (401–301 BC)* (1953), better informed than the preceding on the most recent works; HERMANN BENGTSON, *Griechische Geschichte*, 3rd ed. (1965). ROBERT COHEN, *La Grèce et l'héllenisation du monde antique*, new ed. (1948), is rich in ideas and precious for its indication of the sources and the present state of the questions. For the literary history of the period, still of use is the old and monumental *Histoire de la littérature grecque*, by the brothers ALFRED and MAURICE CROISET, 4th ed., vol. 4 (1928), but ALBIN LESKY carries forward the essential material in his volume of the *Geschichte der griechischen Literatur*, 2nd ed. (1963; Eng. trans., *A History of Greek Literature*, 1966), and his bibliography is more important, more precise, and more up-to-date than that of the Croisets. The diplomatic history is studied more particularly, and with much justice, by the Swiss Hellenist VICTOR MARTIN, *La Vie internationale dans la Grèce des cités, VIe–IVe s. av. J.–C.* (1940). On the civilization in the largest sense of the term, see the two volumes, whose text and illustrations are equally remarkable, of PIERRE LEVEQUE, *L'Aventure grecque* (1964); and of FRANCOIS CHAMOUX, *La Civilisation grecque* (1963); although dealing with nearly identical subjects, the two books, without supplanting one another, complete one another very well. The following are more detailed or more specialized works: The history of Demosthenes and of Philip is explained with finesse and clarity, despite the complications, by PAUL CLOCHE in *Un Fondateur d'Empire: Philippe II, roi de Macedoine* (1955); and WERNER JAEGER, *Demosthenes: The Origin and Growth of His Policy* (1938). The problems of the education of the era are studied thoroughly by H.I. MARROU, *Histoire de l'éducation dans l'antiquité*, 5th ed. (1960; Eng. trans., *A History of Education in Antiquity*, 1956, reprinted 1964), but not without some prejudices. In the matter of art, out of a superabundant bibliography three works stand out: one old but not out-of-date book that explains the evolution of the art of Greece in all of its forms in relation to its history, ANDRE DE RIDDER and W. DEONNA, *L'Art en Grèce* (1924; Eng. trans., *Art in Greece*, 1927, reprinted 1968); the great work of CHARLES PICARD, *Manuel d'archéologie grecque, la sculpture*, vol. 3–4 (1948–63); and a survey work—intelligent, well informed, and well illustrated—by FRANCOIS CHAMOUX, *Art grec* (1966). On Plato the book that can give the best general idea about his policy and philosophy is that of the great specialist AUGUSTE DIES, *Autour de Platon*, 2 vol. (1927), to which it is necessary to add the more summary *Platon*, by the same author (1930). EDOUARD DELEBECQUE, *Essai sur la vie de Xénophon* (1957), studies the relationships between this disciple of Socrates and his times, and attempts to fix the chronology. The best general work on Alexander is GEORGES RADET, *Alexandre le Grand* (1931, reprinted 1950); the bibliography has been brought up-to-date by W. SESTON.

The Hellenistic Age—323–30 BC: The Cambridge Ancient History, vol. 7–9 (1928–32), by various authors, with good bibliography; MAX CARY, *A History of the Greek World*

from 323 to 146 B.C., 2nd ed. rev. (1951); W.W. TARN and G.T. GRIFFITH, *Hellenistic Civilisation*, 3rd rev. ed. (1952, reprinted 1966), a good conspectus; PIERRE JOUGUET, *L'Impérialisme macédonien et l'hellénisation de l'Orient* (1926; Eng. trans., *Macedonian Imperialism and the Hellenization of the East*, 1928). The best recent account of the political history, with full references, is EDOUARD WILL, *Histoire politique du monde hellénistique (323–30 av. J.-C.)* (1966–67). (*Hellenistic kingdoms—Egypt*): E.R. BEVAN, *A History of Egypt Under the Ptolemaic Dynasty* (1927); P.M. FRASER, *Ptolemaic Alexandria* (1972). (*Seleucid Asia*): E.R. BEVAN, *The House of Seleucus*, 2 vol. (1902, reprinted 1966); ELIAS BICKERMAN, *Institutions des Séleucides* (1938). (*Bactria and India*): W.W. TARN, *The Greeks in Bactria and India*, 2nd ed. (1951), fundamental if rather imaginative; A.K. NARAIN, *The Indo-Greeks* (1957). (*Pergamum*): E.V. HANSEN, *The Attalids of Pergamon*, rev. 2nd ed. (1971); R.B. MCSHANE, *The Foreign Policy of the Attalids of Pergamon* (1964). (*Macedonia*): W.W. TARN, *Antigonos Gonatas* (1913); F.W. WALBANK, *Philip V of Macedon* (1940, reprinted 1967). (*Greek states*): J.A.O. LARSEN, *Greek Federal States* (1968); W.S. FERGUSON, *Hellenistic Athens* (1911); HUMFREY MICHELL, *Sparta* (1952). (*Hellenistic politics, institutions, and society—monarchy*): LUCIEN CERFAUX and J. TONDRIAU, *Le Culte des souverains dans la civilisation gréco-romaine* (1957). (*Organization*): HERMANN BENGTSON, *Die Strategie in der hellenistischen Zeit*, 3 vol. (1937–52). (*Hellenization*): MOSES HADAS, *Hellenistic Culture: Fusion and Diffusion* (1959). (*Economy*): M.I. ROSTOVTZEFF, *The Social and Economic History of the Hellenistic World*, 2nd rev. ed. (1957). (*Cultural activities*): MARGARETE BIEBER, *The Sculpture of the Hellenistic Age*, rev. ed. (1961); T.B.L. WEBSTER, *Hellenistic Poetry and Art* (1964), and *Hellenistic Art* (1967); R.E. WYCHERLEY, *How the Greeks Built Cities*, 2nd ed. (1962). (*Warfare*): W.W. TARN, *Hellenistic Military and Naval Developments* (1930); E.W. MARSDEN, *Greek and Roman Artillery: Historical Development* (1969), and *Greek and Roman Artillery: Technical Treatises* (1971). (*View of world*): J.O. THOMSON, *History of Ancient Geography* (1948). (*Roman period*): HERMANN BENGTSON, *Griechische Geschichte von den Anfängen bis in die römische Kaiserzeit* (1969).

(Jo.Bo./B.R.S./Ru.M./E.D./F.W.W.)

Greek Language

Greek is an Indo-European language whose history can be followed from the 14th century BC to the present day. Its documents cover a longer period of time (34 centuries) than those of any other Indo-European language. There is an Ancient phase, subdivided into a Mycenaean period (texts in syllabic script from the 14th to the 12th centuries BC) and Archaic and Classical periods (beginning with the adoption of the alphabet, from the 8th to the 4th centuries BC); a Hellenistic and Roman phase (4th century BC to 4th century AD); a Byzantine phase (5th–15th centuries AD); and a Modern phase.

Separate transliteration tables for Classical and Modern Greek accompany this article. Some differences in transliteration result from changes in pronunciation of the Greek language; others reflect convention, as for example the χ (*chi* or *khi*), which was transliterated by the Romans as *ch* (because they lacked the letter *k* in their usual alphabet). In Modern Greek, however, the standard transliteration for χ is *kh*. Another difference is the representation of β (*bēta* or *víta*); in Classical Greek it is transliterated as *b* in every instance, and in Modern Greek as *v*. The pronunciation of Ancient Greek vowels is indicated by the transliteration used by the Romans. Υ (*upsilon*) was written as *y* by the Romans, indicating that the sound was not identical to the sound of their letter *i*. Modern Greek υ (*ípsilon*) is transliterated as *i*, indicating that the sound used today differs from the ancient υ. (See the transliteration tables for values of all the Greek letters.)

GENERAL CONSIDERATIONS

In the course of the 2nd millennium BC, groups of Greek-speaking Indo-Europeans established themselves by stages on the Greek peninsula, on most of the islands of the Aegean, and on the west coast of Anatolia; with few exceptions that is still the area occupied by the Greek language today. In the second quarter of the 1st millennium BC a vast "colonial" movement took place, resulting in establishments founded by various Greek cities all around the Mediterranean and the Black Sea, especially in southern Italy and Sicily. This extension of the linguistic area of Greek lasted only a few centuries; in the Roman period, Latin, more or less rapidly, took the place of Greek in most of these ancient colonies. "Colonial" Greek survived longest at Byzantium, as the official language of the Eastern empire.

Relationship of Greek to Indo-European. Ancient Greek is, next to Hittite, the Indo-European language with documents going furthest back into the past. At the time when it comes within view in the 2nd millennium BC, it has already acquired a completely distinct character from the parent Indo-European language. Its linguistic features place it in a central region on the dialect map that can be reconstructed for Common Indo-European; the ancient languages with which it has the most features in common are little known ones such as Phrygian or Macedonian. In the study of Indo-European dialectology, phonetic data are the most readily available and provide the most information. In this respect the position of Ancient Greek is as follows. The original Indo-European vowels of *a* and *o* quality, both short and long, remain distinct, whereas they are completely or partially confused in Hittite, Indo-Iranian, Baltic, Slavic, and Germanic. Greek is the only language that distinguishes by three different qualities (*ĕ, ă, ŏ*) the secondary short vowels resulting in certain positions from the three laryngeal sounds, $*H_1$, $*H_2$, $*H_3$, of Indo-European. (An asterisk preceding a sound or word indicates that it is not attested, but is a reconstructed, hypothetical form. For a discussion of these laryngeal sounds, see INDO-EUROPEAN LANGUAGES.) Greek keeps the distinction between the original voiced stops and voiced aspirated stops of Indo-European (*e.g.*, Indo-European **d* becomes Greek *d*, and Indo-European **dh* becomes Greek *th*), whereas Iranian, Slavic, Baltic, and Celtic confuse them. Greek avoids the general shifts of stop consonants that are displayed, independently, by Armenian and Germanic, as well as the palatalization that affects guttural stops in Indo-Iranian, Armenian, Baltic, and Slavic. In these respects, Ancient Greek is conservative, as are, generally speaking, the western Indo-European languages (Italic and Celtic). On the other hand, it does show innovations. One of these, the devoicing of the original voiced stops, is shared with Italic, although it is realized in different ways (**dh*- yields Greek *th*-, Latin *f*-, Osco-Umbrian *f*-); but others are foreign to Italic: for example, the weakening of spirants and semivowels at the beginning of words before a vowel, the evolution of **s*- to *h*- (pre-Mycenaean), and **y*- to *h*- (contemporary with Mycenaean).

Comparison of the sounds of Ancient Greek with other Indo-European tongues

Morphological criteria must, of course, be taken into account in defining the position of a language. It should be noted that there are few grammatical innovations shared by Greek and Italic, apart from the extension to nouns of the pronominal ending of the genitive feminine plural **-āsŏm* (Greek *-āōn;* Latin *-ārum*, Umbrian *-aru*, Oscan *-azum*) and of the pronominal ending of the nominative masculine plural **-oi* (Greek *-oi;* Latin *-ī*). The last innovation, however, is not shared with Osco-Umbrian, but is found instead in Germanic (in the strong declension of adjectives) and partly in Celtic. The dialectal individuality of Greek is very clearly marked in the organization of the verb (see below), which is without parallel except for an approximation in Indo-Iranian.

Greek syllabaries. Starting from a foreign script known as Linear A (used in Crete to record a native language known as Minoan), the Greeks devised, toward 1400 BC at the latest, a syllabic script to record their own language. Known as Linear B, this script was deciphered in 1952 by the British architect Michael Ventris and the British classicist John Chadwick. At present just over 100 very short Linear B inscriptions painted on vases have been found at Khaniá, Knossos, Tiryns, Mycenae, Eleusis, Orchomenus, and especially at Thebes. The major source of Linear B inscriptions are the 3,000 to 4,000 unbaked clay tablets found at Knossos (1400–1350 BC—this date has been questioned), at Thebes (probably 1350–1300 BC), and at Mycenae and Pylos (1250–1150 BC). There are no literary texts, and hardly any continu-

ous texts (there is only a small number of real sentences); all that is currently known, and that only in part, is the accounts of the great Mycenaean palaces and their dependencies, compiled in the Greek language.

Signs of Linear B

The Linear B syllabary consists of about 90 signs. There are signs for the vowels *a, e, i, o, u,* but these are hardly used except for initial vowels of words. There are no signs noting consonants in isolation, only signs noting consonant + vowel combinations; thus there is no sign for *t-*, but five different signs for *ta, te, ti, to, tu*. The script does not distinguish *r-* and *l-*; with the exception of *d-*, it does not distinguish between unvoiced, aspirated, and voiced stops (so the sign *ka* can be read in Greek as *ka, cha,* or *ga*). In addition, the scribes used a shorthand spelling and saved time by omissions, mainly of certain consonants (in particular, those that end syllables or words). Consequently, the spellings are often clumsy and ambiguous, such as *ka-po* for *karpos, a-re-ku-tu-ru-wo* for *alektryōn, ka-sa-to* for *xanthōi*. This inconvenient script and the nature of the documents make Mycenaean inscriptions harder to use and less rich in data than the later alphabetic inscriptions; but the information that can be gathered on the state of Greek five centuries before Homer, incomplete as it may be, is of capital importance.

Another syllabary, distantly related to Linear B, was in use in Cyprus much later (7th–3rd centuries BC) to record a native language of the island (Eteocypriot) as well as Greek.

The Greek alphabet. The Mycenaean script dropped out of use in the 12th century when the Mycenaean civilization was destroyed by the Dorian invasions. For nearly four centuries the Greeks seem to have been illiterate.

In the 8th century at the latest, starting from a Semitic model (which had separate signs for the consonants, but none for the vowels), a new system of writing was created by Greeks—the alphabet. For this purpose the list of Semitic consonants was adapted to the needs of Greek phonology, but the major innovation was the invention of five letters with the value of vowels—α(*a*), ε(*e*), ι(*i*), ο(*o*), υ(*u*). The use of the alphabet spread very quickly from east to west across the Greek world. The earliest datable inscriptions, both from around 725 BC, come from Athens (the Dipylon vase) and the colony of Ischia in the Tyrrhenian Sea (the so-called Nestor's cup).

During the period from the 8th to the 5th centuries BC, local differences caused certain details in the forms of the letters to vary from one city to another. Moreover, the primitive Greek alphabet underwent various reforms —the creation of new letters, first φ(*ph*), χ(*ch*), ξ(*ks*), and ψ(*ps*), and η(*ē*) and ω(*ō*). From the 4th century BC on, the alphabet became uniform throughout the Greek world as the result of the general adoption of the form it had taken in Asiatic Ionia.

Greek alphabetic inscriptions

Greek alphabetic inscriptions are numbered in tens of thousands: dedications, epitaphs, decrees, laws, treaties, religious rules, judicial decisions, and so forth. The majority are of Hellenistic or Roman date. The less numerous Archaic inscriptions (8th–5th centuries BC) are of particular interest for their contribution to the knowledge of the dialects (see below). It is only in Hellenistic papyri, and later in Byzantine manuscripts, that the great works of ancient literature (the originals of which have disappeared) come into view in the form of copies, some further and some less far removed from the originals.

The Greek alphabet, still in use today in Greece in the form it reached in the Hellenistic period, has enjoyed an extraordinary success as a direct or indirect model for other alphabets (notably the Latin alphabet); on it are based the writing systems employed in a great part of the modern world.

ANCIENT GREEK

History and development. Only from the 4th century BC, in the Hellenistic period, did Greek approach great unity throughout the area it covered (see below *Koine*). In the preceding ten centuries there were numerous Greek dialects, which differed in phonetic and morphological details, but which were mutually intelligible. The features shared by the local speech of different regions allow the delineation of dialect groups, of which the Greeks themselves were aware. The classifications of modern scholars modify in various ways the classifications made by the ancients, but still retain these as their basis. Among the dialects there are a West group, an Aeolic group, an Ionic-Attic group, and an Arcado-Cypriot group (the last group was neglected by the ancient Greeks because neither Arcadian nor Cypriot gave rise to a literary language). Modern scholars have never questioned the isolation of the West group, but they have tried in various ways to combine the other three into two divisions (*e.g.*, by considering Aeolic and Arcado-Cypriot as varieties of "central" Greek, or by considering Arcado-Cypriot and Ionic-Attic as varieties of "southern" Greek).

In regard to the dialects, two very different situations must be distinguished: that established for the period between the 14th and the 12th centuries BC and that for the period between the 8th and the 4th centuries BC.

In Mycenaean times the carriers of "West Greek" had not yet reached Greece; they did not irrupt into it until the end of the 2nd millennium. In continental Greece (north and south of the Isthmus of Corinth) and on certain Aegean islands (notably Crete), only varieties of Greek other than West Greek were spoken. The tablets reveal a somewhat artificial chancellery language current in the palace offices and taught as a written language in scribal schools. Based essentially on a dialect of the type that was eventually called Arcado-Cypriot, it shows great uniformity in time (during the two centuries or thereabouts covered by documents) and in space (from Knossos to Thebes). Certain fluctuations in details, however, which can be shown to vary between scribes (even at the same site and at the same date), permit the assumption that, behind this official written form of Greek, there must have been various forms of spoken Greek. The problem of the genesis of the dialects other than West Greek does not now present any provable solution.

Influence of Dorian invasions and Greek colonization

There followed two great events that upset the dialectal distribution within the Greek world. First, the Dorian invasions brought speakers of West Greek into northern Greece, then into the Peloponnese, and finally into the Aegean. Some pre-Dorian Greek populations were expelled from their homes and emigrated eastward to the west coast of Anatolia and to Cyprus. Others, who remained where they were, became more or less thoroughly Dorian in speech. It has long been thought that some of the features that Thessalian and, even more, Boeotian (both of which are Aeolic) shared in the 1st millennium with West Greek can be attributed to "recent" influences; on the other hand, some Doric dialects of the 1st millennium (*e.g.*, in Crete) show sporadic traces of features attributable to an Arcado-Cypriot substratum. The other subsequent event, which is of a different sort, was the great colonization movement that began in the 8th century BC. Each group of emigrants took with them the speech of their mother city and planted it in the new foundation. Thus there developed on the shores of southern Italy a totally new grouping of Greek dialects, side by side—Asiatic Ionic at Siris; Euboean Ionic at Rhegium and Cumae; Laconian Doric at Tarentum and Heraclea; Achaean at Sybaris, Croton, and Metapontum; Locrian at Locri Epizephyrii; and so on.

Toward the middle of the 1st millennium BC the geographical distribution of the dialects (insofar as they are known directly through inscriptions) is briefly as follows:

West Group: (1) Doric proper: Messenia, Laconia (colonies—Tarentum, Heraclea); the Argolid; the territory of Corinth (colonies—Corcyra, Anactorium, Syracuse); the Megarid (colonies—Megara Hyblaea, Selinus, Byzantium); the Sporades (colony—Cyrene); Crete; Rhodes (colonies—Gela, Acragas). (2) North-West Greek: Elis, Achaea (colonies—Ithaca, Sybaris, Croton, Metapontum); Aetolia; Phocis; Locris (colony—Locri Epizephyrii).

Aeolic Group: (1) Boeotia; (2) Thessaly; (3) Lesbos and Asiatic Aeolis.

Ionic-Attic Group: (1) Attica; (2) Euboea (colonies—Catana, Zancle, Rhegium, Cumae); (3) Cyclades; (4) Asiatic Ionia (colonies—Siris, Phocaea, foundations in Pontus [Black Sea]).

Arcado-Cypriot Group: (1) Arcadia; (2) Cyprus; (3) Pamphylia.

Table 1: Classical Greek Alphabet and Numerals

letters: capital	letters: lower-case	letters: combinations	name	equivalent: EB preferred	equivalent: alternatives	approximate pronunciation
A	α, α*		alpha	a		bother
		αι		ae in proper nouns, ai in common words	e	ice
		αυ		au		now
B	β		beta	b		baby
Γ	γ		gamma	g		go
		γγ		ng		angle
		γκ		nk	nc	ink
		γξ		nx		thanks
		γχ		nch	nkh	Ger. München
Δ	δ, ∂*		delta	d		dog
E	ε		epsilon	e		bet
		ει		ei	e or i	day
		ευ		eu		fury
Z	ζ		zeta	z		adz
H	η		eta	ē	e	day
		ηυ		ēu	eu	youth
Θ	θ, ϑ*		theta	th		thin
I	ι		iota	i		even or pin
K	κ		kappa	c in proper nouns, k in common words		kin
Λ	λ		lambda	l		lily
M	μ		mu	m		maim
N	ν		nu	n		not
Ξ	ξ		xi	x		ax
O	ο		omicron	o		obey
		οι		oe in proper nouns, oi in common words	oe	boy
		ου		ou		food

letters: capital	letters: lower-case	letters: combinations	name	equivalent: EB preferred	equivalent: alternatives	approximate pronunciation
Π	π		pi	p		pin
P	ρ		rho	initial, rh; medial, r		rose
		ρρ		rrh		arrow
Σ	σ‡		sigma	s		sand
T	τ		tau	t		tie
Y	υ		upsilon	y	u	Fr. rue
		υι		ui		we
Φ	φ, φ*		phi	ph		fifty
X	χ		chi	ch	kh	Ger. Buch
Ψ	ψ		psi	ps		perhaps
Ω	ω		omega	ō	o	bone

numerals

Greek	Arabic	Greek	Arabic	Greek	Arabic
α′	1	ιε′	15	ο′	70
β′	2	ις′	16	π′	80
γ′	3	ιζ′	17	ϙ′†	90
δ′	4	ιη′	18	ρ′	100
ε′	5	ιβ′	19	σ′	200
ϛ′†	6	κ′	20	τ′	300
ζ′	7	κα′	21	υ′	400
η′	8	κβ′	22	φ′	500
θ′	9	κγ′	23	χ′	600
ι′	10	κδ′	24	ψ′	700
ια′	11	λ′	30	ω′	800
ιβ′	12	μ′	40	ϡ′†	900
ιγ′	13	ν′	50	,α	1,000
ιδ′	14	ξ′	60		

*Old-style character. †Special character. ‡Final, ς.

Literature and dialects

This linguistic situation in the first half of the 1st millennium BC resulted in literature developing on a dialect basis. The Homeric epic in the state in which it became fixed by writing displays a mixture of Aeolic and Ionic features. Choral lyric is especially Doric in colouring. Prose developed first in Ionic surroundings (Herodotus, Hippocrates), then in Attica (Thucydides, Plato). Attic is the language of dialogue in tragedy, but alongside Attic comedy there also developed in Sicily a Doric comedy. Personal poetry employs, depending on the author, Ionic (Hipponax), Lesbian (Alcaeus, Sappho), Boeotian (Corinna), and other dialects. It was only in the Hellenistic and Roman periods that Ionic-Attic became clearly dominant, though in poetry of the later periods there were artificial imitations of the early genres.

Within the alphabetical period of Ancient Greek (8th–4th centuries BC), previous to Koine, there is no break between what is termed Archaic Greek (8th–6th centuries) and what is termed Classical Greek (5th–4th centuries). The Classical period is that in which the progressive elaboration of Archaic data brought Greek literature, as well as Greek art, to perfection.

In the linguistic subdivision of Ancient Greek the effects of substratum languages played only a minor part. In their penetration into Greece toward the beginning of the 2nd millennium BC, the Hellenic peoples found earlier populations established there, about whom Greek tradition gives only vague hints, and whose language or languages are unknown. From this "pre-Hellenic" stratum, Greek vocabulary made numerous borrowings (*kyparissos* "cypress," *pyrgos* "tower," and so on), and it received from it a number of place names (*e.g.*, Korinthos); but there is no reason to think that the divergent characters of the Greek dialects (in phonetics and morphology) could be connected with different substrata. The native "barbarian" languages also had little effect on colonial Greek in the 1st millennium, and these contacts show up only in a few local borrowings.

On the other hand, there is a connection between the facts of civilization (in the political and cultural fields) and the evolution of the language. In the Mycenaean period an evident unity of civilization and the organization in the palaces of record offices and scribal schools allowed the use of a stable and uniform chancellery language. In the first half of the 1st millennium, political subdivision and rivalry between cities allowed dialectal peculiarities to strike deep roots. The special conditions in which epic developed, however, resulted in presenting the "noble" literary genres from their inception with a model of dialect mixture. From the 5th century BC onward, the prestige of Ionic and Attic literature and the political authority of Athens opened the way to Ionic-Attic predominance, and this was eventually imposed by the Macedonian conquest.

Linguistic characteristics. *Phonology.* The phonological systems of Ancient Greek differ noticeably from one period to another and from one dialect to another. The system that has been chosen to serve as an example here is that which may be attributed to Old Attic of the 7th–6th centuries BC.

In Old Attic, there are seven vowel qualities: *i*, open and closed *e*, *a*, open and closed *o*, and *u*, each of which has a

long and a short form, except open *e* and open *o*, which have only the long form. Diphthongs originally included *ei*, *ai*, *oi*, and *eu*, *au*, *ou*, but very soon *ei* began to evolve toward long closed *ē* and *ou* toward long closed *ō*. In addition, there is a rare diphthong *ui*, and usually at the end of words the diphthongs *-ēi*, *-āi*, *-ōi*, with preponderant first elements, which later were reduced respectively to long open *ē*, long *ā*, and long open *ō*.

The consonantal structure is characterized by relative richness in stops (sounds produced by momentary complete closure at some point in the vocal tract): unvoiced *p*, *t*, *k*, aspirated *ph*, *th*, *ch*, voiced *b*, *d*, *g*; and by few spirants: only *s* and *h* sounds (*h* restricted to initial position before a vowel). There is also a voiced affricate sound, *dz;* two liquid sounds, *l* and *r;* and two nasals, *m* and *n*. The guttural nasal is not distinctive, but is only a variant of the sound *n* in front of a guttural stop. Neither *y* nor *w* occur as distinctive sounds. All of the consonants except *h* and *dz* can be doubled between vowels. The only consonant sounds normally allowed at the end of the word are *-s*, *-n*, and *-r*.

Word accent

Apart from some unaccented monosyllabic or disyllabic terms of minor importance, each word is marked by a rise in the musical pitch of the voice (accent) on one of the vowels (one of the last three vowels, if the word has more than three syllables). Short vowels, if they carry the accent, have only a rising tone (noted from the Alexandrian grammarians onward by the sign of the acute accent); long vowels or diphthongs may have either a rising tone (noted by the acute accent) or a rising tone followed by a falling tone (noted by the circumflex accent). Within a phrase the vowel of a final syllable with a normally rising tone is weakened in accent (noted by the grave accent). The position and nature of the accent in the word are governed by rules so strict that they do not usually permit variations that would serve to differentiate two otherwise identical forms. There are, however, examples of such a differentiation: *oîkoi* ("houses") is a nominative plural form, and *oíkoi* ("at home"), an adverb of place; *tómos* means "a cut" and *tomós* "cutting"; and so on.

The accent (which is not associated with stress) does not play any part in the rhythm of the language. This rhythm (and that of poetry, which is a stylized form of it) is based upon the distribution in the sentence (and in the verse) of short and long syllables. For a syllable to be short, it must end in a short vowel; syllables ending in a long vowel, or closed syllables (*i.e.*, those ending in a consonant), are long. The rhythm of Ancient Greek is therefore quantitative.

Morphology. Every nominal (noun) or verbal form combines a "root" that carries the sense of the word and a certain number of grammatical markers that serve principally to define the function of the word in the phrase.

The category of gender, which differentiates masculine, feminine, and neuter, is expressed only in the substantive (noun), adjective, and pronoun. The category of person (1st, 2nd, and 3rd person) is restricted to the personal pronoun and the verb. There are three numbers—singular, dual, and plural—that are distinguished in both the noun and the verb. The survival of the dual is an archaism; although a living form in the Mycenaean period, it tends to be replaced by the plural in the 1st millennium. Attic is one of the dialects in which it is best preserved down to the threshold of the Hellenistic period.

Not counting the vocative case, the Greek declension in the Mycenaean period still contained at least six cases: nominative, accusative, genitive, dative, locative, and instrumental. Between the Mycenaean period and the 8th century the locative and the instrumental ceased to exist as living cases, their functions having been taken over by the dative.

Role of aspect in the Greek verb

The most original feature of Greek morphology is the structure of the verb system, which is determined fundamentally by the category of aspect. It is organized around three principal themes (or tenses) for each verb: the "present" theme for the durative aspect, the aorist theme for the punctual aspect, and the perfect theme for the aspect of completion. Each of these themes provides an indicative and, apart from the imperative, two nonassertive moods (subjunctive and optative). The expression of time exists only in the indicative; there is, on the one hand, a future theme, and on the other, an aorist indicative theme (which always represents past time). There are also past tenses attached respectively to the present indicative (imperfect) and to the perfect indicative (pluperfect). The past tense has as its distinguishing marks a certain set of endings, called secondary (which it shares with the optative), and the presence of a special preverb called the augment. The Greek verb has two voices (active and mediopassive), which are expressed (leaving aside the aorist passive) by the opposition of two series of endings. Finally, in each voice, complete series of participles and infinitives were established corresponding to the present, future, aorist, and perfect themes.

Syntax. A relatively free word order occurs in Greek. Above all, the creation of the definite article (post-Mycenaean and post-Homeric), and the various ways in which the nominal forms of the verb (participles and infinitives) come into play, confer on the Greek sentence a suppleness unmatched in other languages.

Vocabulary. If one considers the roots of words, it seems that, although the essential basis of the vocabulary is of Indo-European origin, a fairly considerable number of terms are borrowings. Most of these loans were taken from the idioms of the populations living in Greece prior to the arrival of the Greeks; many such words had already penetrated into Greek in the 2nd millennium, for there are forms found in Mycenaean that correspond to plant names such as *elaia* "olive," *pyxos* "box tree," and *selinon* "celery"; animal names such as *leōn* "lion" and *onos* "ass"; names for objects such as *asaminthos* "bathing tub," *depas* "goblet," and *xiphos* "sword"; and names of materials such as *elephās* "ivory," *chrysos* "gold," and *kyanos* "dark blue enamel."

The most important fact is that from the verbal and nominal roots (of whatever origin) the language extracted a vocabulary full of nuances and of great scope (by using preverbs, and by forming compounds and derived words). At all periods the lexical creativity of Greek has been very active, thus giving it a vocabulary of extraordinary richness. (M.Le.)

KOINE

The fairly uniform variety of spoken Greek that gradually replaced the local dialects after the breakdown of old political barriers and the establishment of Alexander's empire in the 4th century BC is known as the Koine (*hē koinē dialektos* "the common language"), or "Hellenistic Greek." Attic, by virtue of the undiminished cultural and commercial predominance of Athens, provided its basis; but as the medium of communication throughout the new urban centres of Egypt, Syria, and Asia Minor, it absorbed numerous non-Attic elements and underwent some degree of grammatical simplification. Numerous inscriptions enable scholars to trace its triumphant progress at the expense of the old dialects, at least as the language of business and administration, although some rural dialects are reported to have survived as late as the 2nd century AD. Other sources of information for the Koine are the translation of the Septuagint made in the 3rd century BC for the use of the Hellenized Jewish community of Alexandria, the New Testament, and the writings of a few people (*e.g.*, the historian Polybius and the philosopher Epictetus) who favoured it over Attic. As the everday colloquial language of urban Egypt it may be studied in papyri going back to the 4th century BC. The Koine may be dated very crudely from the period of Alexander's conquests in the 4th century BC to approximately the reign of Justinian in the 6th century AD.

Sound changes from Ancient Greek to the Koine

The Koine replaced the Attic sound *tt* by the *ss* characteristic of Ionic and other dialects (*e.g.*, *glōssa* for *glōtta* "tongue") at an early date, but its main phonological significance lies in its gradual simplification of the rich vowel system of Classical Greek. Ancient *ei* (ει), *i*(ι), and *ē*(η) sounds merged as *i*, and *ai*(αι) was monophthongized to *e*; *oi*(οι) became pronounced as the sound symbolized

by *ü*, thus merging with simple *y* (pronounced as in French *tu*). The second element of *au* (αυ) and *eu* (ευ) was changed to *v* or *f* (compare ancient *autós* to modern *aftós* "he"). These shifts, combined with the loss of length distinctions, led to a new six-term system of vowel sounds: *i, u, ü, e, o, a*. The loss of *h* also belongs to the Koine period, and there is evidence that the change of the ancient aspirates and voiced stops to fricative (spirant) sounds was well under way. As a result of this latter process, Classical *ph, th, ch* (pronounced as in English "pin," "tin," "kin") acquired the fricative articulations of "fin," "thin," and the final element of Scottish "loch," (or German *Buch*); *b, d, g* became the voiced fricative sounds *v, dh* (as in "that"), and *gh* (as in Spanish *fuego*).

Grammar too began to move in the direction of Modern Greek in this period. Nouns in consonant stems acquired the endings of the *-a* declension; *e.g., thygatēr*, "daughter," accusative *thygatera*, was remodelled after items such as *khōra, khōran* "country." The dual number was lost in nouns, verbs, and adjectives, as was the optative mood of verbs. Confusion arose between the perfect and aorist tense forms, leading to the loss of one or the other (the former in most verbs).

In vocabulary there were numerous borrowings from non-Attic dialects, and Attic words acquired new meanings; thus *opsaria* "fish" and *brechei* "it rains" for Classical Greek *ichthyes* and *hyei* both occur in the New Testament (*cf.* Modern Greek *psárya, vrékhi*).

This gradual divergence from the language of Plato and Demosthenes was viewed as a species of linguistic decadence by an influential school known as the Atticists, who unceasingly castigated the use of Koine forms by writers. It was thus that there developed a rift between the everyday spoken language and an archaizing, specifically written language. It became fashionable to publish manuals of "good usage" in which the Attic equivalents of Koine innovations were recommended for the student's imitation.

BYZANTINE GREEK

During the period of the Byzantine Empire (*i.e.*, until the fall of Constantinople in 1453) the language of administration and of most writing was firmly rooted in the Atticist tradition; it is this archaizing style that is often referred to as "Byzantine Greek." The spoken language continued to develop apace, however, and its course can be followed to some extent in the writings of the less educated chroniclers (such as Malalas, 6th century) and hagiographers. Furthermore, the increasing political and military disintegration characteristic of the last few centuries preceding the fall of Constantinople brought with it a general decline in educational level, and works appeared that reflect quite closely the colloquial language of the time, although learned and pseudolearned elements are never absent. While the differences between the *Chronicle of the Morea* (13th century), for example, and present-day spoken Greek are quite minor, Byzantium failed to produce a writer of the stature of Dante, capable of establishing once and for all the living vernacular as a worthy vehicle for great literature.

Changes occurring in the Koine

Most of the phonological and grammatical developments that separate present-day Greek from the Koine occurred during this period. The frequent misuse of the dative case of nouns shows that it went out of use in the spoken language, and the infinitive was replaced by various periphrastic constructions. (Periphrastic constructions involve the use of function words and auxiliaries.) In the early period numerous words (mostly Latin) were imported: the chronicler Malalas has (in their modern form) *pórta* "door," *kámbos* "plain," *saíta* "arrow," *paláti* "palace," *spíti* "house" (from *hospitium*), and hundreds of other borrowings, not all of which have survived. The later period is characterized by the richness of its compound words, usually from native roots. Some of these continued ancient patterns, such as that in which a modifying noun is linked to its head noun by *-o-* (*thalassovrákhi* "sea rock," *vunópulo* "mountain lad"); but coordinative compounds of the type common today are also found (*e.g.*, Modern Greek *andróyino* "man and wife," *makheropíruna* "knives and forks"). Semantic shift was another source of innovation: *álogho* "horse," previously meant "irrational"; *skiázome* "I fear," earlier was "I am in shadow"; and *(u)dhén* "not," was, in Classical Greek, "nothing."

MODERN GREEK

History and development. Modern Greek derives from the Koine via the local varieties that presumably arose during the Byzantine period, and is the mother tongue of the inhabitants of the Kingdom of Greece and of the Greek majority in the Republic of Cyprus. Before the exchange of populations (1923) there were Greek-speaking communities in Turkey (Pontus and Cappadocia), and it remains the language of the Greek community of Istanbul. Certain villages in Calabria in southern Italy are also Greek-speaking. Three main varieties may be distinguished: (1) the local dialects, which may differ from one another virtually to the point of mutual unintelligibility, (2) the standard colloquial Greek spoken in all the urban centres of Greece, known as demotic, and (3) Katharevusa (from *katharós* "pure"), a strictly literary language.

Local dialects. Of the local dialects, Tsakonian, spoken in certain mountain villages in eastern Peloponnese, is quite aberrant and shows evidence of descent from the ancient Doric dialect (*e.g.*, it often has an *a* sound for the early Greek *ā* that went to *ē* in Attic, later to *i*). The Asia Minor dialects also display archaic features (*e.g.*, Pontic *e* for ancient *ē* in certain word elements). It is not certain whether southern Italian Greek represents a survival from ancient times or was reimported there during the Byzantine period. Apart from these peripheral varieties, the modern dialects may be grouped for practical purposes as follows:

Divisions of the modern dialects

1. Peloponnesian, differing but slightly from the dialects of the Ionian isles, forms the basis of standard demotic. It shows very few specifically local innovations in its phonology, although its verb morphology is less conservative than that of the island dialects.

2. Northern dialects, spoken on the mainland north of Attica, in northern Euboea, and on the islands of the northern Aegean, are characterized by their loss of unstressed *i* and *u* and the raising of unstressed *e* and *o* sounds to *i* and *u*. Thus, standard *kotópulo* "chicken" becomes *kutóplu, émine* "he stayed" becomes *émni*. They also mark certain 1st and 2nd person plural past tense verb forms with *-an* (*ímastan* "we were," Athenian *ímaste*) and use the accusative for indirect object pronouns in instances in which the southern dialects have the genitive (*na se pó* "let me tell you," standard *na sou pó*).

3. Old Athenian was spoken in Athens itself until it became the capital of the modern state (1833), and on Aegina until early in the 20th century; it survives in Megara and in the Kími district of central Euboea. Its salient feature is the replacement of the Byzantine *ü* sound (from ancient *y, oi*) by *u* rather than normal *i*; it changes the *k* sound before the vowels *e* and *i* to *ts* and fails to contract the vowels *i* and *e* to a *y* sound before vowels (ancient *sykéa* becomes *sutséa* "fig tree," standard *sikyá*).

4. Cretan softens *k* to a *č* sound (as in "church"), *kh* to *š* (as in "she") before *i* and *e*, and *y* to *ž* (as the *s* in "pleasure")—*e.g., če* "and," *šéri* "hand," *žéros* "old man," standard *ke, khéri, yéros*.

5. The southeastern dialects of Cyprus, Rhodes, Chios, and other islands in the area soften *k* to *č* as in Crete, drop voiced fricative consonants between vowels, and retain the ancient final *-n* (*láin* "oil," standard *ládhi*). They also retain the contrast between long and short consonants (*fíla* "kiss!" but *fílla* "leaves"). As is done in Cretan and Old Athenian, they add *gh* to the *-ev-* that occurs at the end of many verb stems (*dhulévgho* "I work").

Demotic. Demotic is the language used for everyday conversation in the towns of mainland Greece, and is understood without difficulty by all speakers. Differences within the demotic form as it is spoken in the various parts of the country are so minor that it may be regarded as the standard spoken language. In all essential respects

Table 2: Modern Greek Alphabet

Greek letters capital	Greek letters lower case	Greek letters combinations	name	equivalents	approximate pronunciation
A	α, ά*		álfa	a	bother
		αι		ai	bed
		αυ		av	Slav, laugh†
B	β		víta	v	van
Γ	γ		ghámma	g before a, o, ου, ω and consonants other than γ, ξ, and χ; y before αι, ε, ει, η, ι, οι, υ, and υι; n before γ, ξ, and χ	wit, yet, sing
		γκ		initial, g; medial, ng	go, finger
Δ	δ, ∂*		dhélta	dh; d between ν and ρ	then, wondrous
E	ε		épsilon	e	bet
		ει		i	even
		εϊ		eï	day
		ευ		ev	left or revel
Z	ζ		zíta	z	zone
H	η		íta	i	fig
		ηυ		iv	even, leaf
Θ	θ, ϑ*		thíta	th	thin
I	ι		ióta	i	even
K	κ		kápa	k	kin, cook
Λ	λ		lámbdha	l	lily
M	μ		mi	m	maim
		μπ		initial, b; medial, mb	baby, ambush
N	ν		ni	n	not
		ντ		initial, d; medial, nd	dog, fender
		ντζ		ntz	chintz
Ξ	ξ		xi	x	ax
O	ο		ómikron	o	saw
		οι		oi	even
		ου		ou	food
Π	π		pi	p	pin
P	ρ		ro	r	rose
Σ	σ‡		sígma	s	sand
T	τ		tav	t	tie
Y	υ		ípsilon	i initially and between consonants	even
		υι		i	even
Φ	φ, φ*		fi	f	fifty
X	χ		khi	kh	Ger. Buch
Ψ	ψ		psi	ps	perhaps
Ω	ω		oméga	o	bone

*Old-style character. †Pronounced with long *a*. ‡Final, s.

its phonology and grammar follow average Peloponnesian practice, but it has absorbed a vast number of vocabulary items from learned sources. It is also used as the vehicle of poetry and, since the beginning of the 20th century, of fiction, although, except for certain genres such as drama and detective novels, the spoken language is not necessarily reproduced with any fidelity. Indeed, one may speak of a specifically literary demotic that differs from the spoken language in its extensive use of words culled from local dialects, its fondness for innovation in compound formation, and its somewhat eclectic verb morphology.

Use of Katharevusa

Katharevusa. Katharevusa is the purist, archaizing written language of administration; it is also used in technical publications, newspapers, and public notices. Its role in education has varied with the policies of successive governments; since 1967 it has been the official language of education beyond the elementary school. Although the concept of a distinctive written language based on earlier usage goes back to the Atticists, Katharevusa originated in the 19th century as a result of the effort to purify the local dialects of foreign elements and to systematize their morphology (inevitably on the Classical Greek model). Thus, Katharevusa is characterized by its exclusive use of Ancient Greek roots and much Classical inflection, while its syntax and idiom differ but slightly from those of demotic (this is true, at least, of the "simple Katharevusa" current in today's newspapers and periodicals). Katharevusa elements abound in demotic, often with a specialized role; for example, *édhra* (from the Ancient Greek word for "chair") means "professorial chair," the demotic word for "chair," *karékla*, being the term for the article of furniture. Many Katharevusa terms appear almost exclusively in print: *zíthos* "beer" and *ínos* "wine" are common in advertisements, but everyone says *bíra* and *krasí*. Words of Katharevusa origin often show ancient inflections: *kathiyitís* "professor" has the vocative form *kathiyitá*, *fititís* "student" is in the plural *fitité* (*cf.* demotic *traghudhistís* "singer," plural *traghudhistés*, *traghudhistádhes*). Because of its use in newspapers and news bulletins, most Greeks have a good passive knowledge of Katharevusa and it is easily accessible to foreigners with reasonable competence in the classical language.

Linguistic characteristics. *Phonology.* Modern Greek has five distinct vowel sounds (*i, e, a, o, u*) and the glide *y*, most of which are indicated in Greek orthography in more than one way. The consonant sounds are:

Voiceless stops	*p*	*t*		*k*
Voiced stops	*b*	*d*		*g*
Voiceless fricatives	*f*	*th*	*s*	*kh*
Voiced fricatives	*v*	*dh*	*z*	*gh*
Nasals	*m*	*n*		
Liquids	*l*	*r*		

The sounds *f*, *th*, and *kh* derive from ancient aspirated consonants, and the voiced fricative sounds *v*, *dh*, and *gh* from *b*, *d*, and *g*. Modern *b*, *d*, and *g* are usually created by the voicing of *p*, *t*, and *k* after nasals; thus Ancient Greek *pénte* "five" becomes *pénde*. These also occur at the beginning of words in place of ancient nasal + stop sequences (*boró* "I am able" from *emporó*). Other important combinatory changes linking Ancient and Modern Greek include the following: (1) Ancient stop clusters and aspirate clusters both become fricative + stop; *e.g.*, *hepta* "seven," *oktō* "eight," *ophthalmos* "eye" become *eftá*, *okhtó*, *ftarmós* ("evil eye"). (2) Double consonants are simplified except in the southeast, thus *thalassa* becomes *thálasa* "sea." (3) Nasal sounds assimilate to following fricative sounds; thus *nymphē* becomes *níffi* and then (except in the southeast dialects) it changes to *nífi* "bride." (4) The sound *l* is replaced by *r* before consonants; *e.g.*, *adelphos* becomes *adherfós* "brother." (5) Before a vowel, *i* and *e* change to *y*; thus *paidia* becomes *pedhyá* "boys," *mēlea* becomes *milyá* "apple tree." Except for the simplification of double consonants, these statements do not apply to words of Katharevusa origin. Thus in *simfonía*, meaning "symphony" or "agreement," statements (3) and (5) are violated (the true demotic form would be *sifonyá*).

Stress in Modern Greek

Modern Greek has dynamic stress (as in English) and not the ancient musical accent, so that while words may be distinguished by stress placement (*fíli* "friends," *filí* "kiss"), the old distinction between grave, acute, and circumflex is lost (but still represented in written accents).

Grammar. Much of the inflectional apparatus of the ancient language is retained in Modern Greek. Nouns may be singular or plural—the dual is lost—and all dialects distinguish a nominative (subject) case and accusative (object) case. A noun modifying a second noun is expressed by the genitive case except in the north, where a prepositional phrase usually replaces this. The indirect object is also expressed by the genitive case (or by the preposition *se* "to," which governs the accusative, as do all prepositions). Thus:

o yatrós	*ípe tin istoría*	*s ton adherfó tiz dhaskálas*
"The doctor	told the story	to the brother of the teacher"
(nominative)	(accusative)	(genitive)

The ancient categorization of nouns into masculine, feminine, and neuter survives intact, and adjectives agree in gender, number, and case with their nouns, as do the articles (*o* "the," *enas* "a"). In general, pronouns exhibit the same categories as nouns, but the relative pronoun *pu* is invariant, its relation to its own clause being expressed when necessary by a personal pronoun in the appropriate case: *i yinéka pu tin ídhe to korítsi* "the woman *pu* her saw the girl" (*i.e.*, "the woman whom the girl saw").

The verb is inflected for mood (indicative, subjunctive, imperative), aspect (perfective, imperfective), voice (active, passive), tense (present, past), and person (1st, 2nd, 3rd, singular and plural). The future is expressed by a particle *tha* (from earlier *thé[o] na* "[I] want to") followed by the subjunctive. There are also two participles, an indeclinable present active one in *-ondas*, which is confined to certain individual usages (*tróghondas érkhete i óreksi* "in eating comes the appetite"), and a past passive one in *-ménos* (*kurazménos* "tired"). Formally, the finite forms of the verb (those with personal endings) consist of a stem + (optionally) the perfective aspect maker (*-s-* in active, *-th-* in passive) + personal ending (indicating person, tense, mood, voice). Past forms are prefixed by *e-* (the "augment"), although this is usually lost in mainland dialects when unstressed. In the active, the present endings are *-o, -is, -i, -ume, -ete, -un*(e) (*-usi* in southeast dialects), and the past endings are *-a, -es, -e, -ame, -ate* (Peloponnesian, standard, elsewhere *-ete*), *-an*(*e*). In both active and passive paradigms only six tenses are distinguishable in pronunciation. The active forms of "you (singular) write," for example, are:

	Present	Past	Imperative
Imperfective	(1) *ghráfis*	(3) *éghrafes*	(5) *ghráfe*
Perfective	(2) *ghrápsis*	(4) *éghrapses*	(6) *ghrápse*

Ghráfis "you write" represents both the present indicative and imperfective subjunctive; *ghrápsis* "you (may) write" is the perfective subjunctive, and *éghrafes* and *éghrapses* are the imperfect indicative "you were writing" and the preterite "you wrote," respectively.

Aspect in Modern Greek

Aspectual differences play a crucial role. Roughly, the perfective marker indicates completed, momentary action; its absence signifies an action viewed as incomplete, continuous, or repeated. Thus the imperfective imperative *ghráphe* might mean "start writing!" or "write regularly!" while *ghrápse* means rather "write down!" (on a particular occasion). Compare also *tha ghrápho* "I'll be writing" but *tha ghrápso* "I'll write" (once). The difference is sometimes represented lexically in English: *ákuye* "he listened" and *ákuse* "he heard." The passive forms are largely confined to certain verbs active in meaning like *érkhume* "I come," *fováme* "I am afraid," and reciprocal usages (*filyóndusan* "they were kissing"). There are also phrasal constructions representing completed action: *ékho ghrápsi* "I have written" (standard), *ékho ghramméno* (in most dialects). These are, however, much rarer than the corresponding English perfect forms. There is no infinitive; ancient constructions involving it are usually replaced by *na* (from ancient *hína* "so that") + subjunctive. Thus *thélo na ghrápso* "I want to write," *borí na ghrápsi* "he can write." Indirect statement is introduced by *oti* or *pos* (*léi oti théli* "he says that he wants").

Vocabulary. The vast majority of demotic words are inherited from Ancient Greek, although quite often with changed meaning; *e.g.*, *filó* "I kiss" (originally "love"), *trógho* "I eat" (from "nibble"), *kóri* "daughter" (from "girl"). Many others represent unattested combinations of ancient roots and affixes; others enter demotic via Katharevusa: *musío* "museum," *stikhío* "element" (but inherited *stikhyó* "ghost"), *ekteló* "I execute." In addition, there are over 2,000 words in common use drawn from Italian and Turkish (accounting for about a third each), and from Latin, French, and, increasingly, English. The Latin, Italian, and Turkish elements (mostly nouns) acquire Greek inflections (from Italian *síghuros* "sure," *servitóros* "servant," from Turkish *zóri* "force," *khasápis* "butcher"), while more recent loans from French and English remain unintegrated (*spór* "sport," *bár* "bar," *asansér* "elevator," *futból* "football").

(B.E.N.)

BIBLIOGRAPHY

Ancient Greek: M. VENTRIS and J. CHADWICK, *Documents in Mycenaean Greek: Three Hundred Selected Tablets from Knossos, Pylos, and Mycenae, with Commentary and Vocabulary* (1956), a study of both the writing system and the content of the tablets, by the authors of the decipherment; L.H. JEFFERY, *The Local Scripts of Archaic Greece* (1961), a description of all the local varieties of the Greek alphabet from the 8th to the 5th century BC; C.D. BUCK, *The Greek Dialects* (1955), a summary of the dialectal features of Ancient Greek within the scope of a traditional descriptive grammar; A. MEILLET, *Aperçu d'une histoire de la langue grecque*, 7th ed. (1965), the first and still fundamental endeavour to define the characteristics of Greek in a diachronic perspective; E. SCHWYZER and A. DEBRUNNER, *Griechische Grammatik*, 3 vol. (1939–53), a complete and accurate description with exhaustive bibliography; P. CHANTRAINE, *La Formation des noms en grec ancien* (1933), deals with the history of noun suffixes throughout the history of Greek; H. FRISK, *Griechisches etymologisches Wörterbuch*, 2 vol. (1954–70), up to date and wisely selective (but often underrating Mycenaean data).

Koine and Byzantine: For Koine, see F. BLASS and A. DEBRUNNER, *Grammatik des neutestamentlichen Griechisch*, 10th ed. (1959; Eng. trans., *A Greek Grammar of the New Testament and Other Early Christian Literature*, 1961), a translation and revision of a classic work by R.W. FUNK. T.M. DAWKINS, "The Greek Language in the Byzantine Period," in N.H. BAYNES and H.ST.L.B. MOSS, *Byzantium* (1948); and R. BROWNING, *Medieval and Modern Greek* (1969), cover later periods.

Modern Greek: S.A. SOFRONIOU, *Teach Yourself Modern Greek* (1962); and J.T. PRING, *A Grammar of Modern Greek on a Phonetic Basis* (1950), are good elementary introductions. There is no reasonably complete dictionary of literary demotic, but J.T. PRING, *The Oxford Dictionary of Modern Greek* (*Greek-English*) (1965), is adequate for the spoken language. In general, English–Greek dictionaries are intended for Greeks and fail to mark the stylistic level of Greek equivalents; A.N. JANNARIS, *A Concise Dictionary of the English and Modern Greek Languages* (1895, frequently reprinted), has not been improved on. F.W. HOUSEHOLDER, K. KAZAZIS, and A. KOUTSOUDAS, *Reference Grammar of Literary Dhimotiki* (1964), is useful; and for a convenient summary of the development of demotic, see R. Browning (cited above). For dialects, see B.E. NEWTON, *The Generative Interpretation of Dialect: A Study of Modern Greek Phonology* (1972).

(M.Le./B.E.N.)

Greek Law

Any survey of ancient "Greek law" has to begin with a caution: there never was a "Greek law" in the sense of a system of institutions recognized and observed by the nation as a whole as its legal order. There were, however, a number of basic approaches to legal problems, certain methods used in producing legal effects, and a legal terminology, all shared to varying degrees by the numerous independent states constituting the Hellenic world. To the extent that concepts, institutions, and techniques may be attributed to common heritage or otherwise considered as typical expressions of Greek ways of thinking, the postulate of the existence of a "Greek law" appears to be legitimate—though it remains controversial. It should not be forgotten, however, that such common foundations as there were gave rise to a great variety of individual legal systems differing as to their completeness and elaboration and reflecting the tribal (*i.e.* Dorian, Ionian, etc.) and historical backgrounds as

well as the changing social, economic, political, and intellectual conditions of their respective societies.

Disconnected details are known from many of these systems, and knowledge of them will most probably increase as legal research into the Greek inscriptions proceeds. But only two systems of the pre-Hellenistic period are sufficiently documented to admit of real insight into their nature. A famous inscription containing the 5th-century-BC legislation of the Cretan city of Gortyn detailed many institutions of that city. Even more important is the Athenian law of the later 5th and 4th centuries BC. It is not only the best documented and most thoroughly explored system but also seems to have been the most highly developed and therefore perhaps the most truly representative of the legal culture of Classical Greece. This article will be based primarily on the Athenian evidence, but with glances at the other systems, in so far as these feature characteristic deviations.

THE SOURCES

The best sources of information are speeches written by Athenian orators of the 5th and 4th centuries for the purpose of being delivered at trials. Isaeus and Demosthenes were pre-eminent figures but others, such as Antiphon, Lysias, Isocrates, and Hyperides, as well as some anonymous rhetoricians included in the *Corpus Demosthenicum,* also contributed important material still extant. It should be noted that forensic oratory was what came closest in Greek literature to professional writing on legal matters. Such works as Aristotle's treatises on constitutions (*Athenian Constitution, Politics*) or Theophrastus' treatment of private law topics in his "On the Laws" (only a brief fragment of which has been preserved) were mainly confined to mere description or philosophical analysis and criticism or both. Within these limits they do, of course, throw much light on the institutional framework, though not on the practical workings, of contemporary systems.

Sources of knowledge

Information on positive law that can be gathered from other philosophical writings and from general literature is more limited. Plato's *Laws* and Aristotle's *Nicomachean Ethics* and *Rhetoric* suggest some conclusions as to institutions and practices forming the background of the authors' theories. Beyond that, however, legal history profits little from them. Modern research has established that Plato's suggestions concerning legal policy and Aristotle's philosophical analysis of justice exerted no significant influence on court practice in either Classical or Hellenistic times, while the impact that they may have had on legislators of the latest Classical period (*e.g.*, in Athens, Demetrius of Phaleron) and of Hellenistic times is difficult to ascertain. Sporadic, though often valuable, bits of knowledge, however, can be gleaned from poetic and historical literature. Among the former, the epic, the chief source for the Archaic period, and Attic Comedy, especially Menander, deserve mention.

A nonliterary but rich body of sources are inscriptions, coming from every part of the Greek world and covering every period of history beginning before 600 BC. Recording, or referring to, enactments and other juridical acts, such as treaties, as well as private agreements, wills, manumissions, court decisions, etc., these have provided hundreds of texts that offer a most varied picture of the actual life of the law. Their particular value lies in the information they contain on non-Attic laws, for which these inscriptions are often the only evidence available. Save for a few exceptions, such as the law code of Gortyn, juridical research in inscriptions is still in its beginning stage.

A third group of sources, namely, the thousands of papyrus documents discovered in Egypt and, to a much lesser degree, in Mesopotamia and Palestine, is confined to the Hellenistic and Roman periods and will be considered elsewhere (see HELLENISTIC LAW).

GENERAL CHARACTER OF GREEK LAW

Greek legal life of the 5th and 4th centuries was determined by three dominant factors. One, already referred to, was the existence of a multiplicity of city-states (*poleis*), each of which possessed and administered its own set of laws. The second element was the fact that in many, if not most, of the *poleis* (one certain exception was Sparta) the laws were laid down in written statutes, some of them elaborate and more or less complete codes setting forth procedural methods and substantive rules for the administration of justice. This was the result of a great movement for codification that from the 7th century had swept the Greek world. Solon of Athens (594 BC), who had been preceded in 621 by Draco, is the best known of a number of famous lawgivers, other outstanding ones being Zaleucus of Locri Epizephyrii (south Italy) and Charondas of Cantana; Lycurgus of Sparta is considered legendary. A number of enactments rightly or wrongly attributed to Solon still are known from literary quotations rendering them in a modified form that reflects a legislative reform of 403–402 BC. One of the Draconian laws has been preserved in an Attic inscription giving it in a revised version dating from 409 to 408 BC. The law code of Gortyn, which is itself the revised version of an older code, is the only one that comes close to being fully preserved.

The third determining factor for Greek law was the absence of a body of jurisprudence comparable to that of the Romans. Even the Attic orators, for all their practical familiarity with the laws of the city, were mainly interested in presenting arguments suited to persuade the mass juries before whom they had to argue, not in analyzing the legal system with the object of obtaining a deeper insight into its implications. Nor, for that matter, did the philosophers care for the law as it was, their aim being the discovery of abstract standards of justice.

The three factors above had important effects on the general character of Greek law. The first two resulted in a rather stiff positivism. Contrary to views held by scholars until recently, new research has shown that the Athenian dicasts who sat in judgment did not feel free to base their verdicts on vague notions of equity, but adhered, at least in theory, to the literal meaning of the written statutes (*nomoi*), which they were bound by a solemn oath to observe. This somewhat narrow clinging to literal interpretation, combined with the absence of any attempt to deal with statutes or legal situations in an analytical manner, led to the result that Greek law never attained anything comparable to the doctrinal refinement of Roman law, notwithstanding the remarkable technical flexibility that characterized it in Hellenistic times.

INSTITUTIONS OF PRIVATE LAW

Persons, family, succession. *Status.* In the classical Greek *polis* only the grown-up male citizen was in unrestricted possession of public and private rights. Each city-state laid down its own requirements for citizenship; democratic Athens, for example, insisted, as a matter of principle, on legitimate descent from citizens on both sides, although naturalization was allowed occasionally. Resident aliens who were officially admitted, such as the Athenian metics (*metoikoi*), enjoyed full protection of their freedom and property, but might be subject to some discrimination in addition to being deprived of political rights; in Athens they could not live in legitimate wedlock with a citizen or own real property. The exigencies of its maritime trade induced 4th-century Athens to allow nonresident foreign merchants to sue and be sued before special courts sitting for this purpose in Piraeus. Normally, however, nonresident aliens had no access to the courts of the city and depended on a friendly citizen (*proxenos*) to act on their behalf.

Status of women

The status of women varied from city to city. Athens, in accordance with its closely knit family system, subjected them to the head of the house who as their *kyrios* ("master") had power over their share in the family estate and was entitled to give them in marriage. Under a family law less rigidly patriarchal, women might own property and be without a *kyrios;* such was the case in Gortyn and some other non-Attic states.

Slavery existed everywhere. It ranged from unlimited

chattel slavery, making the slave the full property of his master (Athens), to various degrees of serfdom (*e.g.*, Spartan helotry or the status of the villein, *voikeus,* in Gortyn) that curtailed the freedom of the serf and usually subjected him and his family to a personal master, without, however, reducing him to a mere chattel. This, of course, concerned only his legal condition and did not necessarily involve a more tolerable social status. Manumission, apparently permitted everywhere, bestowed personal freedom, in many places by putting the freedman under divine protection, but did not result in his obtaining citizenship. Special obligations were often imposed on freedmen by their manumissors or even by the law itself.

The concept of *oikos*

Family. The family law of the classical *polis* was based on the concept of *oikos.* In its original meaning of "house" the term signified the individual citizen's family, forming a collectivity that rested on special religious ties and duties following from common descent as well as on a common patrimony in which the several members had actual or potential shares. When secularism grew in 4th-century Athens, the term took on its second and more materialistic meaning of "estate." After a citizen's death his *oikos* was carried on by his direct successors, upon whom the estate and the responsibility for the religious duties of the family devolved.

The prerequisites of membership in the *oikos* were not uniform in all the *poleis.* In the democratic period of Athens, it was confined to legitimate sons (*gnēsioi*) and further descendants of the head of the house in the male line, or, male offspring lacking, to the sons of an *epiklēros* (a brotherless daughter). A man without sons was allowed to adopt a son. Elsewhere the rigid restriction of membership to legitimate offspring did not prevail. In the absence of *gnēsioi,* some states granted succession to *nothoi,* who were the sons of concubines, this having also been the rule in predemocratic Athens. In Gortyn, adoption was permitted despite the presence of *gnēsioi,* but the adoptee acquired only the somewhat reduced status of a daughter.

Diversities such as these are indicative of the existence of several types of collective family organization reflecting differences in the basic character of their respective societies. Another confirmation of this inference, which is also suggested by the discrepancies regarding the status of women pointed out above, is found in the rules governing the management of the patrimony in different places. In Athens, where order was founded on the strict principle of paternal power, the father alone as *kyrios* was entitled to dispose of property belonging to the house. Even grown-up sons, though enjoying control of their own possessions, in a marked contrast to their status under the Roman *jus civile,* were denied any influence on the management of the family estate during their father's lifetime. A principle quite opposite to this Athenian concentration of authority emerges, to cite only one example, from numerous inscriptions found at Delphi, where title to the estate was apparently vested in the group as a whole, including the women. Each of its members, alone or conjointly with others, seems to have been in a position to alienate common property, unless checked by the veto of another member possessing equal rights.

Marriage and divorce. The legal principles governing marriage derived from its function of producing legitimate offspring for the perpetuation of the *oikos. Epigamia* (*i.e.*, the ability of the spouses to live in lawful matrimony) was an indispensable prerequisite of a valid marital union. It was sometimes withheld for political reasons; thus democratic Athens, while liberal regarding marriages of kinsfolk and even condoning cohabitation with a half-sister, excluded all noncitizens from *epigamia.* The establishment of a marital union was a concern of the families rather than the spouses themselves. Athens, therefore, regularly required a solemn "giving in marriage" (*engyē*) of the bride by her father or closest male relative acting as her *kyrios,* to be followed by her actual "delivery" (*ekdosis*) to the groom. Monogamy was the norm, although all *poleis* did not go to the same extreme as democratic Athens, which peremptorily excluded from the *oikos* the children of concubines (though possibly allowing a citizen to maintain more than one *oikos*).

In Athens, and other states where the status of women was similarly reduced, the husband became the *kyrios* of his wife and her property, including her dowry (*proix*), the latter consisting of a share in the estate of the bride's *kyrios,* customarily, though not in fulfillment of a legal claim, conveyed by him to the groom. If the marriage was terminated by the death of the wife, title to the *proix* passed to her sons, although the widower might keep it until he died. In all other cases, the husband or his heir was liable for the return of the dowry to the wife's family (*dikē proikos*).

Divorce was permitted to either spouse and might be effected simply by expelling or abandoning the other, no justifying reason being required. A divorced woman returned to her own family, the head of which became again her *kyrios.*

Guardianship. The estate of a fatherless minor (in Athens to 18 years of age) was administered by a guardian (*epitropos*) who was his closest agnate (*i.e.*, relation through male descent), unless the father had appointed a guardian *inter vivos* or by will. The *epitropos* became the temporary *kyrios* of the ward's estate and had to support him. In Athens his management was supervised by the *archōn* (chief magistrate), and every citizen might file a complaint against a negligent guardian. His office having expired, the guardian was liable to the ward for the return of the estate (*dikē epitropēs*). In view of a widespread confusion of concepts, it is necessary to stress that at Athens, though less so in Hellenistic Egypt, there was a marked difference in function and terminology separating guardianship from the domestic power (*kyrieia*) over women.

Succession. The sacral and political importance of the *oikos* and the resulting necessity of securing its perpetuation, as far as possible, was also decisive for the law of succession. In Athens legitimate sons, both natural and adopted, excluded all other persons from intestate (where there is no will) and testate succession. The evidence is vague as to what happened when an Athenian whose father was still alive died intestate and without leaving children, but what is known suggests that the father was entitled to seize the son's property as belonging to the family's *klēros* ("estate"). If the father, too, was dead, the title of *klēronomos* ("possessor of the estate") would pass to the *anchisteis* (agnates) in a legally prescribed order. Women could not inherit under Athenian law. A brotherless daughter (*epiklēros*), however, though not herself entitled to her father's *klēros,* transmitted the right of succession to her sons. The danger of the patrimony thereby passing into the hands of outsiders was averted by the right of the closest *anchisteus* to claim her as his wife, together with the *klēros;* the fact that she might be married to another man was apparently notwithstanding. Analogous regulations existed in Gortyn, Sparta, and probably other places.

Testamentary succession played no large part in Classical Greece. Some *poleis,* such as Gortyn, did not recognize it at all, and everywhere it was excluded by the presence of legitimate sons. An Athenian law attributed to Solon allowed a citizen who had no son to appoint an heir by executing an adoption to take effect after his death. An early custom of dividing an estate among several future heirs widened in Classical times into the practice of drawing up a written will (*diathēkē*) ordering individual bequests to various persons.

Property. Individual ownership was known in the Greek world from at least epic times. Important exceptions in the Classical period were the control over the helots of Sparta, which was vested in the state as a whole, and, apparently, in some city-states other than Athens, certain forms of collective ownership of land and other real property, such as the inalienable ancestral city house of the Gortynian family.

Title to property

Regarding the theoretical construction and the judicial protection of property rights, Greek law remained primitive. Title signified no more than a general relation to the object. Its extent and quality varied in accordance with the purpose for which the relation had been established, ranging from full and permanent ownership through control of a ward's estate or a wife's dowry, and through mortgage to temporary tenancy. Within the limits set by the purpose, title involved the rights to hold or establish actual dominion (*kratēsis*) and to sell or otherwise dispose of the object (*kyrieia*). The holder of a title might under certain circumstances forcibly seize its object. Beyond this he was given only indirect protection, inasmuch as unwarranted resistance to his seizure or wrongful violation of his position gave rise to tort actions.

Modes of original acquisition of title were succession, the gathering of crops, and the taking of war booty. Derivative acquisition required a valid transfer by the rightful holder of the title, who within his own power of disposal might determine the quality of the title to be conveyed. The validity of a title acquired by purchase was, in addition, contingent on the payment of the price ("principle of surrogation"), while, contrary to Roman law, conveyance of the actual possession was neither sufficient nor requisite. A distinctive characteristic of Greek legal methods as compared with those of other ancient nations lay in the fact that in many *poleis* the full effect of a transfer of title to real property depended on varying forms of publicity, such as previous announcement by heralds, a formal statement of the transfer in the presence of private or official witnesses, or a recording of the transaction in a public register, the contents of which were sometimes kept permanently available by inscribing them on stone.

Obligation. *Contract.* As might be expected, contracts played an important part in the economic life of the classical *poleis*, many of which engaged in extensive commerce. Contractual liability, however, resulted only indirectly. Contrary to widely accepted views, recent research has established that Greek law did not attribute a binding effect to agreements entered upon by mere consent. Every contract (*synthēkē*) needed a "real" foundation consisting in a consideration for the loss of which the debtor could be held responsible if by failing to perform his obligation he frustrated the creditor's purpose. Because of this structure, the contractual obligation was necessarily one-sided, and the creditor had no claim to specific performance but only a right to resort to a *praxis* (enforcement proceeding) that the debtor could not avert, unless he paid a ransom usually amounting to double the damages. Historically, this system seems to have evolved from the delict (or tort), which the Greeks called *blabē*. Since, however, parties were at liberty to lay down the prerequisites of liability according to their own wishes and to fix the amount of the damage beforehand, in the epoch of the orators liability for nonfulfillment of a contractual obligation had ceased to be considered as one arising from tort. One of the important factors in this change was the principle that whatever a man voluntarily acknowledged ("homologated") should be *kyrion; i.e.*, "irrefutable" (literally "determining"). In Athens, this was confirmed by express legislation.

The development of a highly diversified stock of standard contracts was probably an achievement of professional draftsmen of the Hellenistic period, but the foundations were laid in Classical times. One of the oldest transactions was *misthōsis* (lease), its earliest and always most important object being agricultural land. Another method known already to the Archaic period, and possibly modelled after a Semitic pattern that may have been Phoenician, was to submit to liability through the acceptance of an *arrha*. Its chief area of application was sale. The *arrha* was a down payment, a multiple of which the seller had to refund as a penalty if he failed to fulfill his part of the bargain. It functioned as an indirect means of obtaining delivery that the buyer could not enforce directly, since sale was held to be strictly a cash transaction that created no obligation apart from a warranty (*bebaiōsis*) of the title conveyed. Basically, *arrha* was, of course, one form of the general contract of credit. Following the simple pattern of the loan (*daneion*) and bearing interest if agreed upon, this was the predominant type of transaction; on account of its flexibility it was, in fact, excellently suited to the needs of business. It was used to finance the extended sea trade in the particular form of the maritime loan, *i.e.*, a sum of money secured by ship or cargo or both and to be repaid on condition of safe completion of the voyage, thus also serving in a primitive fashion the economic function of insurance.

Greek law did not tie the validity of a contract to any particular form. An oral understanding arrived at in the presence of witnesses was sufficient. From the 4th century on, however, written agreements were preferred, and maritime loans were always embodied in a document (*syngraphē*).

Security. Security was known under Greek law in both personal and real forms. In the personal form, the surety (*engyos*) vouched for a defendant's appearance in court or for the fulfillment of a debtor's promise and could be held liable for the same penalty as the latter. The prevalent type of real security in Classical times was the lien (*hypothēkē*), usually drawn up as an "equity of redemption" (*prasis epi lysei*). The debtor might for the time being remain in possession of the pledge, but the creditor had a right to seize it as a forfeit after default. It was, however, possible to mortgage the same object to several creditors, each of whom would be entitled to lay hands on it but who had to satisfy in full any preceding creditors. Agreements permitting creditors to seek satisfaction by selling the lien did not come into use before the Hellenistic period.

Torts. Under Athenian law the most important private tort was *blabē*, the causing of damage, not necessarily confined to physical destruction, to another person's property. The penalty for willful infliction was twice the amount of the detriment. Other torts, basically related to but technically distinguished from *blabē*, were illegal breaking into another's domain by the use of force (*bia*) and hindrance of a rightful seizure (*exoulē*, literally "ejectment"). *Hybris* (overbearing conduct), in addition to being a public crime, gave rise in Athens to two private tort actions: one, *dikē aikeias*, could be brought for physical injury, the other, *dikē kakēgorias*, for some specified types of libel. The penalty was a fine to be determined individually from case to case. Even murder, punishable by death or exile, was basically a private tort —a relic of an archaic custom of taking private revenge —inasmuch as the prosecution of the murderer was restricted to and incumbent on the victim's relatives.

CRIMINAL LAW

Offenses under Greek law

The strongly collectivistic spirit of the classical *polis* explains the character of its public criminal law. In the democratic period of Athens, criminal conduct was a common concern of all the citizens, each of whom was entitled to bring a prosecution. Punishment was confined to conduct classified as criminal by positive legislation, but the list of offenses was long. It ranged from deeds immediately detrimental to the state and its religion, such as treason, cowardice, posing as a citizen, or acts of impiety (*asebeia*), through those jeopardizing the democratic order (*e.g.*, introducing legislation conflicting with the existing legal order) to acts disturbing the domestic peace, such as theft. Even acts revealing an uncivil attitude, such as *hybris* or undue treatment (*kakōsis*) of parents, women, or orphans, were classified as criminal. The penalties exacted were death, exile, sale into slavery, and fines.

COURTS AND PROCEDURES

At the present stage of research, the only system sufficiently known to warrant description is that of 4th-century Athens. It is important to note that its structure was peculiar to Athens, although the basic concept of *dikē* (originally, an act of seizure) was common to all the Greeks.

The Athenian judiciary. In the democratic period justice was administered by magistrates, popular courts (*dikastēria*), and the Areopagus. The functionaries received the actions and arranged the trials that took place before the courts, with each functionary having a specific jurisdiction: the archon over matters pertaining to family and succession, the "king" (*archōn basileus*) over religious matters (including murder), the *thesmothetai* ("determiners of customs") and others over the rest. A special jurisdiction was that of the *polemarchos* (literally, "general") over the metics. The trial competence of the dicasteries rested on the principle, first introduced within certain limits by Solon and made universal after the establishment of full democracy, that the citizenry in its totality should judge the affairs of its members. The dicasts were selected by lot, every citizen over 30 years old being eligible. In rare cases of great political importance, the whole *hēliaia* (*i.e.*, the popular assembly organized as a court of 6,001 men) was convened. Normally sections of the *hēliaia* (specifically called *dikastēria*), composed of 1,501, 1,001, or 501 men in criminal cases and 201 men in civil cases, were charged with the decision.

Murder cases were argued before the Areopagus, a body composed of former archons. Probably transformed from an original council of the nobility, it was a relic of the predemocratic period.

Procedure. In the Greek view, the trial served to determine the justification of a claim to seize the defendant's person or belongings or both by way of an enforcement proceeding (*praxis*). The claim (*dikē*) might be raised by the plaintiff in pursuance of a private right or as a "public" (*dēmosia*) *dikē* for the purpose of obtaining the defendant's punishment. The filing of a public *dikē* (technically called a *graphē*) was open to every citizen. Apart from this, the differences between private and criminal procedures were slight.

Both private *dikai* and *graphai* had to be initiated by summoning the defendant (who might be under arrest) to the magistrate having jurisdiction in the matter and by filing a written complaint with the latter, who would subject it to a preliminary examination (*anakrisis*). Parties to a civil suit concerning pecuniary affairs were then sent to a public arbitrator (*diaitētēs*). If one of them refused to accept the award or if the matter was not subject to compulsory arbitration, the case was referred to a discastery presided over by the magistrate. The dicasts, after listening to the arguments and evidence submitted by the parties, found their decision, which could only be a choice between the two proposals made by the parties, by secret ballot without debate. Their judgment was final between the parties, but the loser might bring a private tort action (*dikē pseudomartyriōn*) against a witness whose false deposition had influenced the verdict. A victorious plaintiff in a private lawsuit had to enforce the judgment himself by attaching property of the defendant.

HISTORICAL IMPORTANCE OF THE GREEK LAW

In distinct contrast with the Greek philosophy of justice, the positive law of ancient Greece had little influence on later developments. Its concepts and methods did, of course, widely determine the legislation and practice of Hellenistic monarchies, and a few institutions of Greek origin, such as the "Rhodian" maritime law of jettison or certain methods of documentation (mostly Hellenistic, to be sure), were adopted by the Romans. Contrary to views held some decades ago, however, the late Roman law, and with it west European legal doctrine, did not undergo any notable degree of Hellenization. Only in the customs of isolated places in Greece itself do some ancient traditions seem to survive; their extent is still a problem for legal historians.

BIBLIOGRAPHY. P. VINOGRADOFF, *The Jurisprudence of the Greek City*, vol. 2 of his *Outlines of Historical Jurisprudence* (1922), is largely obsolete but still deserves mention among the books written on Greek law in the English language. Comprehensive presentations of more recent date include J.W. JONES, *The Law and Legal Theory of the Greeks* (1956); and A.R.W. HARRISON, *The Law of Athens*, vol. 1, *The Family and Property* (1968). An extensive discussion of procedural matters, unfortunately not fully satisfactory from a legal point of view, is found in the philological work by R.J. BONNER and G. SMITH, *The Administration of Justice from Homer to Aristotle*, 2 vol. (1930–38, reprinted 1968). A small but also valuable work is G.M. CALHOUN, *Introduction to Greek Legal Science* (1944), dealing with the intellectual background of Greek law. Detailed bibliographical data may be found in all these works.

(H.J.W.)

Greek Mythology

The body of stories concerning the gods, heroes, and rituals of the ancient Greeks is the subject matter of Greek mythology. That myths contained a considerable element of fiction the more critical Greeks, such as the philosopher Plato in the 5th–4th centuries BC, recognized; in general, however, the myths were viewed in the popular piety of the Greeks as true accounts. Greek mythology has subsequently had extensive influence on the arts and literature of Western civilization, which fell heir to much of Greek culture.

Nature and significance. Although men of all countries, eras, and stages of civilization have developed myths that explain the existence and workings of natural phenomena, recount the deeds of gods or heroes, or seek to justify social or political institutions, the myths of the Greeks have remained unrivalled in the Western world as sources of imaginative and appealing ideas. Poets and artists from ancient times to the present have derived inspiration from Greek mythology and have discovered contemporary significance and relevance in classical mythological themes.

Importance of Greek mythology in the Western world

Myths helped to illuminate and make Greek religion intelligible to worshippers by providing a wealth of religious background detail conceived in simple and picturesque terms. The Romans identified their own functionally perceived deities with their more fully anthropomorphized Greek counterparts, which resulted in a wholesale, though mainly literary, adoption of Greek mythology.

During the Renaissance, the view of myth was reinterpreted in terms of allegory (a symbolic story form), and, in the course of its reinterpretation, Greek mythology enriched literature from the Renaissance period onward with classical allusions. In the modern world, not only art and literature have benefitted from Greek mythology but scientists also have retained the ancient practice of naming new stars and planets (and later, even rockets) after Greek gods and heroes. Indeed, the names of Greek mythological figures have survived in innumerable contexts—from advertising to the cinema—in modern times.

The conquests of Alexander the Great in the 4th century BC and, later, those of the Romans helped to spread Greek mythology throughout the ancient Mediterranean world, profoundly influencing Mediterranean culture. Even the myths of the Egyptians and Babylonians were retold in typical Greek style. The Roman poets, particularly Virgil and Ovid, were most influential in bequeathing Greek mythology to Western posterity. Without the rich heritage of Greek mythology, the literature and art of the Renaissance and modern world would have been infinitely poorer. English and German literature and, above all, French are especially indebted to Greek myth. Psychoanalysts such as Sigmund Freud as well as dramatists (*e.g.*, Jean-Paul Sartre in France and Eugene O'Neill in the United States) have reinterpreted classical myths in the light of modern contemporary experience. Greek mythology has also provided artists with an unparalleled variety of exciting motifs.

Sources of myths: literary and archaeological. *The Homeric poems: the Iliad and the Odyssey.* Homer, who lived probably in the 9th or 8th century BC, is the oldest known literary source of Greek myth and legend: both of his great epics, the *Iliad* and the *Odyssey*, belong to a tradition reaching back into Mycenaean times (*c.* 1400–1100 BC). In them, Homer collected the many myths of the gods, arranged them in a coherent whole, assembled

the gods on Mt. Olympus, and related in detail their activities, loves, and hates. Unlike Hesiod, another great poet, who flourished *c.* 800 BC, Homer was not primarily concerned with cosmogony (origin of the world), and he thus refers only incidentally to Oceanus and his wife Tethys as the authors of creation and to the Titans (offspring of Uranus and Gaea) as imprisoned within the ground. The aim of an epic poet was to entertain; thus, much of what Homer says about the gods is not intended to be taken seriously, and even more about them he assumed. The myth that the god Apollo spread plague by discharging his arrows is a type of religious myth. But others, like the goddess Hera's punishments, are more akin to folktales. Homer remains a primary source of legends connected with the city of Troy in Asia Minor, as well as of older sagas, such as those of the heroes Heracles (in Roman mythology Hercules), Meleager, and Jason.

The works of Hesiod: Theogony and Works and Days. The fullest and most important source of myths about the origin of the gods is the *Theogony* ("Divine Genealogy") by Homer's near contemporary, the Boeotian poet Hesiod. The grim tale of mutilation, revolt, and struggles against older gods and monsters bears some relation to Near Eastern cosmogonical myths. Much else is a bare catalog of matings and births of gods, rivers, planets, winds, and more abstract phenomena. The story of Prometheus' tricking of Zeus, the leading god of Olympus, in giving him bones rather than meat is an etiological (explanatory) myth pertaining to the customs of Greek sacrifice. On the other hand, the myth of Pandora (the first woman, who opened an urn, thus releasing various evils upon mankind) is clearly based on folktale. Both myths reappear in the *Works and Days* (a farmer's calendar, in epic form, of somewhat later date) as well as the myth of the Four Ages of the world, further genealogies, and other myths.

Cosmogonic and etiological interests

Other literary works. Fragmentary post-Homeric epics, of varying date and authorship, filled the gaps in the accounts of the Trojan War recorded in the *Iliad* and *Odyssey;* the so-called Homeric Hymns (shorter surviving poems) are the source of several important religious myths. Many of the lyric poets preserved various myths, but the odes of Pindar of Thebes (flourished 6th–5th centuries BC) are particularly rich in myth and legend. The works of the three tragedians—Aeschylus, Sophocles, and Euripides, all of the 5th century BC—are remarkable for the variety of the traditions they preserve. In Hellenistic times (323–30 BC) Callimachus, a 3rd-century-BC poet and scholar in Alexandria, recorded many obscure myths; and Euhemerus, a novelist of the 3rd century BC, suggested that the gods were originally human, thus giving his name to this view, known as Euhemerism. Apollonius of Rhodes, another scholar of the 3rd century BC, preserved the fullest account of the Argonauts in search of the Golden Fleece. In the period of the Roman Empire, the *Library* of the pseudo-Apollodorus (attributed to a 2nd-century-AD scholar), the antiquarian writings of the Greek biographer Plutarch, and the works of Pausanias, a 2nd-century-AD geographer, as well as the *Genealogies* of Hyginus, a 2nd-century-AD mythographer, have provided valuable sources in Latin of later Greek mythology.

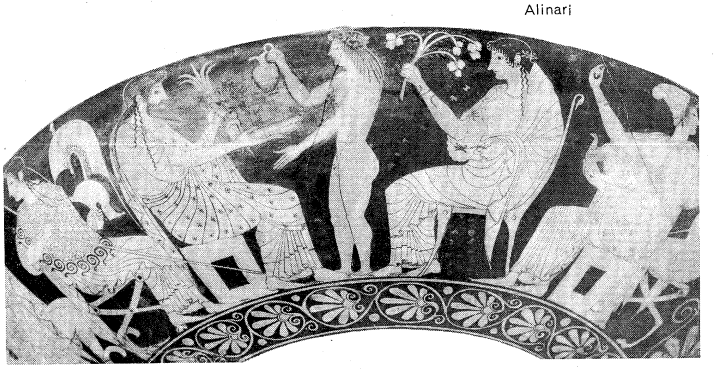
Archaeological discoveries. The discovery of the Mycenaean civilization by Heinrich Schliemann, a 19th-century German amateur archaeologist, and the discovery of the Minoan civilization in Crete (from which the Mycenaean ultimately derived) by Sir Arthur Evans, a 20th-century English archaeologist, helped to explain many of the questions raised by scholars about Homer's epics and provided archaeological proofs of many of the mythological details about gods and heroes. Unfortunately, the evidence about myth and ritual at Mycenaean and Minoan sites is entirely monumental, because the Linear B script (an ancient form of Greek found in both Crete and Greece) was mainly used to record inventories, though the names of gods and heroes have been doubtfully revealed.

Significance of Mycenaean and Minoan archaeological discoveries

Geometric designs on pottery of the 8th century BC depict scenes from the Trojan cycle, as well as the adventures of Heracles. The extreme formality of the style, however, renders much of the identification difficult, and there is no inscriptional evidence accompanying the designs to assist scholars in identification and interpretation. In the succeeding Archaic (*c.* 750–*c.* 500 BC), Classical (*c.* 480–323 BC), and Hellenistic periods, Homeric and various other mythological scenes appear to supplement the existing literary evidence.

Forms of myth in Greek culture. *Religious myths.* Greek religious myths are concerned with gods or heroes in their more serious aspects or are connected with ritual. They include cosmogonical tales of the genesis of the gods and the world out of Chaos, the successions of divine rulers, and the internecine struggles that culminated in the supremacy of Zeus, the ruling god of Olympus. They also include the long tale of Zeus's amours with goddesses and mortal women, which usually resulted in the births of younger deities and heroes. The goddess Athena's unique status is implicit in the story of her motherless birth (she was born directly from Zeus); and the myths of Apollo explain that god's sacral associations, describe his remarkable victories over monsters and giants, and stress his jealousy and the dangers inherent in immortal alliances. Myths of Dionysus (a wine god whose cult involved orgiastic rites), on the other hand, demonstrate the hostility aroused by a novel faith. Some myths are closely associated with rituals, such as the account of the drowning of the infant Zeus's cries by the Curetes, attendants of Zeus, clashing their weapons, or Hera's (the queen goddess of Olympus) annual restoration of her virginity by bathing in the spring Canathus. Some myths about heroes and heroines also had a religious basis. The tale of man's creation and moral decline forms part of the myth of the Four Ages (see below). His subsequent destruction by flood and regeneration from stones is partly based on folktale.

Alinari



The gods on Olympus: Athena, Zeus, Dionysus, Hera, and Aphrodite; detail of a painting on a Greek cup. In the Museo Municipale, Tarquinia, Italy.

Legends. Myths were viewed as embodying divine or timeless truths, whereas legends (or sagas) were quasi-historical. Hence, famous events in epics, such as the Trojan War, were generally regarded as having really happened, and heroes and heroines were believed to have actually lived. Earlier sagas, such as the voyage of the Argonauts, were accepted in a similar fashion. Most Greek legends were embellished with folktales and fiction, but some certainly contain a historical substratum. Such are the tales of more than one sack of Troy, which are supported by archaeological evidence, and the labours of Heracles, which suggest Mycenaean feudalism. Again, the legend of the Minotaur (a being part human, part bull) could have arisen from exaggerated accounts of bull leaping in ancient Crete.

Themes of legends and folktales

In another class of legends, heinous offenses, such as attempting to make love to a goddess against her will, deceiving the gods grossly by inculpating them in crime, or assuming their prerogatives, were punished by everlasting torture in the underworld. The consequences of social crimes, such as murder or incest, were also described in legend (*e.g.*, the story of Oedipus, who killed his father



Heracles fighting with the Amazons, detail from a volute crater attributed to Euphronius, c. 500 BC. In the Museo Archeologico, Arezzo, Italy.
Alinari

and married his mother). Legends were also sometimes employed to justify existing political systems or to bolster territorial claims.

Folktales. Folktales, consisting of popular recurring themes and told for amusement, inevitably found their way into Greek myth. Such is the theme of lost persons —whether husband, wife, or child (*e.g.*, Odysseus, Helen of Troy, Paris of Troy)—found or recovered after long and exciting adventures. Journeys to the land of the dead were made by Orpheus (a hero who went to Hades to restore his dead wife Eurydice to the realm of the living), Heracles, Odysseus, and Theseus (the slayer of the Minotaur). The victory of the little man by means of cunning against impossible odds, the exploits of the superman (*e.g.*, Heracles), or the long-delayed victory over enemies are still as popular with modern writers as they were with the Greeks. The successful countering of the machinations of cruel sires and stepmothers, rescues of princesses (who are often witches) from monsters, or temporary forgetfulness at a crucial moment are also familiar themes in Greek myth. Recognition by tokens, such as Odysseus' scar or peculiarities of dress, is another common folktale motif. The babes-in-the-wood theme of the exposure of children and their subsequent recovery (though usually as the result of a dream or an oracle) is also found in Greek myth. Wishful thinking was responsible for the theme of victory over death, the related visit to the land of ghosts, and, of course, the happy ending, as noted, for example, in the Argonaut saga.

Types of myths in Greek culture. *Myths of origin.* Myths of origin represent an attempt to render the universe comprehensible in human terms. Greek creation myths (cosmogonies) and views of the universe (cosmologies) were more systematic and specific than those of other ancient peoples. Yet their very artistry serves as an impediment to interpretation because true myth is usually embellished with folktale and fiction told for its own sake. Thus, though the aim of Hesiod's *Theogony* is to describe the ascendancy of Zeus (and, incidentally, the rise of the other gods), the inclusion of such familiar themes as the hostility between the generations, the enigma of woman (Pandora), the exploits of the friendly trickster (Prometheus), or exciting struggles against powerful beings or monsters like the Titans (and, in later tradition, the Giants) merely enhances the interest of an epic account.

According to Hesiod, four primary divine beings first came into existence: Chaos (Space), Gaea (Earth), Tartarus (the Abyss), and Eros (Love). The creative process began with the forcible separation of Gaea from her doting consort Uranus (Heaven) in order to allow her progeny to be born. The means of separation employed, viz., the cutting off of Uranus' genitals by his son Cronus, bears a certain resemblance to a similar story recorded in Babylonian epic. The crudity is relieved, however, in characteristic Greek fashion by the friendly collaboration of Uranus and Gaea, after their divorce, in a plan to save Zeus from the same Cronus, his cannibalistic sire.

According to Greek cosmological concepts, the Earth was viewed as a flat disk afloat on the river of Ocean. The Sun (Helios) traversed the heavens like a charioteer and sailed around the Earth in a golden bowl at night. Natural fissures were popularly regarded as entrances to the subterranean house of Hades, home of the dead.

Myths of the ages of the world. Egyptians and Sumerians knew of a paradise free from death and suffering reserved for the gods. Hesiod, on the other hand, in the *Works and Days*, came near to propounding a philosophy of history in explaining man's decline by the myth of the Four Ages. The gods created the men of the Golden Age during the reign of Cronus. They never grew old, were free from toil, and passed their time in jollity and feasting. When they died, they became guardian spirits on Earth.

The decline of man

Why the Golden Age came to an end Hesiod failed to explain, but it was succeeded by the Silver Age. After an inordinately prolonged childhood, the men of the Silver Age began to act presumptuously and neglected the gods. Consequently, Zeus hid them in the Earth, where they became spirits among the dead.

Zeus next created the men of the Bronze Age, men of violence who perished by mutual destruction. At this point the poet somewhat illogically intercalates the Race of Heroes, of whom the more favoured (who were related to the gods) reverted to a kind of restored Golden Age existence under the rule of Cronus (forced into honorable exile by his son Zeus) in the Isles of the Blessed.

The final age, the antithesis of the Golden Age, was the Iron Age, during which the poet himself had the misfortune to live. But even that was not the worst, for he believed that a time would come when infants would be born old, and there would be no recourse left against the universal moral decline. The presence of evil was explained by Pandora's rash action in opening the fatal urn.

Myths of the gods. Myths about the gods described their births, victories over monsters or rivals, love affairs, special powers, or connections with a cultic site or ritual. As these powers tended to be wide, the myths of many gods were correspondingly complex. Thus, the Homeric Hymns to Demeter, a goddess of agriculture, and to the Delian and Pythian Apollo describe how these deities came to be associated with sites at Eleusis, Delos, and Delphi, respectively. Similarly, myths about Athena, the patroness of Athens, tend to emphasize the goddess' love

of war and her affection for heroes and the city of Athens; and those concerning Hermes (the messenger of the gods), Aphrodite (goddess of love), or Dionysus describe Hermes' proclivities as a god of thieves, Aphrodite's lovemaking, and Dionysus' association with wine, frenzy, miracles, and even ritual death. Poseidon (god of the sea) was unusually atavistic, in that his union with Earth and his equine adventures appear to hark back to his pre-marine status as a horse or earthquake god. Many myths, on the other hand, are trivial and lighthearted. Typical of such are the amusing descriptions of conjugal friction between Zeus and Hera in the *Iliad*, which, though conceivably reflecting religious hostility between Zeus worshippers and Hera worshippers in prehistoric times, were mainly the products of fiction and folktale. Similarly, the beauty contest described in the famous Judgment of Paris (son of the king of Troy) appears to be little more than a fairytale in essence.

Combinations of myths, legends, and folktales

As time went on, an accretion of minor myths continued to supplement the older and more authentic ones. Thus, the loves of Apollo, virtually ignored by Homer and Hesiod, explained why the bay (or laurel) became Apollo's sacred tree and how he came to father Asclepius, a healing god. Similarly, the presence of the cuckoo on Hera's sceptre at Hermione or the invention of the panpipe were explained by fables. Such etiological myths proliferated during the Hellenistic era, though in the earlier periods genuine examples are harder to detect.

Of folk deities, the Nymphs (nature goddesses) personified nature or the life in water or trees and were said to punish unfaithful lovers. Water nymphs (Naiads) were reputed to drown those with whom they fell in love, such as Hylas, a companion of Heracles. Even the gentle Muses (goddesses of the arts and sciences) blinded their human rivals, such as the bard Thamyris. Satyrs (youthful folk deities with bestial features) and Silens (old and drunken folk deities) were the Nymphs' male counterparts. Like sea deities, Silens possessed secret knowledge that they would reveal only under duress. Charon, the grisly ferryman of the dead, was also a popular figure of folktale.

Myths of heroes. Hero myths included elements from tradition, folktale, and fiction. The saga of the Argonauts, for example, is highly complex and includes elements from folktale and fiction, but the information that the fleet mustered at Colchis may be regarded as genuine legend. Episodes in the Trojan cycle, such as the departure of the Greek fleet from Aulis or Theseus' Cretan expedition and death on Scyros, may belong to traditions dating from the Minoan–Mycenaean world. On the other hand, events described in the *Iliad* probably owe far more to Homer's creative ability than to genuine tradition. Even heroes like Achilles, Hector, or Diomedes are largely fictional, though doubtlessly based on legendary prototypes. The *Odyssey* is the prime example of the wholesale importation of folktales into epic. All the best known Greek hero myths, such as the labours of Heracles and the adventures of Perseus, Cadmus, Pelops, or Oedipus, depend more for their interest on folktales than legend. Certain heroes—Heracles, the Dioscuri (the twins Castor and Pollux), Amphiaraus (one of the Argonauts), or Hyacinthus (a youth loved by Apollo and accidentally killed)—may be regarded as partly legend and partly religious myth. Thus, whereas Heracles, a man of Tiryns, may originally have been a historical character, the myth of his demise on Oeta and subsequent elevation to full divinity is closely linked with a cult. In time, Heracles' popularity was responsible for connecting his story with the Argonauts, an earlier attack on Troy, and with Theban myth. Similarly, the exploits of the Dioscuri are those of typical heroes: fighting, carrying off women, and cattle rustling. After their death they passed six months alternately beneath the Earth and in the world above, which suggests that their worship, like that of Persephone (the daughter of Zeus and Demeter), was connected with fertility or seasonal change.

Myths of seasonal renewal. Certain myths, in which goddesses or heroes were temporarily incarcerated in the underworld, were allegories of seasonal renewal. Perhaps the best known myth of this type is the one telling how Hades (Latin Pluto), the god of the underworld, carried Persephone off to be his consort, causing her mother Demeter, the corn goddess, to allow the earth to grow barren out of grief. Because of her mother's grief, Zeus permitted Persephone to spend six months above and six months below ground. The meaning is patent. During the torrid Greek summer the earth lies bare and unproductive. When rain falls in autumn and the fields are ploughed, the land begins to bloom once more. This simple myth seems to have inspired the Eleusinian Mysteries (an agricultural and salvatory religion), whose annual celebration helped men to find a compromise with death.

The mysterious myth of the concealment of the infant Zeus in Crete has, as already mentioned, ritual associations. In other words, the child, born of Rhea, one form of the earth mother, was in origin a fertility spirit. His counterparts were Adonis and Attis. Adonis, son of Myrrha (the balsam tree), was loved by Aphrodite and killed by a boar. Zeus therefore arranged that Adonis should spend a third of his life above ground, a third below, and a third where he wished. Hence, female devotees ritually lamented the season of his passing. Attis similarly was born of the almond tree and the beloved of Cybele. When Attis died as a result of a frenzy in which he emasculated himself, Cybele persuaded Zeus to preserve his body and keep one finger (a phallic symbol) alive.

A well-known Attic myth describes how Cecrops' daughters, significantly styled Agraulos (Shining), Herse (Dew), and Pandrosos (All Dew), were given the snake monster Erichthonius, the product of the fire god Hephaestus' attempt to rape her, to keep in a chest for Athena. Snakes are phallic symbols and are associated with the earth, thus suggesting a fertility motif.

Myths involving theriolatry. Many Greek myths involve animal transformations, though there is no proof that theriolatry (animal worship) was ever practiced by the Greeks. Gods sometimes assumed the form of beasts in order to deceive goddesses or women. Zeus, for example, assumed the form of a bull when he carried off Europa, a Phoenician princess, and appeared in the guise of a swan in order to attract Leda, wife of a king of Sparta; Hera, in another incident, transformed Io, a maiden loved by Zeus, into a heifer. Poseidon took the shape of a stallion to beget the wonder horses Arion and Pegasus, as did Boreas (the North Wind) to mate with mares.

Deities appearing in nonhuman form

Though such animals are involved in significant mythical events, they are not necessarily the objects of theriolatry, as noted above. They usually served other purposes in the story, such as being agents of change, symbols of strength and passion, objects to test the courage of heroes, agents of escape, or forms of punishment. Snakes licked the ears of the seer Melampus, thus enabling him to understand the language of birds. Tiresias, a Theban seer, changed his sex after killing a female serpent in the act of mating and was thus able to settle the argument that had arisen between Zeus and Hera as to whether man or woman gained more pleasure from love. Bulls represented unbridled strength and passion and were not especially associated with any particular deity. Poseidon was responsible for the bull that captivated Pasiphae, wife of King Minos of Crete, and begot the Minotaur (half human, half bull). The same beast was reputed to have been brought by Heracles from Crete and to have been captured by Theseus, who sacrificed it to Apollo at Marathon. Mythical land or sea monsters, as well as lions and boars, served to test the heroes' courage. Notable were the famous Calydonian boar hunt, in which many heroes took part, and Heracles' capture of the Erymanthian beast. Other animal transformations are also known. Leto, the mother of Apollo and Artemis, assumed the form of a wolf in order to escape Hera. Lycaon, king of Arcadia, was changed into a wolf by Zeus for practicing human sacrifice. The huntsman Actaeon was changed into a stag by Artemis for surprising her in the act of bathing.

Other types. Other types of myth exemplified the belief that the gods sometimes appeared on Earth disguised

Deities appearing in human form

as men and women and rewarded any help or hospitality offered them. Baucis, an old Phrygian woman, and Philemon, her husband, for example, were saved from the flood by offering hospitality to Zeus and Hermes, both of whom were in human form. The punishment of men's presumption in claiming to be the gods' superiors, whether in musical skill or even the number of their children, is described in several myths. The gods' jealousy of their musical talents appears in the beating and flaying of the flute-playing Satyr, Marsyas, by Athena and Apollo, as well as in the attaching of ass's ears to King Midas for failing to appreciate the superiority of Apollo's music to that of the god Pan. Jealousy was the motive for the slaying of Niobe's many children, because of Niobe's flaunting her fecundity to the goddess Leto, who had only two offspring. Similar to such stories are the moral tales about the fate of Icarus, who flew too high on homemade wings, or the myth about Phaethon, the son of Helios, who failed to perform a task too great for him (controlling the horses of the Sun).

Transformation into flowers or trees, whether to escape a god's embraces (such as Daphne, a nymph transformed into a laurel tree), as the result of an accident (such as Hyacinthus, a friend of Apollo, who was changed into a flower), or because of pride (*e.g.*, the beautiful youth Narcissus who fell in love with his own reflection and was changed into a flower), were familiar themes in Greek myth.

Also popular were myths of fairylands, such as the Garden of the Hesperides (in the far west) or the land of the Hyperboreans (in the far north), or encounters with monstrous or outlandish people, such as the Centaurs or Amazons.

Greek mythological characters and motifs in art and literature. The earliest visual representations of mythological characters and motifs occur in late Mycenaean and sub-Mycenaean art. Though identification is controversial, Centaurs, a Siren, and even Zeus's lover Europa have been recognized. Mythological and epic themes are also found in Geometric art of the 8th century BC, but not until the 7th century did such themes become popular both in ceramic and sculptured works. During the Classical and subsequent periods, they became commonplace. The birth of Athena was the subject of the east pediment of the Parthenon in Athens, and the legend of Pelops and the labours of Heracles was the subject of the corresponding pediment and the metopes (a space on a Doric frieze) of the Temple of Zeus at Olympia. The battles of gods with Giants and of Lapiths (a wild race in northern Greece) with Centaurs were also favourite motifs. Pompeian frescoes reveal realistic representations of Theseus and Ariadne, Perseus, the fall of Icarus, and the death of Pyramus.

Resurgence of Greek mythological motifs

The great Renaissance masters added a new dimension to Greek mythology. Among the best known subjects of Italian artists are Botticelli's "Birth of Venus," the Ledas of Leonardo da Vinci and Michelangelo, and Raphael's "Galatea."

Greek mythology formed the staple of most Greek poetry and epic, as well as of dramatic works. It also influenced the thoughts of philosophers and historians to a marked degree. Homer mingled myth with legend, and Hesiod attempted to give a consistent account of mythology in general or employed it for didactic or social ends. Pindar, unlike the other lyric poets, took an independent line and toned down the more unsavoury features. The tragedians manipulated myths in the interests of drama or to suit their inner vision, and Aristophanes exploited mythological parody. During the Hellenistic Age, poets such as Apollonius, Callimachus, and others in the *Palatine Anthology* passed on a wealth of Greek myth to the Romans.

Through the medium of Latin and, above all, the works of Ovid, Greek myth influenced medieval poets such as Petrarch and Boccaccio in Italy and Chaucer in England; Dante in Italy during the Renaissance; and, later, the English Elizabethans and John Milton. Racine in France and Goethe in Germany revived Greek drama, and nearly all the major English poets from Shakespeare to Robert Bridges turned for inspiration to Greek mythology. More recently, classical themes have been reinterpreted by such major dramatists as Jean Anouilh, Jean Cocteau, and Jean Giraudoux in France, Eugene O'Neill in America, and T.S. Eliot in England and by great novelists such as James Joyce (Irish) and André Gide (French). The German composer Christoph Gluck (18th century), the German-French composer Jacques Offenbach (19th century), and many others have set Greek mythological themes to music.

Conclusion. The importance of Greek mythology in the intellectual, artistic, and emotional history of Western man can hardly be overestimated. To deny its value and significance would be tantamount to denying the achievements of Western culture itself and the genius of the great writers and artists who have been inspired by it. The origins of myth are still obscure, but most scholars would agree that it provides an unerring guide to the thought and psychology of a people. The directness, clarity, and transcendent humanity of the Greeks permeate all their works, and the imaginative appeal of their mythology, rather than any real or fancied symbolism, has stimulated creative effort. Men of all eras have been moved and baffled by the deceptive simplicity of Greek mythmaking, whether in word or stone.

BIBLIOGRAPHY. W.H. ROSCHER, *Ausführliches Lexikon der griechischen und römischen Mythologie*, 6 vol. (1884–1937), the authoritative encyclopaedia of Greek mythology; M.P. NILSSON, *The Mycenean Origin of Greek Mythology* (1932), a pioneer work, and *Cults, Myths, Oracles and Politics in Ancient Greece* (1951), an excellent survey; KAROLY KERENYI, *Die Mythologie der Griechen* (1951; Eng. trans., *The Gods of the Greeks*, 1951), contains detailed data; and *Die Heroen der Griechen* (1958; Eng. trans., *The Heroes of the Greeks*, 1959), a dictionary of saga; H.J. ROSE, *A Handbook of Greek Mythology*, 6th ed. (1960), the most comprehensive handbook in English; RHYS CARPENTER, *Folktale, Fiction and Saga in the Homeric Epics* (1946), a lively comparative account; J.E. FONTENROSE, *Python: A Study of Delphic Myth and Its Origins* (1959), a massive comparative account with full bibliography; M. GRANT, *Myths of the Greeks and Romans* (1962), discussion of chief myths and their subsequent history; R. GRAVES, *The Greek Myths*, 2 vol. (1955–64), a comprehensive account; J. POLLARD, *Helen of Troy* (1965), a popular account of Trojan saga; P. WALCOT, *Hesiod and the Near East* (1966), a discussion of Oriental origins of Greek myth; G.S. KIRK, *Myth: Its Meaning and Functions in Ancient and Other Cultures* (1970), a comprehensive critical account; A.G. WARD *et al.* (eds.), *The Quest for Theseus* (1970), a full, illustrated account.

(J.R.T.P.)

Greek Religion

Greek religion is concerned with the beliefs of the ancient Hellenes about gods, conceived in human terms, and the way they worshipped them. It is not the same as Greek mythology, which is concerned with traditional tales, though the two are closely interlinked. Curiously, for a people so religiously minded, the Greeks had no word for religion itself—the nearest term being *eusebeia* (piety).

Greek religion, in its developed form, lasted for more than a thousand years, from the time of Homer (probably 9th or 8th century BC) to the reign of the emperor Julian (4th century AD), though its origins may be traced to the remotest eras. During that period its influence spread as far west as Spain, east to the Indus, and throughout the Mediterranean world. Its effect was most marked on the Romans, who identified their deities with the Greek. Under Christianity, Greek heroes and even deities survived as saints, while the rival madonnas of southern European communities reflected the independence of local cults. The rediscovery of Greek literature during the Renaissance and, above all, the novel perfection of classical sculpture produced a revolution in taste that had far-reaching effects on Christian religious art. The most striking characteristic of Greek religion was the belief in a multiplicity of anthropomorphic deities, coupled with a minimum of dogmatism. Though pious and parochial in their attitude toward local cults, the Greeks welcomed such foreign deities as the Thracian Bendis

and Cotys (Cotytto), the Phrygian Cybele and Sabazius, and the Egyptian Isis and Sarapis.

The Greeks' reticent attitude toward their own religious beliefs constitutes in itself an obstacle to knowledge. Firsthand information about the details of ritual, such as the precise working of the Delphic oracle, is lacking; and of some important religious events, like the mysteries at Eleusis, little is known, since the initiates were bound to secrecy; and late Christian accounts are naturally biassed. Also, the early history of the Dionysiac cult is relatively obscure, apart from covert references in Homer, fanciful vase paintings, and the dramatic description given in Euripides' *Bacchae*. The main sources of knowledge are inevitably literary, for although archaeology has thrown much light on the externals of religion, it has not and cannot hope to reveal the feelings and attitudes of worshippers. The best sources are historical and antiquarian (the evidence of Pausanias in the 2nd century AD is uniquely valuable), though poetry and philosophy contribute a great deal. Inscriptions, too, are of vital importance in revealing the identities of buildings, cults, calendars, and sacred writings and are often the sole source of information.

HISTORY

The roots of Greek religion. Greek religion evolved during Neolithic times in the original home of the Greeks in Central Europe, where the Indo-European sky god, variously known as Zeus, Jupiter, and Dyaus, was believed to control the weather. As waves of Greek-speaking peoples moved south into the peninsula during the 2nd millennium BC, they absorbed Pelasgian (pre-Greek) cults such as those of the primitive oracle at Dodona, the river and wind gods, and horse-headed Demeter. Whether Greek deities were ever visualized in animal form is uncertain, though each possessed its familiar beast or bird. Some, like Hestia, the hearth goddess, in whose honour a perpetual fire was kept burning in public buildings, were only vaguely personified; while others, like Apollo, Hermes, and Dionysus, were sometimes represented as shaped stones, columns, stocks, or posts.

Indo-European nature deities

In southern Greece the newcomers developed a powerful suzerainty, centred at Mycenae, and came into contact with the Minoan nature religion. As a result, Zeus himself acquired a Cretan origin as well as a consort in the powerful Argive goddess Hera. Athena, the palace guardian, became Zeus's daughter, while Artemis, the "mistress of wild beasts," was said to be Apollo's twin. The period and direction of Apollo's advent are still unresolved, but the Dorians who destroyed the Mycenaean civilization about 1100 BC apparently assisted in popularizing his worship. The Dark Ages that followed the Dorian invasion ended in the 8th century BC, though the traditions of previous eras were doubtless preserved by bards. Kings were succeeded by feudal aristocracies. The gods of Homer's *Iliad* and *Odyssey* appear to mirror the society for which he composed: wealthy, urbane, and easy living. Hesiod (fl. *c.* 800 BC), on the other hand, in his *Works and Days*, preached piety and ritualism and regarded Zeus as the principle of justice in an unjust world.

The Archaic period. Sometime before Homer composed his epics, the orgiastic cult of the nature divinity Dionysus reached Greece from Thrace and Phrygia. Because the god's name is Greek and has been recognized in a Mycenaean tablet, it has been suggested that his worship represented not so much a novelty as a reversion to Minoan religion. His devotees, armed with *thyrsoi* (wands tipped with pinecones) and known as Maenads ("mad ones") or Bacchants, were reputed to wander in *thiasori* (revel bands) about mountain slopes, such as Cithaeron or Parnassus, right down into Roman imperial times. They were also supposed, in their ecstasy, to practice the *sparagmos*, the tearing of living victims to pieces and feasting on their raw flesh (*ōmophagia*). This orgiastic cult of Dionysus, however, was subsequently civilized and tamed. Tragedy developed from the choral song of Dionysus.

Dionysus and the Eleusinian mysteries

Since the nobles owned the land they naturally tended to control the agrarian cults, and this practice continued until comparatively late times. Besides, they were also clan chieftains; and long after the political reorganization of Athens in the late 6th century BC, the worship of clan deities, such as Athena Phratria (Protectress of the Clan) and Zeus Phratrios (Protector of the Clan), remained popular. The 7th and 6th centuries BC were characterized by the revolt of the underprivileged against the nobles. Popular leaders, known as tyrants, seized power in the city-states; and some, like the Peisistratids at Athens, enhanced their glory by building temples and reviving religious festivals. Political unrest was reflected in spiritual unrest. Initiation into the mysteries of Demeter and Kore at Eleusis fulfilled a deep religious need, while oracles, with their claim to reveal the future, were eagerly consulted by individuals and states. Delphi became the cultic centre not only of Greece but of Lydia also. In addition to the official cults, credence was placed in soothsayers and the ability of adepts to evade death and undertake psychic journeys.

The Classical period. During the 6th century BC the rationalist thinking of Ionian philosophers had offered a serious challenge to traditional religion. At the beginning of the 5th century, Heracleitus of Ephesus and Xenophanes of Colophon heaped scorn on cult and gods alike. The Sophists, with their cynical probing of accepted values, continued the process, as did the dramatists Euripides and Aristophanes. Despite these attacks, the civic cults at Athens and elsewhere attained their apogee. The old Peisistratid temples destroyed during the Persian sack of 480 BC were gradually replaced by larger and more imposing edifices, culminating in the Parthenon of Pericles.

Festivals were expressive of religion's social aspect and attracted large gatherings (*panēgyreis*). Being mainly agrarian in origin, they were seasonal in character, held often at full moon and on the 7th of the month in the case of Apollo, and always with a sacrifice in view. Many were older than the deity they honoured, like the Hyacinthia and Carneia in Laconia, which were transferred from local heroes to Apollo. The games were a special kind of festival, sometimes forming part of other religious events.

Popular religion flourished alongside the civic cults.

Gad Borel—Boissonnas



Painting showing a dead man (right) receiving an offering of a votive ship and two calves, and the priestess (left) pouring a libation. From a terra-cotta sarcophagus from the necropolis of Ayia Triádha, Crete, *c.* 1400 BC. In the Archaeological Museum, Crete.

Peasants worshipped the omnipresent deities of the countryside, such as the Arcadian goat-god Pan, who prospered the flocks, and the Nymphs (like Eileithyia, they aided women in childbirth) who inhabited caves, springs (Naiads), trees (Dryads and Hamadryads), and the sea (Nereids). They also believed in nature spirits such as satyrs and sileni, who were mimed in drama, and equine centaurs. Among the more popular festivals were the rural Dionysia, which included a phallus pole; the Anthesteria, when new wine was broached and offerings were made to the dead; the Thalysia, a harvest celebration; the Thargelia, when a scapegoat (*pharmakos*) assumed the communal guilt; and the Pyanepsia, a bean feast in which boys collected offerings to hang on the *eiresiōne* ("wool pole"). Women celebrated the Thesmophoria in honour of Demeter, and commemorated the passing of Adonis with laments and miniature gardens, while images were swung from trees at the Aiora to get rid of an ancient hanging curse. Magic was widespread. Spells were inscribed on lead tablets. Statues of Hecate, goddess of witchcraft, stood outside dwellings, while Pan's image was beaten with herbs in time of meat shortage.

The Hellenistic period. Though the city-states were vanquished by Alexander the Great, loyalty to the old deities remained remarkably tenacious, and colonists spread their worship abroad. Apollo (and the Sibyl) reached Italy during the 8th century; other cults followed, and the process of identification with native counterparts began. It was always more marked in mythology than cult, and some native deities (*e.g.*, Quirinus) were little affected. Cybele was introduced from Phrygia in 204 BC, but the lodges of Bacchus were suppressed in 186 BC.

The Hellenistic era was characterized by the popularity of mystery religions. The Cabeiri, or Great Gods, of Samothrace were patronized both by the Ptolemies (the Macedonian dynasty ruling Egypt) and the Romans, while the Egyptian cults of Isis and Sarapis, though already Hellenized, quickly spread over the classical world. Syncretism—*i.e.*, the blending of deities' functions and identification with foreign gods—was a feature of the post-Alexandrian world. Astrology was studied seriously by astronomers and philosophers, but on a lower plane it descended into quackery. Christianity then arose at a time of political and religious crisis. Papan views of the afterlife remained unsatisfactory, whereas the new creed that recognized neither national nor class barriers brought the hope of personal salvation to all.

BELIEFS

The gods. The Greeks regarded the gods as the immortal controllers of natural forces, which were suitably departmentalized. Zeus, as weather god, was known by such cult titles as Maemactes (Stormy) and Cataebates (Striker), and in times of drought his mountain shrines were visited. Poseidon was god of the sea, Hades (or Pluto) of the underworld; Demeter was goddess of the harvest, Hera of marriage, and Athena of courage, craft (Ergane), and victory (Nike). Apollo visited men with plagues or alternatively sanctioned purifications, Ares personified war, Themis justice, and Aphrodite and Eros roused sexual passion. The individual spheres were nonetheless elastic, so that Zeus also protected guests under the title of Xenios, guarded the house as Herkeios or Ktesios and, more sinisterly, was associated with lycanthrophy (human assumption of the wolf form) and cannibalism as Lycaeus. Apollo was a god of herdsmen (Nomios) and prophecy (Pythios) and Poseidon was a god of horses (Hippios) as well as a causer of earthquakes. Other powerful deities were Nemesis at Rhamnous; Athena, surnamed Alea, at Tegea; Aphaea in Aegina; Artemis Orthia at Sparta; and Despoina (the Mistress) in Arcadia.

The gods in Homer and Hesiod

Homer assembled the major deities, who gave unity to Greek beliefs, on Olympus for his own epic purposes. As portrayed by him the gods are fair of form and move, for all their suprahuman size, with the lightness of birds. Freed, for the most part, from the trammels of cult, they have leisure to indulge their whims. Often they appear to give visible expression to human emotions or, like the Muses (who had cults in Pieria and on Helicon), personified inspiration. Of these cynical and arrogant gods, who were yet terrible on occasion, only Zeus is nobly drawn.

Hesiod composed a detailed genealogy of the gods in his *Theogony,* which appears to depend upon Oriental prototypes. The treatment is entirely mythical, and the gods are derived ultimately from Gaea (Earth) and Uranus (Heaven).

Superior to the gods was *moira* (fate). Even Zeus was unable to interfere with its workings. Moira and Ker (doom) come close in meaning, and the heroes' fates were literally in the balance when Zeus hung out his scales. Sometimes Zeus was the instrument of *moira,* but he never controlled it in the final outcome. When they could not be more specific, the Greeks referred to a *daimōn* (power) as being responsible for some event. According to Hesiod, *daimōnes* were the guardian spirits of those who perished during the Golden Age, but in later Greek religion they were regarded as beings intermediate between men and gods.

Hero cults

Offerings were made to dead Mycenaean kings, but the cult of heroes was a post-Homeric phenomenon. Heroes were regarded as inferior to gods, although some, such as Heracles, the Dioscuri (the heavenly twins), and Asclepius (the healer), were not far removed from divinity. Major heroes included Ajax of Salamis, who was worshipped in an ancestral capacity; Theseus, king of Athens, and Pelops of Olympia, whose shrines were places of asylum and pilgrimage; and Triptolemus at Eleusis. In addition, personalized powers were also revered: Cyamites (Bean-Man), Amynos (Protector), or Echetlaeus (Plow-Handle), who was said to have fought at Marathon. Hero cults were managed by elected officials known as *orgeōnes.* Cults of heroines, such as Helen, were also known.

Cosmogony. Hesiod stated that Chaos was the primal condition of the universe. Then came Earth, visualized as a flat disk, girdled by the stream of Ocean, Tartarus (Hell), and Eros (Love), the active element. From these were born Erebus (the dark void) and Night. Earth bore Heaven and by him became the mother of the Titans, Pontus (Sea), and the river of Ocean, as well as other beings. The Orphics modified Hesiod's system by introducing the novelty of the cosmic egg. From the egg sprang Phanes, the primal deity, who married Night and begat Heaven and Earth.

Man. For the Greeks the universe was man centred. Man consisted of a body (*sōma*) and a soul (*psychē*). The *psychē* was the death rattle, the last exhalation of breath, visualized as a pale replica of the living person; it flew away gibbering to the house of Hades, attracted, according to Plutarch, by the music of the Sirens, whose monstrous forms crowned tombs. The beliefs of certain ascetics—known as followers of Orpheus, the legendary minstrel—that the body was the prison of the soul and that everyone ultimately shared in the divine gave rise to the dualism familiar in Plato and Christianity.

Man's relation to the gods had two aspects: formal and personal. As a member of the community, he shared in the sacrifices proffered for the fortune of the state; he also shared in the god's displeasure if there was any deviation from the accepted ritual or hint of physical pollution. The manslaughter of suppliants was regarded as especially heinous and could endanger the general weal. Private sacrifices were also made and offerings of various kinds, including vases, statuettes, or even large statues that were dedicated in the sacred precincts (temene).

In addition to the Olympians, the Greeks worshipped chthonian (underworld) deities such as Zeus Meilichios (the Gracious, a euphemistic title), who rid Athens of pollution. Like Erichthonius, the ancient Attic hero, and certain domestic deities, he was sometimes represented in snake form. Domestic cults included the formal acceptance of a newborn child into the family by parading it round the hearth (Amphidromia) in the presence of the

Moirai (Fates), which are still honoured by women in modern Greece. Libations were made to the gods, including the Agathos Daimon (Good Spirit), who protected the household, while the Dioscuri were sometimes invited to share a feast (Theoxenia). Outside each dwelling stood the tutelary pillar of Apollo Agyieus (of the highway) or a "herm" (stock embellished with the head and genitals of Hermes), which was set up in the streets of Athens and elsewhere to bring good luck.

Immortality and death

Eschatology. In Homer only the gods were immortal, but Elysium was reserved for their favoured sons-in-law, who were also exempt from death. Heracles alone gained a place on Olympus by his own efforts. The ordinary hero hated death, for the dead were regarded as strengthless doubles who had to be revived with draughts of blood, mead, wine, and water in order to enable them to speak. They were conducted, it was believed, to the realm of Hades by Hermes; but the way was barred, according to popular accounts, by the marshy river Styx. Across this, Charon ferried all who had received at least token burial, and coins were placed in the mouths of corpses to pay the fare. Originally only great sinners like Ixion, Sisyphus, and Tityus, who had offended the gods personally, were punished in Tartarus. But the doctrines of the Orphics influenced Pindar, Empedocles and, above all, Plato. According to the latter, the dead were judged in a meadow by Aeacus, Minos, and Rhadamanthus and were consigned either to Tartarus or to the Isles of the Blest. Long periods of purgation were required before the wicked could regain their celestial state, while some were condemned forever. The dead were permitted to choose lots for their next incarnation, but usually their choice was unwise. Subsequently they drank from the stream of Lethe, the river of oblivion, and forgot all.

WORSHIP, PRACTICES, AND INSTITUTIONS

Sacred writings. Greek religion was not based on a written creed or body of dogma. Nevertheless, certain sacred writings survive in the form of hymns, oracles, inscriptions, and instructions to the dead. Most elaborate are the Homeric Hymns, some of which may have been composed for religious festivals, though their subject matter is almost entirely mythological. Delphic inscriptions include hymns to Apollo but, like the Epidaurian hymn by Isyllus to Asclepius, they are not concerned with liturgy. Delphic oracles are quoted from literary sources but appear, on the whole, to be retrospective concoctions, like the Hebraic–Hellenistic collection of Sibylline prophecies. Questions scratched on folded lead tablets have been found at Dodona, and detailed instructions to the dead, inscribed on gold leaf and possibly of Orphic inspiration, have been found in Greek graves in southern Italy. More recently papyrus fragments of similar character have been recovered from graves in Macedonia and Thessaly.

Shrines and temples. In the earliest times deities were worshipped in awesome places such as groves, caves, or mountain tops. Mycenaean deities had no temples of their own but shared the king's palace. Fundamental was the precinct (temenos) allotted to the deity, containing the altar, temple (if any), and other sacral or natural features, such as the sacred olive in the temenos of Pandrosos on the Athenian Acropolis. *Naoi* (temples—literally "dwellings"—that housed the god's image) were already known in Homeric times and, to judge from models discovered at Perachora, were of wood and simple design. Poros and marble replaced wood by the end of the 7th century BC when temples became large and were constructed with rows of columns on all sides. The image, crude and wooden at first, was placed in the central chamber (cella), which was open at the eastern end. No ritual was associated with the image itself, though it was sometimes paraded. Hero shrines were far less elaborate than temples and were provided with pits for offerings. Miniature shrines also were known.

Oracles and divination

Most oracular shrines included a subterranean chamber, but no trace of such has been found at Delphi, though the Pythia was always said to "descend." At the oracle of Trophonius, discovered in 1967 at Levádhia, incubation was practiced in a hole. The most famous centre of incubation was that of Asclepius at Epidaurus. His temple was furnished with a hall where the sick were advised by the demigod in dreams. Divination was also widely practiced in Greece. Augurs interpreted the flight of birds, while dreams, and even sneezes, were regarded as ominous. Seers also divined from the shape of altar smoke and the conformation of victims' entrails.

Priesthood. Even in the state cults, priesthoods were frequently ancestral prerogatives. Eteobutads organized the cult of the hero-king Erechtheus at Athens; Praxiergids superintended the washing of Athena's robes at the Plynteria; Clytiads and Iamids officiated at the altar of Zeus at Olympia. Although there was no official clergy, since the religious and secular spheres were not sharply divided (*e.g.*, the wife of the *archōn basileus* was ritually wed to Dionysus at the Anthesteria), professional assistance was available at sacrifices. There was no necessary correspondence between the sex of deities and that of priests. Hera and Athena favoured priestesses, but Isis and Cybele favoured priests. Apollo again inspired the Pythia (priestess) at Delphi but a priest at Ptoon. The mysteries at Eleusis were stage-managed by the Eumolpids and Kerykes. The latter assembled the *mystai* (initiates), while the former provided the Hierophant, who revealed the mysteries in the torchlit Anaktoron (king's shrine) within the great Telesterion, or entrance hall.

Festivals. The precise details of many festivals are obscure. Among the more elaborate was the Panathenaea, which was celebrated at high summer, and every fourth year (the Great Panathenaea) on a more splendid scale. Its purpose, besides offering sacrifice, was to provide the ancient wooden image of Athena, housed in the "Old Temple," with a new robe woven by the wives of Athenian citizens. The Great Panathenaea included—in addition to a procession—a torch race, athletic contests, mock fights, and bardic recitations. Next in importance was the Great Dionysia, celebrated at Athens in spring. At the end of the ritual the god's image was escorted to the theatre of Dionysus, where it presided over the dramatic contests. It, like its rural counterpart, included phallic features.

Olympic Games

The Olympic Games formed part of the great festival of Zeus held every fourth summer in the god's sacred precinct—the Altis beside the river Alpheius in the western Peloponnese. A truce was proclaimed in order to permit any warring Greeks to compete, and the celebrations lasted five days. Sacrifice and libation were made at the altar of Zeus, where omens were taken and oracles proclaimed, and at the tomb of Pelops and the altar of Hestia. Competitors and judges took the oath to observe the rules, processions were held, bards recited, and winners were honoured at state banquets. The richer and more famous were immortalized by lyric poets, such as Semonides, Bacchylides, and Pindar. Though women were banned, girls competed at the festival of Hera. The games held in honour of Zeus at Nemea, Apollo at Delphi, and Poseidon at the Isthmus followed the Olympian pattern.

Rites. Sacrifice was offered to the Olympian deities at dawn at the altar in the temenos, which normally stood east of the temple. Representing as it did a gift to the gods, sacrifice constituted the principal proof of piety. The gods were content with the burnt portion of the offering, while the priests and worshippers shared the remainder of the meat. Different animals were sacred to different deities—*e.g.*, heifers to Athena, cows to Hera, pigs to Demeter, bulls to Zeus and Dionysus, dogs to Hecate, game and heifers to Artemis, horses to Poseidon, and asses to Priapus—though the distinctions were not rigorously observed. The practices of ritual washing before sacrifice, sprinkling barley grains, and making token offerings of hair are described by Homer. Victims were required to be free of blemish, or they were likely to offend the deity. Sacrifice also was made to chthonian powers in the evening. Black victims were offered, placed in pits, and the meat was entirely consumed. Sacrifice preceded battles, treaties, or similar events. Human sacri-

fice, though legendary and even historical instances were reported, appears, if it was practiced at all, to have been the exception. Bloodless sacrifices were made to some deities and heroes.

Prayers and processions

Prayers normally began with compliments to the deity, followed by discreet references to the petitioner's piety, and ended with his special plea. If made to an Olympian, the suppliant stood with his arms raised palm upward. Processions formed part of most *panēgyreis* (gatherings) and festivals. The Panathenaic procession set out from the Pompeion (sacred storehouse) at dawn, headed by *kanēphoroi* (maiden basket bearers), who carried the sacred panoply. Elders bore boughs (*thallophoroi*) while youths (*ephēboi*) conducted the victims for sacrifice, and cavalry brought up the rear. The robe was spread on the mast of a wheeled ship.

The procession to Eleusis to restore the sacred objects, brought by the *ephēboi* to the Eleusinium some time previously, followed the wooden image of Iacchus (a personification of the ritual cry), which was escorted by its own priest, the *iacchagogos*, and officials. The *mystai* wore myrtle crowns and carried corn sheaves. Whatever the nature of the mysteries, those initiated returned in a mood of exaltation. Adepts (*epoptai*) were later admitted to more solemn rites (to see an ear of wheat, scoffers said).



Painted Greek vase showing a Dionysiac feast, 450–425 BC. In the Louvre, Paris.
Andre Held—Ziolo

Religious art and iconography. Art illuminates Greek religion, though the scenes represented are often obscure and sometimes completely baffling. On a well-known sarcophagus from Ayía Triádha in Crete, a priestess dressed in a skin skirt assists at a sacrifice, flanked by wreathed axes on which squat birds. The significance of the scene has been much discussed. The birds have been regarded as epiphanies of deities, giving sacral meaning to the transformations in Homer. Again, since goddesses appear to preponderate in Minoan–Mycenaean art, while male deities are represented on an inferior scale, this has been thought to reflect the general superiority of goddesses in many parts of Greece. In the earliest period, terra-cotta statuettes of deities were small and crude, while the old cult images were made of wood and commonly attributed to Daedalus. When artists turned to bronze and marble, deities were fashioned with increasing realism. Gods were shown naked and often bearded, while goddesses were robed until the 4th century BC. Since religious architecture was dominant in ancient Greece, it attracted the leading artists. Phidias was commissioned to construct the colossal chryselephantine statue of Athena for the Parthenon. He also made the image of Zeus for the temple at Olympia, as well as the bronze Athena on the Acropolis. Though the cult statue has disappeared, the masterly rendering of the Panathenaic procession on the Parthenon frieze reflects the dignity of contemporary religious feeling. A famous mural by Phidias' contemporary Polygnotus on the hall of the Cnidians at Delphi depicted scenes in the underworld with gruesome realism, according to Pausanias' description.

The sublimity of 5th-century religious art was rarely maintained in succeeding eras. The grave majesty of the Demeter of Cnidus is a notable exception, as indeed is the Aphrodite from Melos; but for all the liveliness of Scopas' sculpture from the temple of Athena Alea at Tegea and the moving perfection of Praxiteles' Hermes, their work patently heralds a religious decline.

Apart from cult statues and dedications like the Acropolis *korai* ("maidens"), the gods frequently were represented on the pediments, metopes, and friezes of temples, usually in mythological scenes. For the details of ritual, vase painting has proved a fruitful source of information. Dionysiac subjects are common, though usually imaginary, but cult scenes and fertility customs also appear.

The evolution of Greek religion

The history of Greek religion, including beliefs, forms of worship, and visible expression in art, has been traced from the earliest times to its final absorption in Christianity. It falls naturally into stages because, unlike more authoritarian and rigid faiths, men's view of the gods and of the nature of piety and virtue were always changing. Though its essence was stable, Greek religion was never static but was always evolving. It was a polytheism that, paradoxically, contained within it the seeds of a monotheism that was already latent in Homer and Hesiod. Homer's gods form a hierarchy under Zeus, who in Hesiod is the guarantor of right and justice. Aeschylus, the tragedian, also saw in Zeus something like Providence. Poets and philosophers often referred to God, without further specification, as the ultimate arbiter of the universe. Nevertheless, polytheism never wholly perished. Festivals and processions still form an important part of Christian pageantry, while the healing shrine at Epidaurus finds a counterpart in Lourdes. Mysticism is derived in great measure from the mysteries, and the doctrine of free will is at least Greek in spirit. By insisting always on the primacy of man, the Greeks left him at liberty to worship God as he pleased.

BIBLIOGRAPHY

General: M.P. NILSSON, *Greek Popular Religion* (1940; reissued as *Greek Folk Religion,* 1961), a sound and detailed survey, *Grekisk religiositet* (1946; Eng. trans., *Greek Piety,* 1948, reprinted 1969), a general survey, *The Minoan-Mycenaean Religion and its Survival in Greek Religion,* 2nd rev. ed. (1950), the best account of origins, and *Geschichte der griechischen Religion,* 2nd ed. (1955), the standard history; H.J. ROSE, *Ancient Greek Religion* (1948), a brief but masterly sketch; W.K.C. GUTHRIE, *The Greeks and Their Gods* (1950), the best general account; *The Religion and Mythology of the Greeks* (1961), a brief sound sketch of origins; J. POLLARD, *Seers, Shrines and Sirens* (1965), Greek religion in the 6th century BC; G.S. KIRK, *Myth* (1970), deals exhaustively with such problems as myth and religion.

Oracles and divinations: W.R. HALLIDAY, *Greek Divination* (1913, reissued 1967), still the best account in English; P. AMANDRY, *La mantique apollinienne à Delphes* (1950), the most authoritative account; H.W. PARKE and D.E.W. WORMELL, *Delphic Oracle,* 2 vol. (1956), the fullest account in English; R. FLACELIERE, *Devins et oracles grecs* (1961; Eng. trans., *Greek Oracles,* 1965), a useful account including divination; H.W. PARKE, *Greek Oracles* (1967), the best account in English, and *The Oracles of Zeus: Dodona, Olympia, Ammon* (1967).

Mysteries and eschatology: E. ROHDE, *Psyche: The Cult of Souls and Belief in Immortality Among the Greeks* (1925; pub. orig. in German, 1890), the fundamental work; W.K.C. GUTHRIE, *Orpheus and Greek Religion,* 2nd rev. ed. (1952), the best work on Orphism; I.M. LINFORTH, *The Arts of Orpheus* (1941), a hypercritical account; E.R. DODDS, *The Greeks and the Irrational* (1951), the best account since Rohde; G.E. MYLONAS, *Eleusis and the Eleusinian Mysteries* (1961), a good general survey; K. KERENYI, *Eleusis: Archetypal Image of Mother and Daughter* (1967), a psychological account; W.F. JACKSON KNIGHT, *Elyson* (1970), ancient views of the afterlife.

Cults and festivals: L.R. FARNELL, *Cults of the Greek States,* 5 vol. (1896–1909), the best critical survey in English, and *Greek Hero Cults and Ideas of Immortality* (1921), a formal and critical account; M.P. NILSSON, *Griechische Feste* (1906, reissued 1957), the standard work on non-Attic festivals; A.B. COOK, *Zeus: A Study in Ancient Religion,* 3 vol. (1914–40), a

monumental compendium of all the evidence; L. DEUBNER, *Attische Feste* (1932, reissued 1966), the standard work on Attic festivals; A.D. NOCK, "The Cult of Heroes," *Harvard Theol. Rev.*, 37:144–174 (1944), a masterly survey; E.J. and L. EDELSTEIN, *Asclepius*, 2 vol. (1945–46), the best account in English; K. KERENYI, *Der göttliche Arzt* (1956; Eng. trans., *Asklepios: Archetypal Image of the Physician's Existence*, 1959), a psychological account; L. DREES, *Olympia* (1967, Eng. trans., 1968), a full popular account of the festival.

Art and architecture: V.J. SCULLY, *The Earth, the Temple and the Gods: Greek Sacred Architecture* (1962), a full, if somewhat fanciful account of temple siting; H. BERVE and G. GRUBEN, *Greek Temples, Theatres, and Shrines* (1963), a detailed survey of the chief buildings; B. BERGQUIST, *The Archaic Greek Temenos: A Study of Structure and Function* (1967), a scholarly survey.

(J.R.T.P.)

Greenland

Greenland (Danish Grønland), the largest island in the world and, since 1953, an integral part of the Kingdom of Denmark, lies in the Northern Hemisphere, on a northeastern extension of the structural shelf fringing the continent of North America. Its icy, inhospitable environment, lying mostly within the Arctic Circle and extending to within less than 500 miles (805 kilometres) of the North Pole, makes the territory appear to be of little importance to man. Yet it possesses considerable potential, as attested by the flow of investment following the abolition, in 1951, of the governmental commercial monopoly. Indeed, on close examination it can be characterized as one of the few northern parallels to the developing regions of the Southern Hemisphere.

Physical dimensions

Greenland forms a wedge-shaped mass of about 840,000 square miles (2,175,600 square kilometres), of which over 700,000 square miles (1,800,000 square kilometres) are ice covered. Its maximum north–south extension is 1,650 miles (2,650 kilometres), and it is over 750 miles (1,200 kilometres) across at its widest point, at about 70° N. The length of its indented coastline has been calculated at 24,430 miles (39,310 kilometres)—almost exactly equal to that of the circumference of the earth at its Equator. Its southernmost tip is Cape Farewell (Kap Farvel), 59°45′ N, while Cape Morris Jesup, at 83°39′ N, was, until a 1969 recalculation of the position of Kaffeklubben Island, some 20 nautical miles to the east, considered as the nearest land to the North Pole. The westernmost point, Cape Alexander, at 73°05′ W, is farther west than the American city of Boston, while its easternmost point, Nordost Rundingen, at 12°09′ W, is almost as far east as Ireland. At the narrowest point, Greenland is only 16 miles (26 kilometres) from Ellesmere Island in the Canadian north. A submarine ridge, no deeper than 600 feet (183 metres), connects the island physically with North America, from which it is separated, from north to south, by the Nares Strait, the Robeson Channel, the Kennedy Channel, the Kane Basin, Smith Sound, Baffin Bay, Davis Strait, and the Labrador Sea. To the north lies Lincoln Sea and the ice masses of the Arctic Ocean, while on the east and south lie the Greenland Sea, Denmark Strait, and the Atlantic Ocean. At a depth of about 2,000 feet (610 metres) another submarine ridge (the Yermak Plateau) connects Greenland with Spitsbergen, while at the same depth the Faeroe-Iceland Ridge snakes across the ocean floor to Scotland and the European continent.

THE LANDSCAPE

Structurally, Greenland is an extension of the Laurentian Shield, the vast, rough plateau dominating the Canadian north and made up of hard Precambrian gneiss and granite rocks, among the most ancient of the earth's surface. In addition, the northern shoulders of Greenland, notably the mountains of Peary Land and their westward extension to Ellesmere Island, are structurally linked, via Spitsbergen, to the Caledonian orogeny of Europe.

From the dawn of geological time, the vast mass of Greenland, now largely obscured by ice, has been molded by a series of earth movements and related processes. At the earliest period, volcanic activity took place along the eastern edge. Later, sand and clay were washed out from ancient mountains into a long structural trough along the east coast. This became filled with layers of sandstone, limestone, and shale many thousands of feet thick, which were then uplifted to form the fold mountains of east Greenland. In more recent times, sandstones were laid down on both the east and west coasts, and coal of the Cretaceous Period, about 100,000,000 years ago, is found in the coastal areas of Disko Island and the Nûgssuaq Peninsula. Later still, some 50,000,000 years ago, further volcanic activity formed horizontal layers of hardened lava streams, which can still be seen as black bands interlacing exposed rock faces. Basaltic survivals of this period are at the Scoresbysund hot springs area in the east, at Julianehåb in the south, and on Disko Island, and Nûgssuaq and Svartenhuk peninsulas in the west. The many fossils in the Cretaceous sandstones and coal deposits testify that a temperate climate then existed.

The onset of the ice age, a million or so years ago, totally changed the Greenland environment. Vast accumulations of ice slid from the interior to the coasts, scouring away the rock masses beneath through debris frozen to the base of the slowly moving ice. Lakes formed on the coastal periphery, and plains also developed from sediments washed out from under the advancing ice; it was these sand and clay surfaces that were used for military airfields during World War II. Today, the indented, island-strewn coast forms, in part, a narrow, ice-free fringing strip. Long, deep fjords—drowned former valleys—reach far into both the east and west coasts in complex systems, offering magnificent, if desolate, scenery. The Scoresbysund network on the east coast is the largest, with a length of about 185 miles (300 kilometres) and a breadth of 125 miles (200 kilometres). A range of mountains 7,000 feet (2,130 metres) high also runs along the east coast, with Mt. Gunnbjorn, 12,139 feet (3,700 metres), marking the highest point in Greenland. Among many parts of the coast, the ice sheet fronts directly on the sea, with large chunks breaking off the tongueing glaciers and sliding into the waters as icebergs.

The ice sheet

The ice sheet itself, the major feature of the Greenland landscape, is the largest ice mass outside of Antarctica. It is contained by coastal mountains on the east and, to a lesser extent, on the west, although the rock floor far beneath its surface is at, or slightly beneath, current sea level. The average depth of the ice is 5,000 feet (1,500 metres), with a maximum of 10,000 feet (3,000 metres). Layers of snow falling on its barren, windswept surface become compressed into ice layers, which constantly move outward to the peripheral glaciers—the Jakobshavn Glacier, often moving 100 feet (30 metres) a day, is among the world's fastest. Nunataks—lofty, isolated peaks—emerge occasionally around the rim of the ice sheet, whose total area is 708,100 square miles (1,833,900 square kilometres).

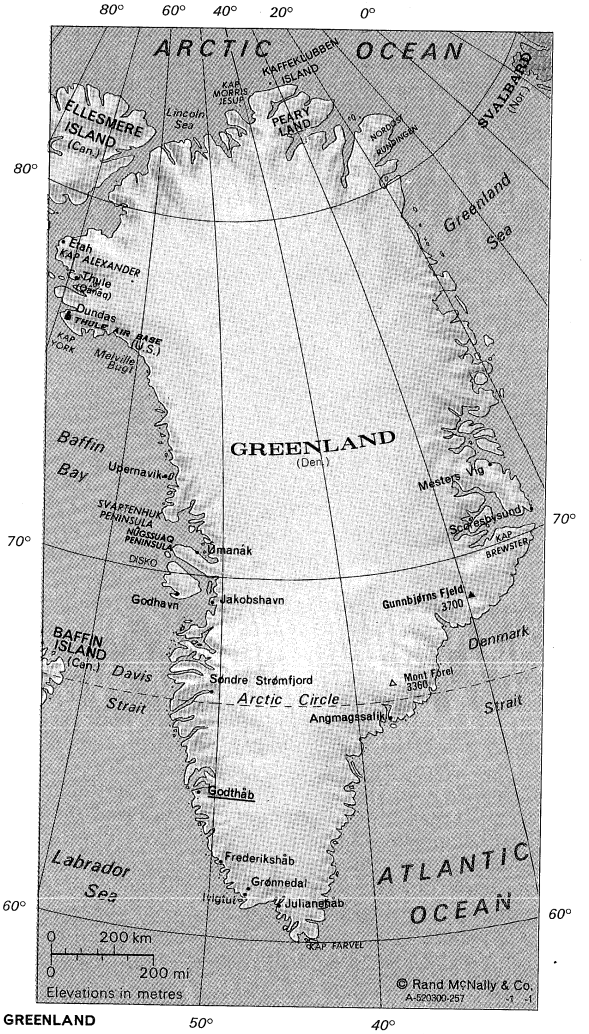
Climate. The climate of Greenland is bleak and Arctic, modified only by the slight influence of the Gulf Stream in the southwest. Rapid changes, from dazzling sunshine to impenetrable blizzards, are common and result from the eastward progression of low-pressure air masses over a permanent layer of cold air above the ice. January average temperatures at Ivigtut are 18° F (−7.8° C) with July readings of 49.5° F (9.7° C), while at Thule, the American military base in the north, the corresponding figures are −7.5° F (−21.9° C) and 41° F (5° C). Precipitation decreases from 40 inches (1,000 millimetres) per year in the south to eight inches (200 millimetres) per year in the north, with summer rainfall concentrated in the southwest; snow can, and does, fall in any month. Although summers can be quite pleasant on the southwest coast, the inland ice is uniformly cold, with a July average of 10° F (−12.2° C) and a February mean of −53° F (−47.2° C). In 1966 American scientists drilled a 4,500-foot (1,310-metre) ice core near Thule, detailing climatic fluctuations up to 100,000 years ago and enabling correlations with the fossil "raised beaches" of the coasts, which indicate relative changes in sea and land levels. The climate became noticeably warmer during the early part of the 20th century, with a 1930 maximum, associated with northward

Recent climate changes

seal movements and an increase in the cod population. By the late 1960s and early 1970s, sea temperature was dropping again, causing a serious decrease in cod fishing.

Vegetation. The vegetation of Greenland is represented mainly by tundra types, with heather, birch, willow, and alder scrub together with sedge, cotton grass, and lichen. The Greenland summer is rich in plant life. Of the 400 or more species, 300 are of North American origin, 50 or so were imported by the earliest Norse inhabitants, and the rest, on the isolated nunataks, have survived the ice ages. In southernmost Greenland, winter hay is grown for sheep, and along the western coast such vegetables as radishes, cabbage, and lettuce are raised, with potatoes in the extreme south.

Animal life. The rich animal life of the surrounding seas is the basis of existence for Greenlanders. The sea mammals—seals and whales—were formerly the main sources of nourishment, with ring- and black-sided seals predominating. Land mammals are represented by seven species—polar bears, musk-oxen, reindeer, arctic foxes, snow hares, ermines, and lemmings. Half of the breeding birds are native, and most of the remainder are from North America. The most important sea birds are eiders, guillemots, auks, wild geese, ducks, and gulls, while land birds include ptarmigans, ravens, white-tailed eagles, gyrfalcons, peregrine falcons, snowy owls, snow buntings, and longspurs. Among insects, mosquitoes appear, surprisingly enough, in summer. Salmon and trout are found in the rivers, while cod, salmon, flounders, halibuts, and *angmagssat* (capelin) are important saltwater fish.



GREENLAND

MAP INDEX

Cities and towns

Angmagssalik...65·36n 37·41w
Dundas.....76·34n 68·48w
Etah.....78·19n 72·38w
Frederikshåb.....62·05n 49·30w
Godhavn.....69·15n 53·33w
Godthåb.....64·11n 51·44w
Grønnedal.....61·20n 48·00w
Ivigtut.....61·10n 48·00w
Jakobshavn.....69·10n 51·00w
Julianehåb.....60·43n 46·01w
Mesters Vig.....72·15n 24·00w
Qânâq, see Thule
Scoresbysund...70·30n 22·00w
Søndre Strømfjord.....67·00n 50·59w
Thule (Qânâq)...77·28n 69·12w
Ûmanâk.....70·41n 52·09w
Upernavik.....72·50n 56·00w

Physical features and points of interest

Alexander, Kap, *cape*.....78·10n 73·05w
Arctic Ocean....85·00n 20·00w
Atlantic Ocean...62·00n 36·00w
Baffin Bay.....73·00n 65·00w
Brewster, Kap, *cape*.....70·19n 22·05w
Davis Strait.....68·00n 58·00w
Denmark Strait..67·00n 30·00w
Disko, *island*.....69·50n 53·30w
Farvel, Kap, *cape*.....59·45n 44·00w
Forel, Mont, *mountain*.....67·00n 37·00w
Greenland Sea...78·00n 10·00w
Gunnbjørns Fjeld, *mountain*.....68·55n 29·53w
Kaffeklubben Island.....83·40n 31·15w
Labrador Sea....62·00n 57·00w
Lincoln Sea.....83·00n 56·00w
Melville Bugt, *bay*.....75·30n 63·00w
Morris Jesup, Kap, *cape*.....83·39n 33·52w
Nordost Rundingen, *cape*.....81·36n 12·09w
Nûgssuaq Peninsula.....70·10n 53·00w
Peary Land, *physical region*..82·40n 35·00w
Svartenhuk Peninsula.....72·00n 55·00w
Thule Air Base...76·34n 68·48w
York Kap, *cape*..75·53n 66·12w

Table 1: Greenland, Area and Population

	area		population	
	sq mi	sq km	1965 census	1970 census*
Territories				
Nordgrønland	41,200	106,700	1,000	1,000
Ostgrønland	44,800	115,900	3,000	3,000
Vestgrønland	46,000	119,100	36,000	43,000
Total Greenland	131,900†‡	341,700†	40,000	47,000

*Preliminary. †Ice-free; total area of Greenland is 840,000 sq mi (2,175,600 sq km), of which 708,100 sq mi (1,833,900 sq km) is covered by ice cap.
‡Figures do not add to total given because of rounding.
Source: Official government figures.

THE PEOPLE

Ancient and modern influences had, by the 1970s, created a uniquely blended society in Greenland. The strongest element stems from Eskimo culture, whose representatives, many with admixtures of European blood, are known as Greenlanders. By 1970 they numbered over 38,000 out of a total population of 47,000. The pure Eskimo are now found only in the far north, near Thule, and in east Greenland. In terms of origin, the Eskimo are believed to have crossed from North America to northwest Greenland, using the islands of the Canadian Arctic as stepping stones in a series of migrations that stretched from 4000 BC to AD 1000. Each wave of migration was represented by different cultures, beginning with Independence cultures (named after a nearby fjord), which followed the musk-ox and reindeer north to Peary Land. About the same time, the Sarqaq culture spread along the west coast and even reached east Greenland, although the lack of hunting grounds impeded movement around the northern rim of the island. Later migrations brought the Dorset, Dundas (Thule), and Inugsuk cultures.

The Eskimo contribution

Greenlanders today retain a clear linguistic and cultural identity (the Eskimo refer to themselves collectively as simply *inuit*, "people"). The Eskimo dialects, taken together, form a language group differing from all others. A Greenlander grammar (1851) and a dictionary (1871) built up an Eskimo literary language, and Eskimo is the predominant spoken language of Greenland. The material culture, especially in the north, includes such beautifully constructed practical hunting equipment as the kayak, often considered the most elegant boat in the world. The Eskimo, with his snow igloo, his dog team, and his hunting instruments, survives effectively in one of the most inhospitable places in the world.

The second element in Greenland society is represented by European, specifically Danish, influence. Apart from the early Norse settlers—whose colonies became extinct by the early 15th century as a result of climatic deterioration and who left only an archaeological legacy—the Eskimo remained the sole inhabitants of the island until 1721, when a Danish missionary initiated new settlement. Greenland was a closed society until 1951, and

Danish immigration was small, consisting largely of administrators. Faeroese and Danish settlement increased during the 1950s and 1960s. The modernization of Greenland society was given great impetus by developments in air communications and global strategy during and following World War II. Lying between the great superpowers of the Northern Hemisphere, Greenland became of prime importance in radar and air communications and in weather observations. One result was the construction of the giant United States base at Thule in 1951, although American personnel had little direct contact with, or influence on, the indigenous population.

Permanent settlement is extremely dispersed and entirely confined to the coastal fringe. Greenland entered the 1970s with 154 inhabited places, made up of 19 towns, 117 villages, and 18 weather stations and airports. Over the decade of the 1960s, urban population increased by some 14 percent, to a total of nearly 70 percent; this resulted from the superior fishing facilities of the towns and an associated decline in the village-based seal hunting. Population has grown in recent decades: in 1805 (the first census) it was 6,046; in 1925, 13,600; in 1950, 23,642; in 1960, 33,140; and in 1970, 45,600, of whom 7,200 were Danes. The capital, Godthåb, has a population (1969) of 7,200. A half dozen or so other towns have populations of 1,500 to 2,000. Half of the population is below 15 years of age. The population growth has been aided by improved medical facilities, resulting in a declining overall death rate (8.3 per 1,000 population in the early 1970s, contrasting with 25.5 per 1,000 during the period 1945–49), coupled with an unchanged high birth rate of around 40.0 per 1,000.

The contemporary population

THE ECONOMY

Seal hunting, once the mainstay of the Greenland economy, declined drastically in the early part of the 20th century, partly as a result of overintensive hunting and partly due to climatic amelioration. By the early 1970s, sealing was still the occupation of some 5,000 people, a fifth of whom were actually hunters, based mainly in Ûmának, Upernavik, Thule, and east Greenland. Catch value exceeded 3,600,000 kroner (7.518 kroner = $1 U.S.; 18.043 kroner = £1 sterling, December 1, 1970) with an export value approaching 5,400,000 kroner. Fishing, which flourished in the warmer waters, replaced seal hunting at the most important economic activity, and cod fishing, canning, and freezing have played major roles in modernizing Greenland economy and sustaining population growth. Some 10,000 people made a living from fishing in the early 1970s, with about 3,000 fishermen directly engaged in catching, using a fishing fleet of over 1,000 motorboats. The state owns 60 fishing plants, eight of which are of major importance, and there are also nine private plants. The decline in the cod catch consequent upon the lowering seawater temperature occasioned some concern in the late 1960s: the total catch dropped from 45,000 tons in 1967 to 34,000 tons in 1968, for example, causing a fall in export value from 76,000,000 kroner to 70,000,000 kroner. Shrimp and salmon are also a valuable export, and they, too, suffered drops in annual value of up to 20,000,000 kroner in the later 1960s.

Hunting and fishing

Animal husbandry is marginally possible in the milder southwestern portion of Greenland, where two-thirds of a total of 20,000 sheep perished in the severe winter of 1966–67 through a combination of hostile environment and insufficient fodder. Reindeer, first introduced from Norway in 1953, numbered about 8,000 by 1970.

Mining operations have been erratic: deposits may be rich, but extraction and transportation costs have often been high. The Ivigtut cryolite mine, opened in 1864, closed down its production in 1963, when it proved no longer economically feasible to operate what had been the world's largest natural deposit of this mineral. The similarly unprofitable coal-mining operations on Disko Island also became increasingly marginal by the 1970s. The zinc and lead mines at Mesters Vig, opened 1956, were exhausted by 1962. On the other hand, extensive explorations by international companies, utilizing the state licenses available in the 1950s and '60s, have discovered considerable potential in terms of uranium, molybdenum, zinc, lead, and other important deposits, and—given the possibility of improved extractive technology—the mineral prospects of Greenland remained to the economic forefront at the start of the 1970s.

Mining

Greenland trade has continued to fall largely into government hands, in spite of the ending of the official state trading monopoly in 1951. The Royal Greenland Trade Department (Den Kongelige Grønlandske Handel; KGH) continues to supply the country with necessary consumer and other goods, helping to overcome excessive transport costs. By the 1970s total sales amounted to well over 200,000,000 kroner, a substantial portion of total imports. Industrial expansion and economic development have caused a great excess of imports over exports (see Table 2). Nearly half of the local economic enterprises are conducted through the Greenland Technical Organization (Grønlands Tekniske Organisation; GTO), which hands its small factories, workshops, and other enterprises over to local, private ownership after setting them up. A good proportion are now in Greenlander, as opposed to Danish, ownership.

Table 2: Trade, 1960 and 1970
(000,000 kroner)

imports	1960	1970	exports	1960	1970
Consumer goods, clothing	32	117	Fish products	24	83
Fuel	12	29	Sheep and reindeer products	1	1
Building materials	14	59	Fur	2	7
Machinery, etc.	19	71	Cryolite	18	7
Transportation	3	18	Ore	13	0
Miscellaneous	28	103	Miscellaneous	1	7
Total	108	396*	Total	58	105

*Figures do not add to total given because of rounding.

TRANSPORTATION

By the early 1970s, the transportation system of Greenland, both internal and external, had undergone a complete transformation linked with the development of a market economy and with technological improvements, notably in air communications. Freight continues to be largely shipped by sea, while passenger traffic is usually airborne. The Denmark–Greenland connection is open all the year round, both by ship and by plane. Shipping connects freight fish products from, and oil, coal, and salt to, Greenland, while passenger vessels ply the west coast. The airport at Søndre Strømfjord, in addition to its function as a stopover on transpolar flights to North America and Asia, is the focus of flights into Greenland from Europe. Mail and passengers are then transported to their subsidiary destinations by the use of helicopters, although in the wilder areas of the northwest coast and east Greenland, dog-drawn sledges are of fundamental importance, with about 10,000 dogs still in use. At the start of the 1970s, over 13,000 passengers were transported between Denmark and Greenland, 12,000 of whom went by air. Along the west coast, 33,000 passengers were carried annually by ship and about 23,000 by air. Mail is carried by the governmental postal service, while the vast size of the island has led to the establishment of an ultramodern telephone, telegram, Telex, and radiotelephone network, with more than 90 separate stations, quite apart from the sophisticated military network associated with NATO and the North American defense system.

ADMINISTRATION AND SOCIAL CONDITIONS

Links with Denmark

Greenland, whose inhabitants enjoy equal political rights with Danish citizens, is administered from the Greenland ministry in Copenhagen, represented by a governor (*landshøvdingen*). There are three administrative subdivisions. Overall planning is the responsibility of the Greenland Provincial Council (Grønlandsrådet), made up of members of the National Council (Landsrådet) and the Danish Parliament, which makes recommendations

to the Minister for Greenland. The National Council administers public finances, which include about 50,000,000 kroner annually from alcohol and tobacco duties. Bills with reference to Greenland must be submitted to the council before their introduction in Parliament. The president of the council is now elected from its members, a significant move toward local autonomy made in 1966; before that date the post was held by the governor.

Greenland elects two members of the Danish Parliament, the electorate consisting of all Danish subjects (*i.e.*, Greenlanders and Danes) over the age of 21 who are resident in Greenland. The 16 members of the National Council each represent separate districts and voting for these posts, as in the case of the smaller local community councils, is restricted to those who have voted in parliamentary elections. With the growth of population, separate political parties are beginning to emerge.

Justice. The High Court (Landsret) consists of a judge, a graduate of the faculty of law of the University of Copenhagen, and two lay judges. The 16 district courts are headed by lay judges, from whom appeal can be made to the High Court, and, ultimately, to the Supreme Court in Copenhagen. There are 14 police districts headed by a chief constable. Although the changing patterns of social life have introduced a rising crime rate, with alcoholism a problem, there were no actual prisons, apart from temporary detention facilities, in Greenland by the early 1970s. Fishing inspection vessels are based at Grønnedal.

Education. The educational system consists basically of seven-year elementary schools, with provision for kindergarten classes. In addition, there are some high school facilities. The system is headed by the governor, the head of the church (most Greenlanders are Evangelical Lutherans), two National Council members, and a professional director. There are 17 school districts. The number of pupils in the Danish-financed, but largely Greenlander-staffed, system doubled over the decade of the 1960s, and the pressure of a growing and youthful population necessitated travel to Denmark for advanced studies. A state technical college was opened in 1970.

Health and welfare. The health service is state supported and without charge, utilizing Danish-trained staff in 17 districts, each with a doctor and a small hospital. The central hospital is located in Godthåb. Tuberculosis, responsible for 31 percent of all deaths in 1953, caused only three deaths 15 years later, at the end of an intensive campaign to eliminate the disease. The welfare service, which provides a wide range of assistance, was expanded in 1968, with the state underwriting 30 percent of welfare expenses and the National Council the remaining 70 percent. Social services expenses achieved the high rate of 1,000 kroner per person in the early 1970s. Although the developing market economy has stimulated expansion, tripling the standard of living in the period from 1950 to 1970, the standard of living remains at half of Denmark's. An increasing tax burden has appeared.

CULTURAL LIFE AND INSTITUTIONS

The cultural heritage of the Greenlanders finds most prominent expression in the Eskimo material culture: kayaks, umiaks (large boats paddled by women), sledges, harpoons, and soapstone lamps, all of which have a close relationship to the environment. Cultural values, in the wider sense, arose from the design concepts developed during the construction of these artifacts and found expression in sculptures (some made for the commercial market) in the mediums of ivory, wood, and soapstone.

Crisis in cultural identity

In Eskimo legends of today a relationship to the past and to other Eskimo groups around the Arctic is maintained. Modern poetry attempts to revive old traditions but points to a break between Greenlandic and Danish culture, with the Greenlander feeling the danger of losing his cultural identity in the developing market economy and modern European values. A folk high school, opened in 1963, attempts to support Greenlandic traditions.

Press and broadcasting. The press of Greenland dates back to the 1860s, when the newspaper *Atuagagdliutit* (literally "Something to Read") made an appearance in Godthåb. It is still published, in Greenlandic and Danish, as a fortnightly publication, contrasting with its erratic appearances of a century ago. There is also an embryonic local press, a product of the recent movement to the towns. Broadcasting, which started in World War II, expanded considerably with the building of a new station in 1958. This provides a wide range of programming, under Greenlandic management. In the early 1970s, the introduction of television was under consideration.

Problems and prospects. Although visits between Greenlanders and Canadian Eskimo have increased in recent decades, there is, as yet, no evidence of a pan-Eskimo movement linked to the resurgence of ethnic consciousness among North American Indians. The industrialization of fishing activities, started in the 1950s, has produced radical changes in Greenlandic society and a large population growth. The huge investments made in this program were threatened, in the late 1960s, by the departure of cod for warmer waters following a slight drop in seawater temperature. A return of seal hunting might be expected, and hopes for a developing tourist industry—perhaps linked to a conservational theme—could cushion potential blows to the local economy. It appeared, at least in the early 1970s, that long-term mineral development offered the only path for sustained economic growth. If that should fail, a large-scale migration to Europe might develop, a trend already seen in the exodus of part of a very youthful population. If this should occur, educational preparation for the transition to Danish culture and society would appear an unsatisfactory, but necessary, development.

BIBLIOGRAPHY. DANISH COMMISSION FOR THE DIRECTION OF THE GEOLOGICAL AND GEOGRAPHICAL INVESTIGATIONS IN GREENLAND, *Meddelelser om Grønland*, 189 vol., containing many papers in English, German, Danish, French, and Greenland Eskimo (1879–), is the largest existing work about Greenland, written by prominent specialists, and mostly covering geological, geographical, and ethnographic subjects. *Greenland*, 3 vol. (1928), although quite old, remains a useful source of information, notably in matters concerning the physical environment and ethnography; KAJ BIRKET-SMITH, *The Eskimos*, 2nd ed. (Eng. trans. 1959), the most authoritative work on the Eskimos and their culture; PETER FREUCHEN, *Book of Eskimos* (Eng. trans. 1961), a good, entertaining account by a famous Danish explorer; VILHJALMUR STEFANSSON, *Greenland* (1942), a readable work on Eskimo culture by a famous American explorer, now mainly of historic interest; SVEND KLITGAARD, *Greenland* (Eng. trans. 1970), an up-to-date extensive volume covering all aspects of Greenland life. Current information may be found in the *Danish Foreign Office Journal*, and in publications of the Danish Ministry of Foreign Affairs.

(M.Ga.)

Greenland Sea

A major outlying sea of the Arctic Basin, the Greenland Sea lies to the east of Greenland, between latitude 66° and 80° north. It covers 465,000 square miles (1,205,000 square kilometres)—an area almost as large as Tibet—and is of considerable scientific interest, although its inhospitable environment has severely limited its utilization by man. A line linking northeast Iceland; the isolated island of Jan Mayen; and Bjørnøya (Bear Island), lying halfway between the northern tip of Scandinavia and the great archipelago of Svalbard forms the conventional borders between the Arctic Greenland Sea and the more moderate environment of the Norwegian Sea to the southeast. The line so drawn also serves to mark off the course of significant underwater ridges and also to indicate the average edge of Arctic ice. The Greenland Sea is also linked to the North Atlantic via Denmark Strait (between Greenland and Iceland); with the Arctic Basin via the strait between the northern extremities of Greenland and Spitsbergen; and with the Barents Sea through the strait between Bjørnøya and Spitsbergen. Average depth in the sea is 4,750 feet (1,450 metres), with the deepest recorded point at 16,000 feet (4,800 metres).

Scientific importance

The first scientific investigations of the region were carried out from the Norwegian vessel "Vøringen" in 1876–78. Norwegian, Icelandic, and Soviet vessels, among others, have since carried scientific expeditions into the area. In 1909 Fridtjof Nansen, the famous ex-

plorer, had a hand in detailing the complex current system of the region, and his scheme has been brought up to date by a number of Soviet studies. Publications by scholars of a number of nations now make it possible to obtain a reasonably clear scientific picture of the Greenland Sea (for related information see ARCTIC OCEAN; ATLANTIC OCEAN; GREENLAND).

Physical characteristics. The bed of the Greenland Sea deepens irregularly northward and is divided by the submarine Mohns Ridge into the Greenland Basin (Hollow) and, to the south, the North Icelandic Deep. Ridge structures greatly influence major current directions. Bottom sediments, dispersed by currents and ice, are of nonorganic terrestial origin. Silts fill the submarine hollows and gorges; silty sands, gravel, boulders, and other products of erosion coat the shelves and ridges.

The sub-Arctic climate of the region causes a considerable accumulation of ice, and bitter north and northeast winds prevail throughout the year, cooling the sea surface and driving the cold waters southward. Absolute minimum air temperatures reach as low as −57° F (−49° C) off Spitsbergen, while the absolute maximum temperatures reach 77° F (25° C) off Greenland. Averages, however, are 14° F (−10° C) in the south and −15° F (−26° C) in the north, for February, the coldest month; August, the warmest month, experiences averages of 41° F (5° C) in the south and 32° F (0° C) in the north. The number of frosty days rises northward from 225 to 334. Precipitation totals 10 inches (250 millimetres) annually in the north and double this amount in the south. Surface water temperatures range from 30° F (−1° C) in the north, in February, to 43° F (6° C) in the south, in August. Fogs are frequent.

The hydrological system of the region is determined by the East Greenland Current, which brings ice down from the north and birfurcates when it approaches the central ridge. Branches of warm Atlantic currents, notably the West Spitsbergen branch of the Norway Current, also have some effect, resulting in a northward retreat of floating ice and in the formation of a "whale bay" in the ice cover beyond the sea's northern limit. The ice season lasts from October to the following August, and the ice includes Arctic pack ice (several yards thick), sea ice (about a yard thick), and freshwater ice in the form of towering icebergs. In June, as the bay ice breaks up along the coasts, icebergs that have split off from coastal glaciers are released to float out into the Atlantic via Denmark Strait. Tides in the Greenland Sea increase in amplitude near the coast and, in combination with the complex current system, break up the ice sheet and cause a mixing of the various water layers.

Biological characteristics. Because of the rather large amount of dissolved nutrient salts in its waters, the Greenland Sea is quite densely inhabited by lower life forms, which serve as the base of the food chain. Large invertebrates, fishes, birds, and mammals all feed on the smaller invertebrates and small organisms. Fish include cod, herring, redfish, halibut, and plaice; and mammals include seals, whales, and dolphins. Mosses, lichens, and scanty bushes are found around the coasts: they support the few deer and bull musks. Coastal waters are richer in fauna, including many types of gulls and ducks.

The role of nutrient salts

Fishing is weakly developed: with the exception of narrow zones off Iceland and Spitsbergen, which are fished by a number of European nations, the fish supplies of the Greenland Sea are only utilized by the local population.

BIBLIOGRAPHY. The METEOROLOGICAL OFFICE, *Monthly Meteorological Charts and Sea Surface Current Chart of the Greenland and Barents Sea*, 2nd ed. (HMSO, 1959), is recommended as a textbook for studying the meteorological and hydrological regime of the Greenland Sea. See also K. AAGAARD, "Temperature Variations in the Greenland Sea Deep-Waters," *Deep-Sea Res.*, 15:281–296 (1968); K. AAGAARD and I.K. COACHMAN, *The East Greenland Current North of Denmark Strait*, pt. 1, pp. 181–200 (1968); LAUGE KOCH, "The East Greenland Ice," *Meddelelser om Grønland*, vol. 130, no. 3 (1945); and A.F. LAKTIONOV, V.A. SHAMONTEV, and A.V. YANES, *An Oceanographic Sketch of the Northern Part of the Greenland Sea* (Eng. trans. 1963).

(M.M.A.)

Gregory I, Saint and Pope

Considered by many to be the architect of the medieval papacy, Gregory I the Great was known in his own time as a people's pope who was canonized (proclaimed a saint) by popular acclaim. His administrative, social, liturgical, and moral reforms became examples for those who followed him as bishop of Rome and pope of the medieval Roman Catholic Church.

Early life and career. Born in Rome about 540, Gregory was the son of a Gordianus and Silvia, who may have belonged to the eminent patrician family of the gens Anicia. His great-grandfather was Pope Felix III (483–492), and Pope Agapetus I (535–536) also may have been related to him. During his early years in Rome, the Lombards threatened and then invaded Italy (568). In about 572 Gregory, at the age of 32, became *praefectus urbis* (urban prefect; *i.e.*, the administrative president of Rome). Political and social conditions apparently caused him to relinquish this highest civilian office only two years later. Having a great interest in monasticism, Gregory converted the palace at Caelian Hill, which he had inherited as part of a large paternal fortune, into St. Andrew's Monastery, but he did not become its abbot. He then utilized his entire estate for the establishment of six additional monasteries on his other holdings in Sicily. Pope Benedict I (reigned 575–579) assigned him a diaconate in Rome, and in 579 Pope Pelagius II (reigned 579–590) sent him to Constantinople (the capital of the Byzantine Empire) as a papal nuncio, or representative, the curia's (papal administrative office) only foreign post. Gregory probably served there under Emperor Tiberius II (reigned 578–582) and Emperor Maurice (reigned 582–602) until 584, on the whole without much success in securing aid for Rome against the Lombards, who were also in a state of war in Italy with Byzantium.

Interest in monasticism

Election to the papacy. After sincere efforts to evade his election to the papacy Gregory was elected in 590 to that highest ecclesiastical position in the West. He complained in letters that he had been forced to assume the office. He determined to be a pope for the people, and he immediately devoted himself to alleviating the misery of the populace and of the refugees, including 3,000 nuns alone who had fled from the Lombards. Gregory I had grain sent from Sicily and used the revenues from church property to aid those who were starving and living in severe poverty.

He centralized the entire papal administration and vigorously opposed the graft and negligence of those in positions of responsibility, who, according to his view, administered the property of the poor and therefore were obligated to live up to the norms of absolute justice. The corrupt Byzantine officials had to be kept in check with gifts. Gregory became the first pope especially known for his devotion to social concerns, a devotion succinctly stated in one of his letters (Epistle I:44): "We do not want the treasury of the church defiled by disreputable gain." Gregory increased his diversified assistance, aides, and economic advisors in view of the devastation and barbarism of the Lombards.

Centralization of the papacy

The Pope attempted to reform and save the church in Italy, which was endangered spiritually as well as materially. He began his attempt by slowly catholicizing, in spite of their external Arianism (a heresy that denied the essential unity of God the Father and God the Son), the uncivilized Lombards. He did not want to see them destroyed but rather won for the kingdom of God, without breaking with Byzantium. He protested against the oppressive fiscal policies of the Byzantine exchequer, which so harshly taxed the people that they sometimes had to sell their children or emigrate into areas controlled by the Lombards. The Lombards, in turn, so extorted the Pope on their behalf that he called himself the "paymaster of the city."

Romanus, the Byzantine governor of Ravenna, who wanted war instead of the proposed peace of the Pope, ignored the Lombard King Agilulf's (reigned 590–616) stipulations for peace. He acted more badly toward Greg-

ory than toward the Lombards and agitated against him before the emperor Maurice. The letters of Gregory during the Lombard danger, citing the intrigues of Romanus and the accusations of the Emperor, provide a vivid and illuminating interpretation of the history of the time as well as an insight into the character of the Pope. Not until 598 did a temporary peace result in Italy.

Relations with Byzantium. In 602 Phocas, a Thracian centurian in the imperial army, during a period of disorder, managed to get himself elected emperor. He had Emperor Maurice, Empress Constance, the couple's five sons—the oldest was the godson of the Pope—and three daughters all executed. Knowing how to make use of the existing conditions of social and political disorder, Phocas, later known as a hated tyrant, also knew how to manipulate people. He made the Pope sympathetic toward both himself and his Lombard policy. Gregory's greeting and approval of one of the most hated tyrants of that century remains a blemish on his otherwise saintly character. Phocas, who was thus able to act with increasing terror, sought support from Gregory, whose blessing was tantamount to absolution for all offences. This action on the part of the head of the highest moral court of Europe established a precedent that was followed by many popes. The Byzantine Phocas, however, made peace with the Lombards, and thus peace for Italy in its relations with the Lombards was not secured by Gregory. The realization that a peace purchased at the price of agreement with Phocas also would incur negative consequences had not occurred to Gregory.

Phocas recognized the papal primacy of jurisdiction in the church and gave Gregory the impression of subordination. The Roman papacy had always valued such an attitude and in doing so overlooked other matters, including even the character or those with whom it came to terms. Gregory was deceived by Phocas, who conferred on him, rather than on John IV (the Faster), the patriarch of Constantinople, the disputed title of "ecumenical patriarch." The deposed and executed emperor Maurice, a devout humane ruler, had not previously granted the sought after title to the patriarch of Constantinople. The patriarch John, therefore, conferred this title on himself, as had other patriarchs before him, a practice that Pope Pelagius II had previously disputed. Gregory in 595 protested against this designation out of his conviction regarding the primacy of the pope. Instead, Gregory conferred on himself the title "servant of God's servants," a title borrowed from St. Augustine, which in its far too great humility meant, in effect, the opposite.

A reign of anarchy under Phocas spelled the end of the late Roman era. Gregory, with foresight, clearly recognized the approaching importance of the migrating peoples of the West, who were hardly or not at all christianized, and that the future of the church of the West lay with them. The visionary ideals of his conception in practice, however, would be to bring the barbarian powers of the West under the political sovereignty of Byzantium in the sense of a united Christian world under the ecclesiastical authority of Rome. He intensified his influential connections with Theodolinda, the Catholic Bavarian wife of the Lombard king Agilulf, whose son Adaloald only became Catholic in 615, and with Brunhilda, the powerful queen of Merovingian with whom he dealt just as submissively as he had with Phocas.

Missions. In 596, under the protection of Brunhilda, he initiated one of the greatest acts of his pontificate, the establishment of missions in England. His decision to do so may have emanated from his apprehension that the highly spiritual Irish–Scottish monasticism, which was strongly influenced by the Eastern Church and had not joined Rome, might finally take possession of the mission to England. He appointed Augustine (later St. Augustine of Canterbury) abbot of his St. Andrew's Monastery and later the first archbishop of Canterbury—as well as 40 monks—to begin the work in England. In contrast to other regions, Gregory had much regard for the pagan mentalities and customs in England, to which Augustine seldom adhered. The later English missionary monks St. Willibrord (658–739) and St. Boniface (*c.* 672/673–754) were able to conduct their missionary campaigns on the European continent because of the efforts of Gregory in regard to England.

Gregory, however, thought about missions in terms that were not always consistent with the monastic ideal of conversion by peaceful persuasion. He sometimes advocated a war of aggression against heathens in order to christianize them. His letter to Gennadius, the Byzantine governor from Africa, with the demand "to wage numerous wars"—in complete opposition to his peace efforts in Italy—in order to convert the subjugated to Christianity, can be viewed as the earliest conception of a crusade, a "holy war" differing from the spiritual battles of missionary activities. Gregory became, according to some misrepresentations, the model for the warring Pope Gregory VII (*c.* 1025–1085) as well as for Anselm of Lucca (Pope Alexander II) and Bonizone of Sutri, the well-known war theorists and contemporaries of Gregory VII. The earliest war benediction orginated with Gregory; he has become, along with St. Augustine (354–430), a precedent setter for the ecclesiastical war ideology of the Middle Ages. He admonished Brunhilda to prevent pagan sacrifices by means of armed forces.

Relations with the Jews

In regard to the Jews, to whom he offered seemingly economic advantages at conversion, the Pope was essentially tolerant. Had forced conversions been successful, however—such a policy was practiced in Spain by King Recared (died 601), who shortly beforehand (587) had become Catholic, and by the great church leader and adversary of the Jews St. Isidore of Seville (*c.* 560–636)—it is probable that Gregory would have agreed to such a policy.

Other concerns. Gregory did not take any definite action against the slave trade. He bought and sold slaves but would also free them; and sometimes he threatened others with severe punishments for their mistreatment.

He advocated rigid asceticism, and—perhaps because of this—he suffered from a stomach disorder and then from arthritis, which caused him to become almost totally lame in one of the last two years of his life.

He was an organizer, missionary, and manager, but not a politician. With the consolidation of the patrimony of Peter (lands controlled by the papacy), Gregory, without realizing it himself, became the founder of the later Papal States and of the temporal papal authority. According to his view, the patrimony of Peter ought to be at the immediate disposal of the church and of the poor. The view that this state would at one time serve the authoritative demands of the popes and would result in wars conducted by popes for augmentation of their imperialistic policies was inconsistent with his concept of the papacy's role in temporal affairs. He understood his period of rule as one of irrevocable service, as charity over the authority. His epitaph bears his policy's most suitable distinguishing mark: God's Consul.

Gregory did not comprehend that rulers and nations were incapable of following his conception of a *societas reipublicae Christianae* (a society of a Christian republic), which was formalized later in the Middle Ages. He was completely dependent upon the teaching (especially the concept of *The City of God*) of St. Augustine but not, however, predestined to speculative theology. The Pope —in whose views and actions are found the first attempts to subjugate secular authority to ecclesiastic authority and to elevate the priest to an extremely high status—exhibited a strange mixture of withdrawal from the world and energy, idealism and realism, melancholy and trust in God, and otherworldliness and the desire for power.

As a monk, which he always remained, he naturally had the expansion of monasticism especially at heart. Through him the Benedictine monastic principle attained broader support and results. Because of his concern for people, he tried to make their faith more intelligible to themselves by popularizing miracles and the concept of purgatory, as well as by encouraging a reform of the mass—from which came the Gregorian chant. His numerous writings—including his letters—possess little originality, but his *Regulae pastoralis liber*

("Book of Rules for Pastors") became a spiritual and practical guide to medieval bishops. The *Moralia in Job*, a textbook on moral theology and biblical interpretation, also exerted much influence in succeeding centuries. His ecclesiastical training was not extensive; he rejected culture and art as characteristic values; he treated the pre-Christian spiritual life with hostility. Estimations of his character oscillate in history; and he has undergone highly contrary evaluation, ranging all the way from ecclesiastical adulation to sharp criticism.

Gregory died in 604; his body lies buried in St. Peter's basilica in Rome. He had forbidden veneration of his corpse under penalty of excommunication. Since the 8th century he has been honoured as a doctor (teacher) of the church. His festivals of commemoration are March 12 (the date of his death) and September 3.

BIBLIOGRAPHY. The writings of Gregory I are contained in J.P. MIGNE, *Patrologia Latina*, vol. 75–79 (1844); and in the *Bibliothek der Kirchenväter*, 2 vol. (1873–74). Commentaries on his writing and thought may be found in: J.P. MCCLAIN, *The Doctrine of Heaven in the Writings of Saint Gregory the Great* (1956); H. DAVIS (ed. and trans.), "Pastoral Care" in *Ancient Christian Writers*, vol. 11 (1950); H. DELEHAYE, *Servis Servorum Dei* (1923); C. ERDMANN, *Die Entstehung des Kreuzzugsgedankens* (1935, reprinted 1965); K. GAMBER (ed.), *Sacramentarium Gregorianum: Das Stationsmessbuch des Papstes Gregor* (1966); *Morals on the Book of Job* in *Library of the Fathers*, 4 vol. (1844–50); R. RUDMANN, *Mönchtum und kirchlicher Dienst in den Schriften Gregors des Grossen* (1956); N. SHARKEY, *Saint Gregory the Great's Concept of Papal Power* (1950); and L.M. WEBER, *Hauptfragen der Moraltheologie Gregors des Grossen* (1947). Biographical literature includes: P. BATIFFOL, *Saint Gregory the Great* (Eng. trans. 1929); C. CHAZOTTES, *Grégoire le Grand* (1958); F.H. DUDDEN, *Gregory the Great: His Place in History and Thought*, 2 vol. (1905); and W. STUHLFATH, *Gregor I der Grosse* (1913), on his life until his election as pope.

(H.Ku.)

Gregory VII, Saint and Pope

Pope Gregory VII, though now known primarily for his role in church–state relations because of the Investiture Controversy (see below), was first and foremost, in his own correct estimation, a pope devoted to ecclesiastical and spiritual reforms. He was, indeed, one of the greatest and most successful reform popes of the Middle Ages.

Early career

Known as Hildebrand before he became pope, Gregory VII was born near Soana in Tuscany (Italy) about 1020 of a workingman's family. He went to Rome at an early age and began his education at the Monastery of St. Mary on the Aventine Hill, where his uncle was abbot. He apparently became a monk but continued his education at the Schola Cantorum (School of Musicians) in the Lateran Palace. This was a school for clergy and, perhaps, for laymen also, since Gregory mentions that two Roman nobles were educated with him. One of his teachers there, John Gratian, later became Pope Gregory VI (reigned 1045–46). Gregory took Hildebrand into his service and when he was deposed by Emperor Henry III (1017–56) at the Council of Sutri in 1046, Hildebrand went with his fallen patron into exile in Germany.

In Germany Hildebrand found favour with Emperor Henry III and was called back to Rome by Pope Leo IX (reigned 1049–54). He formed one of the groups of reformers that Leo IX was assembling, a group that was to exert a profound influence on the 11th-century church. Hildebrand became the "man behind the throne" during the pontificate of his immediate predecessor, Pope Alexander II (1061–73), having already been an important member of the Roman reform group. He became a cardinal and archdeacon of Rome and was able to satisfy his monastic inclinations by reforming the famous Monastery of St. Paul. He demonstrated his love of people by curbing the activities of the petty nobles who had caused excessive disorder in Rome and the neighbourhood. He also served on several important papal legations. Perhaps the most important of these was his legation to the Synod of Tours at which Berengar professed his faith in the Real Presence of Jesus in the Eucharist.

Elected by acclamation (April 22, 1073) to succeed Alexander II, Hildebrand took the name of Gregory VII. He was consecrated in St. Peter's basilica on June 30, 1073.



Gregory VII, after his expulsion from Rome, lays a ban of excommunication on the clergy "together with the raging king"; drawing from the chronicle of Otto von Freising (c. 1111–58). In the library of the University of Jena, Germany.

Leonard von Matt—EB Inc.

Election as pope

The keynote of Gregory's pontificate was reform and renewal of the church. To understand Gregory's personality and influence it is necessary to realize how deeply he was committed to the spiritual values of his age. From the beginning of his career he was not successful as a politician or a statesman; his specialty was spiritual leadership.

Gregory tried to restrain the marauding Normans of France in their conquest of south Italy (*c.* 1030–71) and to defend the Papal States, but found it difficult to subdue these hard fighting and acquisitive Frenchmen. Deeply interested in healing the still-young schism that had occurred between the Western and Eastern churches in 1054, he tried to encourage the European states to embark on a crusade to help Constantinople and the Eastern Christians, but in this he failed.

As a spiritual leader he was more successful even though he faced a task of awesome proportions. The efforts aimed at ecclesiastical reform by his predecessors, the attempts of the monks based at the Benedictine monastery at Cluny (France) to reform the church spiritually, and the preaching of reformers such as Peter Damian (1007–72) and Cardinal Humbert (*c.* 1000–61) were only partially successful. Gregory promptly began an attack on the chief problems of the church: simony (selling or purchasing ecclesiastical offices) and nicolaitism (clerical marriage or concubinage). He held a synod at Rome every Lent that decreed strong measures against the buyers and sellers of sacred offices and married clergy. He attempted to associate the bishops and the lay rulers with him in his effort to eliminate these problems. Since many bishops had purchased their positions and many also held very loose views of clerical celibacy, Gregory had his work cut out.

Because he found it difficult to work through the bishops, he tended to centralize authority. He used papal legates (representatives) freely and insisted on their precedence over local bishops. He sought uniformity in the Western Church and he discouraged the use of the Mozarabic liturgy (pre-Islāmic and Byzantine influenced rite) in Spain and the use of Slavonic language in the liturgy in Bohemia.

Investiture Controversy

Gregory is chiefly known for his contest with the German emperor Henry IV (1050–1106) over lay investiture (the right of lay rulers to grant ecclesiastical officials the symbols of their authority), a contest that he helped to

precipitate. Gregory's first concern was for reform, and he believed that secular rulers should support church authority in bringing it about. He had seen the beneficent results of the reform-minded emperor Henry III's (1017–56) interference, and he tried hard to work with young Henry IV. It was only when he lost confidence in Henry that Gregory began his attack on lay investiture.

The Pope's Roman Synod of 1075 struck hard at lay investiture and began the long conflict that was to go beyond Gregory's lifetime. At that synod Gregory excommunicated five of Henry's advisers. In late 1075 the situation deteriorated. Henry's defeat of the rebellious Saxons had increased his power and reduced his distractions. In Milan Erlembald, the leader of the Patarines, a lay reform group, was killed and the anti-reform party got the upper hand. Henry now openly showed his hand, gave support to the anti-reform party in Milan, and placed a new bishop in the position of the legitimate bishop, Atto (flourished 1085). He also appointed bishops to Spoleto and Fermo.

In 1075, while Gregory was saying Christmas mass in St. Mary Major, he was attacked, slightly wounded, and carried off by Cencius, a Roman noble. The Romans, who had much admiration for Gregory, rallied to his defense, attacked Cencius' stronghold and forced him to release the Pope, who went back to St. Mary Major to continue his mass. Gregory spared the life of Cencius.

Although Gregory had written to Henry in December 1075, holding out the possibility of negotiations on the issue of lay investiture, Henry gave no satisfaction to the legates that the Pope had sent to Germany. Indeed, he openly defied Gregory and with his bishops renounced obedience to Gregory and bade him step down from the papal throne. Supported by north Italian bishops, Henry sent the Roman Synod of 1076 a letter beginning: "Henry, King not by usurpation, but by the pious ordination of God to Hildebrand now not Pope but false monk."

Excommunication of Henry IV

The reading of such a document aroused indignation in the synod, and Gregory struck back hard. He and the synod excommunicated Henry, and the Pope declared him deposed. Gregory defended his actions against Henry in two letters to Bishop Hermann of Metz: the emperor is in the church and therefore he may be called to account by the pope. Gregory defended this position by arguments from Scripture, the Fathers, and history.

The excommunication had its effect. The number of Henry's partisans dwindled and the restless Saxons once more rose in arms. Plans were set on foot by the magnates to depose Henry and elect another king. Apparently, at the persuasion of Gregory's legates, a more moderate position was taken, though the terms drawn up by the magnates were severe enough. Henry was to leave the decision of his case to the Pope, who was to come to a meeting of the magnates at Augsburg on February 2, 1077. He was expected to repudiate his rebellion against the Pope and to urge his advisers who had been excommunicated to seek absolution. Thus was the stage set for a famous action at Canossa.

Early in 1077 Gregory went north to cross the Alps but found, instead of the guards the Germans had promised, the news that Henry was hastening to Italy. Alarmed, the Pope withdrew to the castle of Canossa, a stronghold of his faithful friend and supporter, Matilda (*c.* 1046–1115), countess of Tuscany. Henry, however, was coming not as a foe but as a suppliant. For three cold January days he stood outside the castle pleading for absolution while Matilda and St. Hugh, abbot of Cluny, added their pleas to his. Gregory was in a quandary. The nobles and bishops of Germany were awaiting his presence at Augsburg to discuss Henry's fate, and here was Henry in the cold begging piteously for absolution. The priest in Gregory prevailed over the politician, and the Pope absolved Henry from excommunication. It is to the Pope's credit as a spiritual leader that he absolved Henry, even though the action was disastrous to his own cause.

Henry promptly regarded himself as legitimate king again, and Gregory had to write somewhat apologetically to the German magnates explaining his action. The Germans cancelled the Augsburg meeting and called for another gathering at Forchheim on March 13. Gregory desired to attend this meeting, but apparently neither Henry nor the leader of the opposition, Rudolf of Swabia (died 1080), really desired the Pope's presence. Gregory, however, sent legates who pleaded with the assembled nobles and bishops not to proceed with an election until the Pope could be present. The magnates went ahead, however, and elected Rudolf of Swabia, thus precipitating a bloody civil war. Gregory tried to mediate between Henry and Rudolf. He recalled his legates, and when Henry imprisoned one of them, the other excommunicated Henry. To prevent the Pope from confirming this excommunication the King sent ambassadors to plead with the Pope. They succeeded and the Pope contented himself with calling for a great meeting to settle the quarrel. For two years, 1078–80, Gregory maintained a mediatior's position and was abused by both sides.

Later reign

By 1080 the Pope was convinced that Henry was intransigent and once more excommunicated him and declared him deposed. This meant war. Henry had the support of his faction in Germany and that of the Lombard (north Italian) anti-reform party. Gregory sought the aid of the formidable Robert Guiscard, duke of Apulia and Calabria (*c.* 1015–85). Henry's German bishops met at Brixen (Italy) and declared Gregory deposed. To replace him they chose Guibert, archbishop of Ravenna, who took the name Clement III (1080, 1084–1100).

The tide began to flow strongly in favour of Henry when Rudolf of Swabia was killed at the Battle of the Elster (1080). Henry, freed from pressure in Germany, came over the Alps, defeated the forces of Countess Matilda, and besieged Rome. Gregory renewed his excommunication of the King. He tried to stir up opposition to Henry in Germany by urging Welf I of Bavaria (died 1101) and the princes to hold an election to replace Rudolf, but this did not deter Henry from besieging Rome in 1081, 1082, and 1083. Still firm, Gregory held a synod at the Lateran in November 1083 to attempt a settlement, but Henry prevented some bishops from attending.

The fathers of the synod, very much aware of the menacing presence of Henry's soldiers across the Tiber, pleaded with Gregory not to renew his excommunication of Henry at this time, whereupon the Pope contented himself with a general excommunication of all who prevented attendance of the synod. All attempts at peace failed, and on March 21, 1084, Henry's troops took the city. Gregory sought refuge in the castle of St. Angelo and suffered the chagrin of seeing Guibert of Ravenna (now Clement III) crowned in St. Peter's. Guibert in turn crowned Henry emperor. Help, however, was on the way. Robert Guiscard, back from an unsuccessful attempt on the Byzantine Empire, marched on Rome and rescued the Pope. Gregory's safety was dearly bought, for in a fight between the Normans and the Romans a large part of the city was burned down. Gregory, now unpopular with the embittered Romans, left with Guiscard. He died at Salerno on May 25, 1085. A biographer placed on his dying lips the words, "I have loved justice and hated iniquity, therefore I die in exile."

In 1584 Gregory XIII beatified him. He was canonized by Paul V in 1606, and Benedict XIII extended his feast to the whole church in 1728. Gregory's feast is kept on May 25.

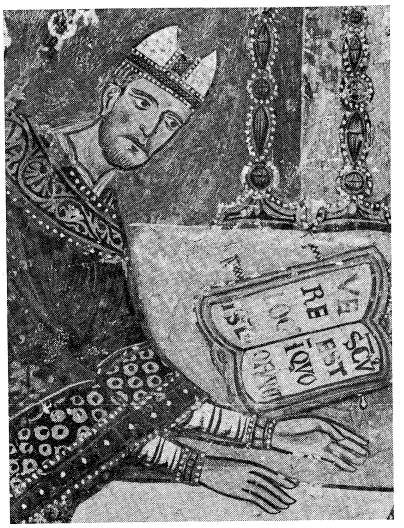
BIBLIOGRAPHY. A primary source is the correspondence of Gregory. His *registrum* and other letters may be found in ERICH CASPAR, *Das Register Gregors VII* (1955); J.P. MIGNE, *Patrologia Latina*, vol. 148 (1888); and selections from the *registrum* in EPHRAIM EMERTON, *The Correspondence of Pope Gregory VII* (1932). AUGUSTIN FLICHE, *La Réforme Grégorienne*, vol. 2 (1924), is devoted entirely to Gregory and has an excellent critical study of the sources. BRIAN TIERNEY, *The Crisis of Church and State, 1050–1300* (1964), is a perceptive account containing some of the key documents. See also WALTER ULLMAN, *The Growth of Papal Government in the Middle Ages* (1955); and SCHAFER WILLIAMS, *The Gregorian Epoch* (1964), an interesting and provocative selection of extracts from historians who disagree on Gregory.

(J.S.Br.)

Gregory IX, Pope

Canon lawyer, theologian, founder of the papal Inquisition, defender of papal prerogative, the irascible Gregory IX (reigned 1227–41) stands as one of the most vigorous popes of the 13th century—the century in which the power of the church reached its zenith.

Alinari



Gregory IX consecrating the chapel of St. Gregory, detail of a fresco, 13th century. In the Lower Church of Sacro Speco, Subiaco, Italy.

Gregory, a nephew of Pope Innocent III, was born prior to 1170 of the family of the counts of Segni and was baptized Ugo or Ugolino. He studied theology at the University of Paris, but his early ecclesiastical career marked him as a diplomat. Shortly after his creation as a cardinal-deacon by his uncle in 1198, he was involved in peace negotiations with Markwald of Anweiler in southern Italy. Twice before 1210 he served Innocent as a papal legate in Germany. In 1206 Innocent promoted him to the cardinal bishopric of Ostia, the port city of Rome. During the pontificate of Pope Honorius III (1216–27), Ugo continued to play a leading role. He enjoyed not only the support of the Pope but also that of the youthful emperor-elect, Frederick II of Hohenstaufen, king of Sicily, whose cause he had supported during the reign of Innocent III. Ugo was a deeply religious man, closely attuned to the great spiritual movements of his time. He was friend to both St. Dominic and St. Francis of Assisi, founders of the first mendicant orders. He served as cardinal-protector of the Franciscans and adviser to St. Clare of Assisi, the founder of the Poor Clares. Like his predecessors, Ugo firmly supported the crusading movement, and it was from his hands that Frederick II took the cross as a symbol of his intention to lead a crusade. Ugo was an austere man of decisive mind and somewhat harsh personality. Even those he loved and admired most sometimes felt the strength of his convictions and the force of his will. But there can be no doubt about his moral integrity and dedication to the church. Still, it was his quickness to anger and his impatience with opposition that marked the character of his pontificate.

When Ugo ascended the papal throne as successor to Honorius III on March 19, 1227, he had already lost patience with the moderate policies of his predecessor. In particular, he had grown increasingly disenchanted with Emperor Frederick II. Frederick's delays in embarking on his promised crusade and his efforts to hold both the imperial throne and the crown of Sicily aroused opposition to him in the Roman Curia. The rupture broke into the open shortly after Gregory's election, when Frederick, who had finally launched his crusade, was forced to return to Brindisi because of an outbreak of plague. Already suspicious of Frederick's sincerity, the Pope excommunicated him on September 29, 1227, and issued a pained and angry encyclical to justify his action. Frederick responded by an attack on the excommunication as unjustified and a denunciation of the Roman Curia.

Excommunication of Frederick II

Nevertheless, Frederick embarked for the East, where he conquered Cyprus and negotiated with the sultan of Egypt for Jerusalem. Gregory was incensed at Frederick's presumption in leading a crusade while under ban of excommunication. Claiming provocation by Frederick's vicar in the Kingdom of Sicily, Gregory raised an army and launched an attack on the kingdom. This war marked the end of the policy of negotiation. Though Frederick's return witnessed the defeat of the papal forces, the deep fears aroused by his policies remained unsettled by the Treaty of San Germano (1230). In 1231 Gregory sharply protested Frederick's issuance of the *Liber Augustalis*, or Constitutions of Melfi, a code of laws for the Kingdom of Sicily. Though there was little in these laws that was actually objectionable, their thrust in the direction of a strong monarchy contained a threat to the church.

During the early 1230s Gregory took advantage of the respite in his struggle with the Emperor to turn his attention more to the internal and spiritual problems of the church. He ordered the canonist Raymond of Peñafort to compile the *Decretals*, a code of canon law based both on conciliar decisions and on papal letters, which he promulgated in 1234. This collection, arranged topically in five books, remained as the fundamental source of ecclesiastical law for the Catholic Church until after World War I. He also entered into negotiations with the Greek Orthodox Church that resulted in a series of conferences at Nicaea in January 1234 but proved abortive. Gregory continued the policies of his predecessors against heresy in southern France and northern Italy. He strengthened the Inquisition and entrusted its operations to the Dominicans. One of these inquisitors, Bernardo Gui, wrote the principal contemporary biography of Gregory IX.

Church reforms

The truce between Gregory and Frederick II was severely strained in 1235 by imperial accusations that the Pope had been working with the Lombards of northern Italy to undermine imperial influence. While Gregory denied the charge, the work of the Dominicans among heretics in northern Italy, many of whom were leagued with Frederick's supporters, did provide a foundation for imperial fears. Frederick's invasion of Sardinia, a papal fief, on behalf of the candidacy of his son Enzio for the Sardinian crown, led to a renewal of the excommunication on March 20, 1239, and caused Gregory to seek supporters in northern Italy. The propaganda war that accompanied the renewed hostilities is noted more for vitriolic than for reasoned argumentation. Gregory accused Frederick of crimes against the church in the Kingdom of Sicily and labelled him a blasphemer. The effort to find a settlement between the secular and the spiritual powers of medieval society received a decisive blow in this struggle. No definitions of separate spheres of authority would ever again overcome the reality of the fears that dominated both the papal Curia and secular powers.

With Frederick's army invading the Papal States, Gregory summoned a general council of the church, which met in Rome on Easter Sunday 1241. The capture of a large number of prelates on their way to the council by Frederick's Pisan allies put an end to this project, at least during Gregory's pontificate. Gregory IX died soon after, on August 22, 1241, his work unfinished. He had attempted to carry on the work of Innocent III and was successful in many of his efforts. Historians have judged him harshly because of his conflict with Frederick II, but too often their judgments have turned on the defects of his personality rather than the objectives of his policy.

BIBLIOGRAPHY. The most valuable account is contained in A. FLICHE and V. MARTIN (eds.), *Histoire de l'église depuis les origines jusqu'à nos jours*, vol. 10, pp. 217–426 (1950), which should be supplemented by ERNST BREM, *Papst Gregor IX bis zum Beginn seines Pontifikats* (1911). As yet there is no standard biography except that of P. BALAN, *Storia di Gregori IX e dei suoi tempi*, 2 vol. in 1 (1872–73), H.K. MANN, *History of the Popes in the Early Middle Ages*, 2nd ed., vol. 13 (1925), is long out of date. Brief discussions of Gregory's con-

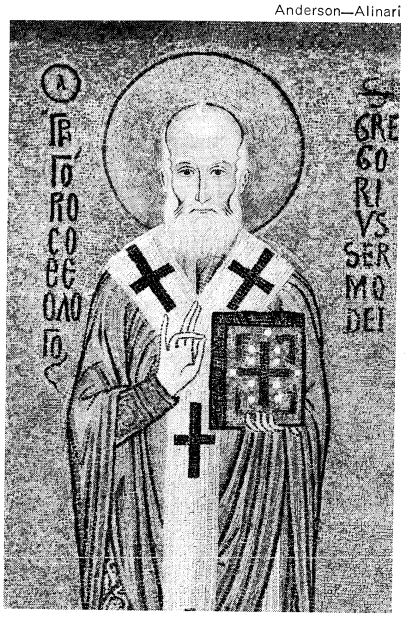
troversies with Frederick II may be found in J.M. POWELL, "Frederick II and the Church: A Revisionist View," *Catholic Historical Review,* 48:485–497 (1963); and (trans. and ed.), *The Liber Augustalis* (1971).

(J.M.Po.)

Gregory of Nazianzus, Saint

Among the Fathers of the Church, Gregory of Nazianzus is distinguished by the clear and vigorous exposition of the trinitarian theology (relating to God the Father, Son, and Holy Spirit) that he and his fellow Cappadocians, Basil the Great (*c.* 329–379) and Gregory of Nyssa (*c.* 335–*c.* 394), formulated; the three are called the Cappadocian Fathers. The dramatic episodes of Gregory's career in church and state give his life a certain romantic interest, of which he himself was not unaware.

Anderson—Alinari



St. Gregory of Nazianzus, detail of a mosaic in the Palatine Chapel, Palermo, Italy, 12th century.

Gregory's father, also named Gregory, was converted to the Christian faith from the monotheistic sect known as the Hypsistarii (Worshippers of the Most High) under the influence of his Christian wife. He was soon afterward consecrated bishop of his native city, Nazianzus (the exact location of which is not known; Cappadocia was in eastern Anatolia), by bishops on their way to the Council of Nicaea in 325. Born some years later (*c.* 330), the younger Gregory thus grew up in a Christian and clerical family. Nevertheless, he received a classical as well as religious education, studying first at Caesarea, the provincial capital; at least briefly at Alexandria; and finally at Athens (*c.* 351–356). He was a close friend of Basil, his fellow student and later bishop of Caesarea, and in his panegyric at Basil's death in 379 he gave a vivid picture of student life of the period. Among Gregory's other contemporaries as a student at Athens was the future emperor Julian, who in his brief two-year reign attempted to revive paganism. Soon after returning to Cappadocia, Gregory joined the monastic community that Basil had founded at Annesi in Pontus. During this time, in order to preserve the thought of the great Alexandrian theologian Origen, many of whose speculative views were under attack, the two friends collaborated in editing the *Philocalia,* an anthology of theological and devotional selections from the works of Origen.

Education

In 362 Gregory accepted ordination to the priesthood to assist his father, though he went to Annesi for further preparation and remained there until the following Easter. For the next 10 years he worked at Nazianzus supporting Basil—who was first presbyter and from 370 to 379 bishop of Caesarea—in his struggles with personal rivals, with Arians (who denied the divinity of Christ and were semi-Origenists), and with the Arian emperor Valens. Basil was attempting to retain control of the church in at least part of the new province of Cappadocia Secunda, which had been created by Valens to diminish orthodox authority. Gregory, under pressure from Basil to assist him in this conflict, reluctantly accepted consecration (372) to the episcopate for the village of Sasima. He never took possession of the bishopric, however, and withdrew with a sense of grievance against Basil for having presumed on their friendship. He briefly administered the church of Nazianzus again after his father's death in 374, but when a successor was installed in that bishopric he retired to a monastery in Isauria, in south central Anatolia.

The death of Valens in 378 at the Battle of Adrianople ended the imperial patronage of Arianism, and after Basil died on the following January 1, Gregory became the outstanding spokesman in Asia Minor of the Nicene party that accepted the decrees of the Council of Nicaea of 325. He was invited to take charge of the Nicene congregation at Constantinople, a city torn by sectarian strife. His Chapel of the Resurrection (Greek Anastasia) became the scene of the birth of Byzantine (from Byzantium, the earlier name of Constantinople) Orthodoxy —*i.e.,* the post-Nicene theology and practice of the majority of Eastern Christianity. Among the sermons he preached there, the *Five Theological Orations* are a striking presentation of trinitarian doctrine, and his memorial addresses and others on special occasions are important historical sources. Though Gregory wrote no commentaries, he was famous for his deep knowledge of Scripture; among his hearers at Constantinople was the biblical scholar Jerome, who gained a greater understanding of the Greek scriptures from Gregory. A religious adventurer, Maximus the Cynic, however, was set up as a rival to Gregory by bishops from Egypt, who broke into the Anastasia at night for a clandestine consecration.

Leading exponent of Nicene theology

When the new emperor, Theodosius, came east in 380, the Arian bishop of Constantinople, Demophilus, was expelled, and Gregory was able to lead his congregation to the Great Church (probably the earlier basilica on the site of the present-day Hagia Sophia). The council (later recognized as the second ecumenical council) that met at Constantinople in 381 was prepared to acknowledge Gregory as bishop of Constantinople; but on the arrival of Bishop Timothy of Alexandria, his position was challenged on technical grounds. Weary of disputes and intrigues, Gregory withdrew after an eloquent farewell discourse. The council, however, supported his policy, condemning old and new heresies, denying all validity to the consecration of Maximus, and forbidding bishops to interfere outside their own areas of authority (a step toward the system of patriarchates). It endorsed the trinitarian doctrine of three equal persons (Father, Son, and Holy Spirit) as taught by Gregory and expressed in the "creed commonly called the Nicene," which is still regarded as authoritative in East and West alike, including most leading Protestant churches.

For the rest of his life Gregory lived quietly on the family property at Arianzus near Nazianzus, except for a brief period as administrator of the church of Nazianzus during a vacancy. He continued his interest in church affairs through correspondence, even during one year when he took a vow of silence for Lent. He wrote to his successor, the amiable but ineffective Nectarius, and others against the heresy of Apollinaris, who denied the existence of a human soul in Christ. His writings of the period include a long autobiographical poem (commonly referred to as *Carmen de se ipso,* "Song Concerning Oneself") and many short poems, mostly on religious subjects. His preserved works include a number of sermons, not improperly called orations, and a large collection of letters. His death is dated about 389, according to a statement of Jerome.

Poetry

BIBLIOGRAPHY. The Benedictine edition of Gregory's works, vol. 1 (1778), and vol. 2 (1840; delay in publication caused by the French Revolution), is reprinted in J.P. MIGNE (ed.), *Patrologia Graeca,* vol. 35–38 (1857–62). *Five Theological Orations* were edited with a valuable commentary by A.J.

MASON (1899). Selected orations and letters were translated by C.G. BROWNE and J.E. SWALLOW in *Nicene and Post-Nicene Fathers*, ser. 2, vol. 7, pp. 185–498 (1894); the theological orations and letters on Apollinarianism are reprinted in *Library of Christian Classics*, vol. 3 (1954), with an introduction by E.R. HARDY. Book 8 of the *Greek Anthology* comprises 254 of Gregory's Epigrams, in *Loeb Classical Library*, ed. and trans. by W.R. PATON, vol. 2, pp. 400–505 (1917).

For biographical information, see CARL ULLMANN, *Gregorius von Nazianz* (1825, 2nd ed. 1867; Eng. trans. by G.V. COX, 1851), still valuable; and ROSEMARY RUETHER, *Gregory of Nazianzus: Rhetor and Philosopher* (1969). There are useful sketches by J.H. NEWMAN, who felt a strong kinship with Gregory, in *The Church of the Fathers* (1840), reprinted in *Essays and Sketches*, vol. 3 (1948); and DOROTHY BROOKE, *Pilgrims Were They All*, ch. 4 (1944). On Gregory's thought, see J.N.D. KELLY, *Early Christian Doctrines*, 3rd ed. (1965); and on the Council of 381 and the Creed, J.N.D. KELLY, *Early Christian Creeds*, 2nd ed., ch. 10, pp. 296–321 (1960).

(E.R.Ha.)

Gregory of Nyssa, Saint

As a philosophical theologian and mystic, Gregory of Nyssa completed the contributions to Christian thought of the other two Cappadocian Fathers, his brother Basil of Caesarea and their friend Gregory of Nazianzus. His ecclesiastical career was less successful than theirs, but Gregory's work as scholar and writer was creative, and in the 20th century it is being rescued from undeserved neglect.

Alinari



St. Gregory of Nyssa, detail of a 12th-century mosaic. In the Palatine Chapel, Palermo, Italy.

Early ecclesiastical career

A younger son of a distinguished family of Caesarea in Cappadocia, Gregory was born about 335 and educated in his native province but was more deeply influenced by his philosophical training than Basil and Gregory of Nazianzus. He began his adult life as a teacher of rhetoric and may have been married—although several references that suggest this are capable of a different interpretation, and the strictures on marriage in his treatise *On Virginity* seem to imply the contrary. In the 360s he turned to religious studies and Christian devotion, perhaps even to the monastic life, under Basil's inspiration and guidance. As part of Basil's struggle with Bishop Anthimus of Tyana—whose city became the metropolis (civil and therefore ecclesiastical capital) of western Cappadocia in 372—Gregory was consecrated as bishop of Nyssa, a small city in the new province of Cappadocia Secunda, which Basil wished to retain in his ecclesiastical jurisdiction. In 375, however, Gregory was accused of maladministration by the provincial governor as part of the Arianizing campaign of the Emperor Valens (an attempt to force the church to accept the views of the heretic Arius, who denied the divinity of Christ). He was deposed in 376 by a synod of bishops and banished. But on Valens' death in 378 his congregation welcomed him back enthusiastically.

Though Basil had considered Gregory unsuited for ecclesiastical diplomacy, after his return to his diocese he was active in the settlement of church affairs in the years that followed. In 379 he attended a council at Antioch and was sent on a special mission to the churches of Arabia (*i.e.*, Transjordan); his visit to Jerusalem on this occasion left him with a dislike for the increasingly fashionable pilgrimages, an opinion he expressed vigorously in one of his letters. In 381 he took part in the General (second ecumenical) Council at Constantinople and was recognized by the emperor Theodosius as one of the leaders of the orthodox communion in Cappadocia, along with Basil's successor at Caesarea. He declined election to the important bishopric of Sebaste, but the care of his small diocese left him free to preach at Constantinople on such special occasions as the funerals of Theodosius' wife and daughter. Under the unlearned Nectarius, the successor of Gregory of Nazianzus at Constantinople, Gregory of Nyssa was the leading orthodox theologian of the church in Asia Minor in the struggle against the Arians.

Contribution as a scholar

Gregory was primarily a scholar, whose chief contribution lay in his writings. Besides controversial replies to heretics, particularly the Arians—in which he formulated the doctrine of the Trinity (Father, Son, and Holy Spirit) that emerged as a clear and cogent answer to Arian questioning—he completed Basil's *Hexaëmeron* ("Six Days"), sermons on the days of the Creation, with *The Creation of Man*, and he produced a classic outline of orthodox theology in his *Great Catechesis* (or *Address on Religious Instruction*). The latter work is especially notable for developing systematically the place of the sacraments in the Christian view of restoration of the image of God in human nature—lost through sin in the fall of Adam. His brief treatise *On Not Three Gods* relates the Cappadocian Fathers' theology of three Persons in the Godhead (*i.e.*, the Trinity) to Plato's teachings of the One and the Many. As a Christian Platonist, Gregory followed the great Alexandrian theologian Origen (*c.* 185–*c.* 254), though not slavishly; most notably, he shared Origen's conviction that man's material nature is a result of the fall and also Origen's hope for ultimate universal salvation. In imitation of Plato's *Phaedo*, Gregory presented his teaching on resurrection in the form of a deathbed conversation with his sister, the Abbess Macrina.

Platonic and Christian inspiration combine in Gregory's ascetic and mystical writings, which have been influential in the devotional traditions of the Eastern Orthodox Church and (indirectly) of the Western Church. His *Life of Macrina* blends biography with instruction in the monastic life. *On Virginity* and other treatises on the ascetic life are crowned by the mystical *Life of Moses*, which treats the 13th-century BC journey of the Hebrews from Egypt to Mt. Sinai as a pattern of the progress of the soul through the temptations of the world to a vision of God. A notable emphasis of Gregory's teaching is the principle that the spiritual life is not one of static perfection but of constant progress. His greatest achievement is his remarkably balanced synthesis of Hellenic (Greek) and Christian traditions, in an age when both were represented by vigorous and acute minds.

Gregory did not, however, neglect his practical and pastoral duties, as is attested by his preserved letters and sermons. Many of the latter were written in praise of the saints venerated in Cappadocia or to celebrate the great days of the church year. Others, such as Gregory's attacks on usury and on the postponement of Baptism, deal with ethical problems of the church in his time. His more intimate discourses on the Lord's Prayer and the Beatitudes (Matt. 5:3–12) combine ethical and devotional interests, as does his commentary on the Song of Solomon. Gregory disliked attending gatherings of bishops but was periodically invited to preach at such occasions. His last public appearance was at a council at Constantinople in 394, and he probably died soon afterward.

BIBLIOGRAPHY. The unsatisfactory 17th-century edition of Gregory's works reprinted in J.P. MIGNE (ed.), *Patrologia Graeca*, vol. 44–46 (1858–62), is being replaced magnificently by W. JAEGER *et al.*, *Gregorii Nysseni Opera* (1952–). The *Life of Moses* has been edited by J. DANIELOU in *Sources Chrétiennes*, 3rd ed., vol. 1 (1968), with French translation; the *Catechetical Oration*, edited with commentary by J.H. SRAWLEY (1903). Treatises and letters were translated by W. MOORE and H.A. WILSON in *Nicene and Post-Nicene Fathers*, ser. 2, vol. 5 (1893); the *Address on Religious Instruction* and *On Not Three Gods* by C.C. RICHARDSON in *Library of Christian Classics*, vol. 3 (1954), with introduction and bibliography; the *Life of Macrina* by W.K.L. CLARKE (1916); *Sermons on the Lord's Prayer and the Beatitudes* by H.C. GRAEF in *Ancient Christian Writers*, vol. 18 (1954).

There is a good sketch and bibliography by J. QUASTEN, "Gregory of Nyssa," in *Patrology*, vol. 3, pp. 254–296 (1960). On Gregory's theology, see J.N.D. KELLY, *Early Christian Doctrines*, 3rd ed. (1965); on his philosophy and mysticism, J. DANIELOU, *Platonisme et théologie mystique: Essai sur la doctrine spirituelle de Saint Grégoire de Nysse*, 2nd ed. (1954), a basic work; selections from his mystical writings chosen by Danielou are in *From Glory to Glory*, trans. by H. MUSURILLO (1961). W. JAEGER, *Early Christianity and Greek Paideia*, pp. 75–102 (1962), is a valuable discussion of Gregory's place in cultural history.

(E.R.Ha.)

Gregory of Tours, Saint

Gregory of Tours, bishop, historian, and saint, provided the major source for knowledge of the 6th-century Franco-Roman kingdom in his *History of the Franks* and himself played a notable part in some of the events he recorded.

Gregory was born at Augustonetum (now Clermont-Ferrand, France) on November 30, 538/539, to an aristocratic family. On both sides his family supplied several of the chief bishops of what today is central France. He was educated at Arvernia in the bishop's household. In 573 he succeeded his cousin as bishop of Tours.

At that moment the Frankish kingdom (which included present Rhineland Germany) was divided into three kingdoms under the three surviving grandsons of the great Clovis, the founder of the Merovingian dynasty in western Europe: Guntram, ruling Burgundy and Provence; Chilperic, ruling Neustria, the western kingdom; and Sigebert, ruling Austrasia, the eastern kingdom, but with appendages at Tours and in the south. Just as Gregory became bishop, one of the fratricidal wars endemic among the family of Clovis broke out; Sigebert was murdered (575), and Chilperic seized Tours.

H. Roger-Viollet



St. Gregory of Tours, statue by Emmanuel Frémiet, 19th century. In the Panthéon, Paris.

Gregory as bishop

Gregory was bishop for ten years under the capricious and tyrannical Chilperic. Outside the city of Tours was the sanctuary of St. Martin, the revered 4th-century bishop of Tours, where famous criminals or persecuted persons sought shelter. Chilperic's own son Merovech sought shelter there from his father; and, in consequence of trying to protect Merovech and the right of sanctuary, Bishop Gregory became embroiled with his king and especially the queen, Fredegund. He was alternately threatened and cajoled with offers of bribes and had to defend himself formally, at the Council of Berny in 580, against the charge of spreading scandal that the Queen had committed adultery with the Bishop of Bordeaux. The council accepted Gregory's denial on oath and acquitted him. Gregory stood up to Chilperic to protect a fellow bishop and to resist the King's unorthodox opinions on the Trinity. "It would be no sensible man but a lunatic that would adopt views like yours," he told the man whom he once described as the Nero and Herod of his time. In 584 Chilperic was murdered, and Tours came under King Guntram of Burgundy. With Guntram, who possessed a certain degree of political ability, his relations were far happier. In 587 the Treaty of Andelot between Guntram and Sigebert's son Childebert II restored Tours to Austrasia, and Gregory came under Childebert. As bishop he restored the great church of St. Martin at Tours, dedicated a number of churches, protested successfully against excessive taxation in his diocese, and arbitrated in Frankish vendettas. His last years were disturbed by a riot that he had to quell in the nunnery of St. Radegunda at Poitiers. He died in 594 or 595 at Tours.

Writings of Gregory

Gregory wrote several groups of lives of saints, called the "Lives of the Fathers," and seven books of miracles. These afford unique evidence of the piety and social life of Merovingian France. In addition, he wrote a commentary on the Psalms (of which only fragments survive) and a treatise on church offices. But his fame rests on his *History of the Franks* (which he called simply the *History;* its complete title is not the original). It was written in three separate sections: the first (books 1–4) covered the period to the death of Sigebert in 575 and was written soon afterward; the second (books 5–6) was composed about 581–584; the third (books 7–10) was written about 590–591. The manuscripts show two texts, a longer one and a shorter one, and perhaps Gregory himself personally edited the additions or subtractions. In an age of dry annalists his history has only one Latin competitor, namely, the work of Bede, the Anglo-Saxon historian. In his pages France of the 6th century comes alive as nowhere else. Gregory's chronicle is an unforgettable portrait of the western kingdoms just after the breakup of the western Roman Empire.

Though he was a compulsive writer, he lamented his confusion over grammar, especially genders and prepositions, and thought that he wrote like a country bumpkin. He once asked his mother about the matter: she said that if he wrote in everyday language, everyone would understand. He tended, nevertheless, to drag in classical quotations when he could. He arranged his presentation by years but was never dull like the contemporary monastic chroniclers. Yet the story is at times disordered and confused. His genius as a writer was for graphic, fast-moving, blunt narrative, understated rather than rhetorical. He had no interest in personal details or the habits of society, and none of his characters (except himself) comes alive. His strength lay primarily in describing dramatic situations.

His philosophy of history (in so far as someone so unphilosophical could have one) saw the world as groaning and the church as fighting the battle to save humanity

from its bonds. He stated his aim thus: "to record the wars of kings with their enemies, of the martyrs with the heathen, and of the churches with the heretics." He looked for villainy to find its just punishment, and virtue its just reward, in this life as well as in the next. He thought an orthodox faith to be important and judged kings partly by whether or not they professed it. Gregory made no effort to paint the church of his day in rosy colours, describing warrior bishops or adulterous and drunken priests with equal impartiality. He generally used his authorities well, though he did not transcribe them well. In describing events of his own time (in books 4–10) he was most successful as a chronicler.

Gregory enjoyed the world, shared deeply in the superstitions of the age, travelled constantly and knew France as well as any man of his time, was affectionate to little children, and gave in his writings the unconscious impression of a likable, down-to-earth, courageous, not very spiritual, but Christian and humane, man.

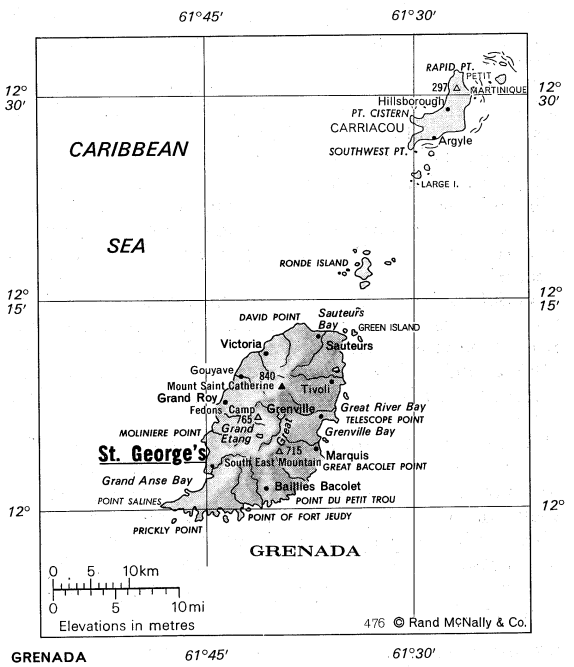
BIBLIOGRAPHY. Editions of Gregory's works may be found in the *Patrologiae cursus completus . . .*, vol. 71, and in B. KRUSCH and W. LEVISON, *Scriptores Rerum Merovingicarum*, vol. 1; rev. ed. by R. BUCHNER, 2 vol. (1955). An English translation of the histories, with a valuable introduction by O.M. DALTON, is *The History of the Franks, by Gregory of Tours*, 2 vol. (1927). For background information, see H.G.J. BECK, *The Pastoral Care of Souls in South-East France During the Sixth Century* (1950); S. DILL, *Roman Society in Gaul in the Merovingian Age* (1926); and J.M. WALLACE-HADRILL, "Gregory of Tours and Bede: Their Views on the Personal Qualities of Kings," in *Frühmittelalterliche Studien*, 2:31–44 (1968).

(W.O.C.)

Grenada

The island of Grenada, also known as the Isle of Spice, is the southernmost of the Windward Islands in the eastern Caribbean Sea about 100 miles (160 kilometres) north of the coast of Venezuela. In 1974 it attained complete independence within the Commonwealth of Nations and membership in the United Nations, the first of the six West Indies Associated States to do so.

Oval in shape, the island is approximately 21 miles (34



GRENADA

kilometres) long and 12 miles (19 kilometres) wide with an area of 120 square miles (311 square kilometres). The southern Grenadines—the largest of which is Carriacou, about 20 miles (32 kilometres) north-northwest, with an area of 13 square miles (34 square kilometres)—are a dependency.

In 1970 the population of Grenada and its dependency was 93,000, and by 1975 it was estimated at 100,000. The capital, St. George's, is on the southwest coast and has a population of about 6,300. It is also the main port, having a fine natural harbour as well as picturesque pastel-coloured houses that rise up the hillsides from the waterfront. The waterfront itself is known as the Carenage because island schooners were once careened (beached for cleaning or repair) there. St. George's is the yachting and charter-boat centre of the eastern Caribbean. The capital

The natural environment. *Relief.* Grenada is volcanic in origin, with a ridge of mountains running north and south—the steeper slopes to the west and a more gradual incline to the east and southeast. The highest point is Mt. St. Catherine (2,756 feet [840 metres]) in the northern part of the interior. The landscape is attractive, with fairly deep, steep-sided valleys, about 10,000 acres (4,000 hectares) of forest, and many plantations of bananas, cocoa, nutmegs, and sugarcane.

There are several short, swiftly flowing streams that supply all towns and most villages with piped clean water. A further source of water supply is Grand Etang, a circular lake covering 36 acres (15 hectares) in the crater of an extinct volcano at 1,740 feet (530 metres) altitude.

Climate. The climate is of the tropical maritime type, with equable temperatures varying with altitude and averaging 82° F (28° C) in the low country. Rainfall is adequate, except in the Point Salines area in the southwest; it varies from an average of 60 inches (1,524 millimetres) in coastal districts to 164 inches (4,166 millimetres) at Grand Etang. The rainy season lasts from June to December. November is the wettest month, but showers occur frequently during the other months. Grenada lies south of the usual track of hurricanes and is seldom visited by high winds.

Vegetation. The island is verdant, with a year-round growing season and a wide variety of tropical fruits, flowering shrubs, and ferns. There are also forests of teak, mahogany, saman (known as the rain tree), and blue mahoe (a strong-fibred tree) in the interior. In addition to bananas, cocoa, nutmegs, and sugarcane, commercial crops include limes, coconuts, coffee, and vegetables. Lesser quantities of spices—pepper, cloves, cinnamon, ginger, and vanilla—are also grown. Sea Island cotton is grown on Carriacou. Fruit crops

Animal life. The animal life is varied and includes domestic livestock and such wild animals as the mona monkey (a small, long-tailed, West African species introduced by slaves from Africa), the manicou (a species of opossum), the agouti (a rabbit-sized rodent, brown or grizzled in colour), the iguana, and a variety of turtles and land crabs. Birds, whose numbers were much reduced by the hurricane of 1955, are again plentiful.

History. Grenada was discovered by Christopher Columbus on August 15, 1498, when he sailed past the island without landing and gave it the name of Concepción. The origin of the name Grenada remains obscure. After its discovery by Columbus, Grenada was dominated for the next 150 years by the warlike Carib Indians, who had earlier killed off the more peaceful Arawaks. In 1609 a company of British merchants attempted to form a settlement but were forced by the Caribs to leave the island. The French governor of Martinique, du Parquet, purchased Grenada from a French company in 1650 and established a settlement at St. George's.

In 1674 Grenada became subject to the French crown, remaining so until 1762, when it capitulated to a British force. The island was formally ceded to Britain in 1763 by the Treaty of Paris. Sixteen years later, in 1779, it was recaptured by the French, only to be restored to Britain by the Treaty of Versailles in 1783.

In the late 18th century the British imported large numbers of slaves from Africa to work the sugar plantations. During 1795 and 1796, when French policy favoured the abolition of slavery, a rebellion against British rule occurred, led by a French planter and supported by the French in Martinique. The rebels massacred a number of the British, including the lieutenant governor, but the uprising was quelled.

The emancipation of the slaves finally took effect in

1833 and was accompanied by less economic and social upheaval than elsewhere because of the rapid growth of peasant proprietorship.

Grenada was the headquarters of the government of the British Windward Islands from 1885 until 1958, when Grenada joined the Federation of the West Indies at its formation. The federation was dissolved in 1962, after which Grenada attempted to federate with the remaining territories in the Windward Islands, as well as with Barbados and the Leeward Islands. On March 3, 1967, however, the island became a self-governing state in association with the United Kingdom.

In the general election of August 1967, the Grenada United Labour Party defeated the Grenada National Party and took office under the premiership of Eric M. Gairy, a trade unionist. The party was re-elected in February 1972.

Grenada became an independent nation on February 7, 1974. The transition was marked by violence, strikes, and political controversy centring upon Gairy, who was named prime minister.

The people. About 95 percent of the population is of African and mixed descent. There are also a few thousand East Indians, descendants of indentured labourers brought to replace the freed slaves; a few hundred descendants of the old French and British settlers; and a handful of more recent immigrants from North America and Europe. The century of French rule left its mark on the island; a form of patois is still spoken by older people in the villages, and many place-names, although now pronounced in an English fashion, are French. English is the accepted language, however, and the literacy rate is 93 percent. A majority of the population, predominantly of French descent, is Roman Catholic; other Christian denominations represented include Anglicans and Methodists.

Grenada, Area and Population

	area		population	
	sq mi	sq km	1960 census	1970 census
Districts				
Carriacou*	13	34	7,000	6,000
St. Andrew	37	96	22,000	22,000
St. David	17	44	9,000	11,000
St. John	15	39	8,000	9,000
St. Mark	9	23	4,000	4,000
St. Patrick	17	44	11,000	11,000
Municipal borough				
St. George's	25	65	27,000	30,000
Total Grenada	133	344†	89,000†‡	93,000

*Includes islands of Carriacou and Petit Martinique.
†Figures do not add to total given because of rounding.
‡Excludes adjustment of total to account for under-enumeration.
Source: Official government figures.

Economy. Grenada's economy is partially dependent on financial help from Britain, its principal trading partner. About 40 percent of all exports (especially bananas) go to the United Kingdom, and about a quarter of all imports (mainly manufactured goods, motor vehicles, and foodstuffs) come from the United Kingdom. The balance of trade is adverse, with imports generally exceeding exports in value.

Agriculture and tourism are the main sources of the island's income. Agriculture is the mainstay of the economy, and about 37 percent of the labour force lives off the land. To a greater extent than in any other West Indian island, Grenada's arable land is divided into small holdings on which peasant proprietors cultivate diversified crops. Less than 1 percent of the farm holdings exceed 100 acres; indeed, 87 percent of all farms are of five acres or less. Because of these small holdings and the generally hilly terrain, mechanical tilling is rare. The three major export crops, bananas, cocoa, and nutmegs, in the past were controlled by cooperative associations, but there is now greater governmental control. Banana exports, until 1969 greater in value than those of cocoa and nutmegs, depend upon preferential terms given by the United Kingdom and are affected by the European Common Market. Exports of lime juice also provide substantial earnings. Copra, and, increasingly, other products processed from the coconut, are also exported; they have surpassed bananas as export earners. A wide variety of tropical fruits are also grown—mangoes, passionfruit, guavas, tamarind, and citrus fruits. The government is encouraging increased production of vegetables, such as peas, tomatoes, sweet potatoes, pumpkins, and maize (corn).

Land holdings

Tourism, a growing factor in the island's economy, is encouraged by the government, and there are a dozen hotels, chiefly near the Grand Anse Beach south of St. George's. During the winter season, from October to March, as many as 100 cruise ships call at Grenada.

Tourism

Other sources of employment are such secondary industries as the milling of sugar, brewing, the distilling of rum (a strong white rum is made for local consumption), the processing of copra, and soapmaking. There is a cotton ginnery on Carriacou. Fishing, a traditional occupation, is limited by outmoded vessels and methods.

Customs and excise duties are the principal base for the country's budget, but taxes are also levied on incomes and professions. Expenditures are primarily for the civil service, education, roadbuilding, and waterworks. In most years there is an appreciable excess of expenditure over revenue.

Canadian, British, and U.S. banks are in operation, as well as the Grenada National Bank and Trust Company, founded in 1969 as a joint enterprise of the government and a group of U.S. investors. The unit of currency is the East Caribbean dollar (ECar$2.33 = U.S. $1; ECar$4.82 = £1 sterling in October 1975).

Transport and communications. Grenada has a network of more than 600 miles (1,000 kilometres) of roads. Bus service is available between the larger towns and villages, and there are several thousand private vehicles. Pearls Airport—providing daily air connections to nearby islands with connecting flights to South America, London, and New York—is located on the northeast coast. Plans are underway to expand air service by enlarging Pearls Airport and establishing a subsidiary airstrip at Point Salines on the southern tip of the island. There is a small airstrip on Carriacou.

The harbour at St. George's has berths for two ocean-going vessels, as well as a yacht basin and service facilities. Several shipping lines maintain regular passenger and cargo services to North America, the United Kingdom, Europe, and neighbouring West Indian islands.

Grenada was the centre for the Windward Islands Broadcasting Service, but its station became independent as Radio Grenada. Two newspapers are published in St. George's.

Administration. A member of the Commonwealth of Nations, Grenada is an independent state and a member of the United Nations, the World Bank, and other international bodies. The governor general of the island is appointed by the British crown on the advice of the prime minister of Grenada.

Prospects for the future. Grenada is faced with the problems of overpopulation and underemployment that are common to the West Indian islands. As emigration to South and Central America, Trinidad, and the United Kingdom has become more difficult, the pressure of people on resources is greater. Birth control, however, is not popular because some 60 percent of the people are Roman Catholics.

BIBLIOGRAPHY. RAYMUND DEVAS, *The Island of Grenada, 1650–1950* (1964), a good but outdated historical survey; *Conception Island* (1932), a history of the Catholic Church in Grenada; KAY FRANCIS, *This is Grenada* (1965), a popular handbook for visitors; A.W. SINGHAM, *The Hero and the Crowd in a Colonial Polity* (1962), a study of conflict between politics and administration; CARLEEN O'LOUGHLIN, *Economic and Political Change in the Leeward and Windward Islands* (1968), a competent survey; M.G. SMITH, *Kinship and Community in Carriacou* (1962), a specialized survey of ra-

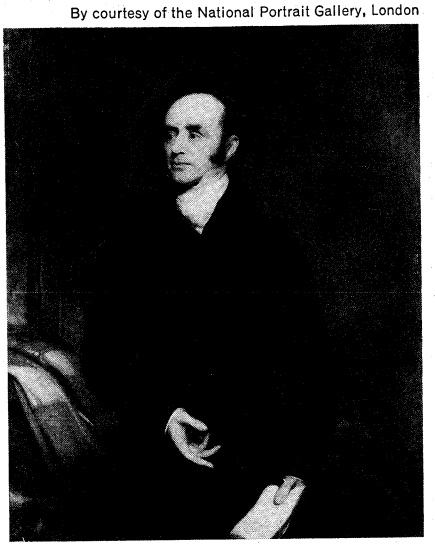
cial characteristics; *West Indies and Caribbean Year Book* a comprehensive annual reference work.

(E.V.B.B.)

Grey, Charles Grey, 2nd Earl

English politician and leader of the Whig party, Lord Grey assumed power in 1830 at the age of 66 with less experience of office than any British prime minister of the 19th century. Yet his achievement in securing the passage of the Reform Act of 1832, which extended the franchise to the middle classes and favoured larger communities in the distribution of seats, has secured him an unchallenged place among English statesmen.

By courtesy of the National Portrait Gallery, London



Grey, painting attributed to T. Phillips, *c.* 1820. In the National Portrait Gallery, London.

His father, the younger son of an old Northumberland family, was a distinguished army general who was raised to the peerage in 1801. Charles, the future second earl, was born on March 13, 1764, at Falloden, Northumberland, and received a conventional aristocratic education at Eton and Cambridge. When only 22 he was elected member of Parliament for Northumberland. Entering the London world in 1786, he gravitated immediately to the fashionable but rakish circle of the leader of the liberal Whig party, Charles James Fox; the politician-playwright Richard Sheridan; and the Prince of Wales. Handsome, witty, and attractive, Grey soon became prominent among the aristocratic Whig set that provided the political opposition to the conservative government of William Pitt (1759–1806). When the French Revolution in 1789 revived the political agitation caused by the American Revolution, Grey was one of the young Whig aristocrats who formed the Society of the Friends of the People (1792) to encourage lower and middle class demands for parliamentary reform. These activities—which at the time were considered radical—followed by the outbreak of war with revolutionary France in 1793, split the Whig party. The emotions generated by the conflict with France, the repressive, though popular, measures taken by the government, and the extreme and often absurd lengths to which Fox carried his pro-French sympathies turned his following into an impotent and discredited minority. Grey's parliamentary reform bill of 1797 was heavily defeated, and for some years afterward Fox's faction of the Whigs virtually withdrew from parliamentary life.

Entry into politics

Grey's marriage in 1794 to Mary Elizabeth Ponsonby, the daughter of a leading Irish liberal family, strengthened his sympathies with the cause of Catholic emancipation; however, it weakened his zeal for politics. A devoted husband with a growing family (numbering 15 children by 1819), Grey found contentment in a close and affectionate home life. In 1801 his bachelor uncle Sir Henry allowed him to use Howick, a country house on the Northumberland coast, as his permanent residence. Howick was four days travel from London, and Grey's dilatoriness in coming south for the parliamentary sessions frequently evoked Fox's good-humoured reproaches. Some of Grey's political extremism had also waned. His criticisms of the government for resuming the war with France in 1803 were noticeably milder than those of his chief. When on Pitt's death in 1806 Lord Grenville formed the so-called government of All the Talents that included the Fox group, Grey (now Lord Howick) became first lord of the Admiralty. When Fox died the same year, Grey took his place as foreign secretary and leader of the Foxite Whigs. The dismissal of the ministry the following year, because of a disagreement with the King over relieving Catholic disabilities, left Grey with an ingrained distaste for office without freedom of action or pledges without certainty of performance. The loss of his seat for Northumberland as a result of his Catholic sympathies, followed by his removal in 1807 to the House of Lords, increased his political detachment. In the political negotiations of 1810–12, which were initiated by the Prince of Wales when he became regent, Grey and Grenville frigidly declined to accept anything less than complete power. The end of the war came with the Pittite Cabinet of Robert Banks Jenkinson, 2nd earl of Liverpool, firmly established in office.

Foreign secretary

Between 1815 and 1830 Grey was patron, rather than leader, of the quarrelsome and divided Whig opposition. While holding that Catholic Emancipation was a condition of any genuine Whig government, he accepted the fact that parliamentary reform must wait until there was solid support for it in the country. He thought the political stability of Britain was endangered both by the reactionary postwar policy of the administration of Lord Liverpool and by the demands for democratic reform put forward by doctrinaire agitators. His private conclusion was that the task of a Whig ministry, if one could ever be formed, would be to produce a measure of reform large enough to satisfy respectable opinion and yet conservative enough to preserve the basic principles of the aristocratic constitution.

In 1830 Grey's opportunity came at last. The grant of Catholic emancipation in 1829 had destroyed the last cohesion of the conservative Liverpool party. The collapse of the Duke of Wellington's ministry in 1830 brought Grey into office on his own terms and with popular backing for a reform of the antiquated parliamentary representative system. But the extent of the changes proposed in his bill of 1831 staggered even his own supporters, and it needed a fresh general election and the coercion of the House of Lords before the bill ultimately passed into law. Grey had misjudged the temper of both houses and involved himself in a painful conflict with the new king William IV when he had reluctantly to ask for enough new peers to be created to carry the bill. He had not, however, misjudged the temper of the country. A wave of popular enthusiasm sustained him during the long battle for reform in 1831–32 and returned a vast liberal majority to the House of Commons in 1833. The epoch-making Reform Act of 1832 was the crowning achievement of the old Whig party to which Grey belonged, and he had shown courage and imagination in forcing it through to the statute book. But the measure that he envisaged as a conservative and healing act of statesmanship was regarded by many of his new supporters as a springboard for further extensive changes in church and state. The strains of the new era produced quarrels and resignations in his cabinet, and Grey retired from politics two years later. He died at Howick on July 17, 1845.

Grey's Reform Bill

BIBLIOGRAPHY. G.M. TREVELYAN, *Lord Grey of the Reform Bill* (1920), a sympathetic study, the only modern biography; C. GREY, *Some Account of the Life and Opinions of Charles, Second Earl Grey* (1861), a well-documented account of Grey's career up to 1817 by one of his sons, valuable for the author's firsthand knowledge; HENRY GREY, 3RD EARL (ed.), *The Reform Act, 1832: The Correspondence of the Late Earl Grey with His Majesty King William IV and with Sir Herbert Taylor, from Nov. 1830 to June 1832*, 2 vol. (1867), essential information on the Reform Bill crisis, 1831–32; G. LE STRANGE (ed.), *Correspondence of Princess Lieven*

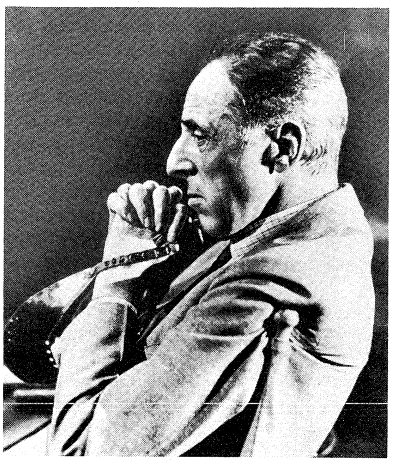
and Earl Grey, 3 vol. (1890), interesting sidelights on public and personal matters; J.R.M. BUTLER, *The Passing of the Great Reform Bill* (1914), detailed study of Grey's major achievement.

(N.G.)

Griffith, D.W.

More than any other individual, the pioneer American director D.W. Griffith developed the techniques through which motion pictures became an art form—an instrument able to express emotions and ideas. A genius in the art of the film, who never worked with a script, he innovated continually in the use of the camera angles and movement, in lighting, and, especially, in editing and tempo, and his influence throughout the world on the most creative directors of the next generation, such as Erich von Stroheim and Sergey Eisenstein, is inestimable. With the premier of his greatest success, *The Birth of a Nation*, in 1915, the previously little-known Griffith became the best known motion-picture director in the industry and, with some publicity assistance, became known as the Shakespeare of the screen.

Wide World Photos



Griffith.

Early life and influences

David Wark Griffith, the son of Jacob Griffith, a former Confederate colonel, was born on January 22, 1875, in Floydsfork, Kentucky, a tiny hamlet, later called Crestwood, not far from Louisville. He received his early education in one-room schools, largely under the tutelage of his older sister, and was subject to the strong influence of his father's imaginative stories of the Mexican and Civil wars and family readings of the works of Dickens, Shakespeare, and Sir Walter Scott. The Griffith family was impoverished upon the death of Jacob Griffith, when David was seven years old. After a brief stay with relatives, the family moved to Louisville. Griffith's formal education was terminated in secondary school by the necessity of contributing to the family's financial needs. He became, successively, an elevator operator in a dry-goods store and a clerk in a bookstore. During the latter clerkship, Griffith was exposed to the literati of Louisville and to the actors and actresses who played at Louisville's Temple Theatre.

Griffith began an acting career with several amateur theatre groups and made his professional debut in small roles with a stock company at the Temple Theatre. A barnstorming career with various touring companies followed, concluding with a Boston engagement in the spring of 1906. Following this engagement, Griffith completed a play, *A Fool and a Girl*, based on his personal experiences in the California hop fields, which was produced in Washington, D.C., in the fall of 1907. The play was a failure, despite the presence of Fannie Ward in the leading role. After the closing of the play, Griffith wrote a second play, *War*, which was based on events that occurred in the American Revolution. This later play remains unproduced.

On the advice of a former acting colleague, Griffith sold some scenarios for one-reel films, first to Edwin Porter, the director of the Edison Film Company, and then to the Biograph Company, both located in New York City. Griffith appeared as an actor in one film for the Edison Company, *Rescued from an Eagle's Nest*, under Porter's direction, and in several films for the Biograph Company. When an opening for a director developed at Biograph, Griffith was hired. During the next five years, from 1908 to 1913, Griffith made over 400 films for Biograph, the majority in the one-reel format, lasting approximately 12 minutes. His first film was *The Adventures of Dollie* (1908), about a baby stolen by and recovered from Gypsies. During the latter part of his employment, he experimented with longer films; his last Biograph film, *Judith of Bethulia* (1913), a biblical story of Judith and Holofernes, based loosely on a play of the same title by Thomas Bailey Aldrich, comprised four reels.

Career at the Biograph

During his Biograph period D.W. Griffith introduced or refined the techniques of motion-picture exposition, including the close-up, a film shot in which a single object or face filled the screen; the scenic long shot, showing an entire panoramic view; and cross-cutting, a technique of editing scenes at various locations together and intermixing them to give the impression to the viewer that the separate actions were happening simultaneously. With the assistance of his brilliant cinematographer, G.W. "Billy" Bitzer, Griffith made effective use of the fade-out and fade-in, a technique in which the screen darkens gradually to black or lightens from black to a full image, to indicate the end or the beginning of the story or of an episode, and the framing of film images through the use of special masks to produce a picture in other than the standard rectangular image. Griffith introduced to the screen young actors and actresses who were to become the motion-picture personages of the future. Included among these were Mary Pickford, Lillian and Dorothy Gish, Mack Sennett, Mae Marsh, Lionel Barrymore, and Harry Carey.

In 1913 Griffith left Biograph and entered into an agreement with Mutual Films for the direction and supervision of motion pictures. From this association, among other films, came *The Birth of a Nation*. With the official opening of the film under the title, *The Clansman*, at Clune's Auditorium in Los Angeles on February 8, 1915, the infant art of the motion picture was revolutionized. The film was subsequently lionized for its radical technique and condemned for its "racist" philosophy. Filmed at a cost of $110,000, it returned millions of dollars in profits, making it, perhaps, the most profitable film of all time, although a full accounting has never been made.

Censorship of *Birth of a Nation*

As a result of public opposition to its alleged racist message, however, *The Birth of a Nation* was censored in many cities, including New York, and Griffith became an ardent opponent of censorship of the motion picture. His next important film, *Intolerance* (1916), was, in part, an answer to his critics.

Success and failure of *Intolerance*

Intolerance, a film of epic proportions, combined four separate stories: the fall of ancient Babylon to the hordes of Cyrus, the St. Bartholomew's Day Massacre of the Huguenots in 16th-century France, the crucifixion of Jesus, and a contemporary story dealing with a wrongfully condemned man. The giant settings, especially the one representing ancient Babylon, have remained a benchmark for motion-picture spectacle, and the opulent settings for 16th-century Paris were almost equally impressive. Griffith interwove the four stories in an increasingly complex manner until all were brought to resolution in a controlled torrent of images that still leaves the viewer breathless. Only the contemporary story was given a happy ending. The film ends with an allegorical plea for the end of war through divine intervention, indicated through superimpositions of heavenly hosts above a flower-strewn battlefield. The film was an artistic success on its presentation in New York on September 5, 1916, but proved to be a financial failure. Nevertheless, tribute has been paid to its seminal influence on the work done by many film directors. Almost unanimously, critics have hailed *Intolerance* as the finest achievement of the silent film.

Most of Griffith's profits from *The Birth of a Nation* were used and lost in the making of *Intolerance*, but he was able to secure the financing for the building of his own studio in Mamaroneck, New York. His films were to be released through United Artists, a motion-picture distributor of which he was a founding partner, with Mary Pickford, Charles Chaplin, and Douglas Fairbanks. Despite making such distinguished films as *Broken Blossoms* (1919) and *Orphans of the Storm* (1921), and an extremely profitable film, *Way Down East* (1920), his studio foundered on the failure of lesser films and the business recession of the first half of the 1920s.

Director at Paramount Pictures

Griffith was subsequently employed as a director by Paramount Pictures and as contract director by United Artists. His view of the American Revolution was realized in *America* (1924), and his next-to-last film, *Abraham Lincoln* (1930), was another view of the American Civil War in a somewhat ponderous biographical style. Despite his past success and the general acknowledgement of his vital contributions to the syntax of the motion picture, Griffith was unable to find permanent employment after *Abraham Lincoln*. His last film, *The Struggle* (1931), a grim study of the degeneration of an alcoholic husband, was an abject failure, withdrawn by United Artists after a brief run. Griffith had produced *The Struggle* independently and, although not destitute, was never again able to finance another film or to find regular employment in the motion-picture industry. He died on July 23, 1948, in Hollywood.

Griffith married twice. His first wife, Linda Arvidson, was an actress. They were married in 1906 and separated in 1911. Griffith won a divorce in 1936 and then married Evelyn Marjorie Baldwin. This second marriage also ended in divorce. There were no children from either marriage.

MAJOR WORKS

The Adventures of Dollie; The Taming of the Shrew; (both 1908). *Edgar Allan Poe; The Curtain Pole; The Voice of the Violin; The Drunkard's Reformation; Resurrection; The Cricket on the Hearth; The Lonely Villa; The Mended Lute; 1776: or, The Hessian Renegades; Pippa Passes; In the Watches of the Night; Through the Breakers; Lines of White on a Sullen Sea; The Red Man's View; A Corner in Wheat* (all 1909). *In Old California; The Unchanging Sea; Ramona; The Usurer; The Message of the Violin;* (all 1910). *His Trust; His Trust Fulfilled; Fisher Folks; The Lonedale Operator; Enoch Arden; Fighting Blood; The Last Drop of Water; The Battle; The Miser's Heart;* (all 1911). *The Mender of Nets; The Goddess of Sagebrush Gulch; A Girl and Her Trust; Old Lena and the Geese; Man's Genesis; The Sands of Dee; A Pueblo Legend; An Unseen Enemy; The Musketeers of Pig Alley; The Massacre; The New York Hat;* (all 1912). *The Battle at Elderbush Gulch; Judith of Bethulia* (all 1913); *The Battle of the Sexes; The Escape; Home Sweet Home; The Avenging Conscience;* (all 1914). *The Birth of a Nation* (1915); *Intolerance* (1916); *Hearts of the World; The Great Love; The Greatest Thing in Life* (all 1918). *A Romance of Happy Valley; The Girl Who Stayed at Home; Broken Blossoms; True Heart Susie; The Mother and the Law; Scarlet Days; The Greatest Question* (all 1919). *The Idol Dancer; The Love Flower; Way Down East* (all 1920). *Dream Street* (1921); *Orphans of the Storm* and *One Exciting Night* (both 1922); *The White Rose* (1923); *America* (1924); *Sally of the Sawdust; Isn't Life Wonderful?; That Royle Girl* (all 1925); *Sorrows of Satan* (1927); *Drums of Love* and *The Battle of the Sexes* (both 1928); *Lady of the Pavements* (1929); *Abraham Lincoln* (1930); *The Struggle* (1931).

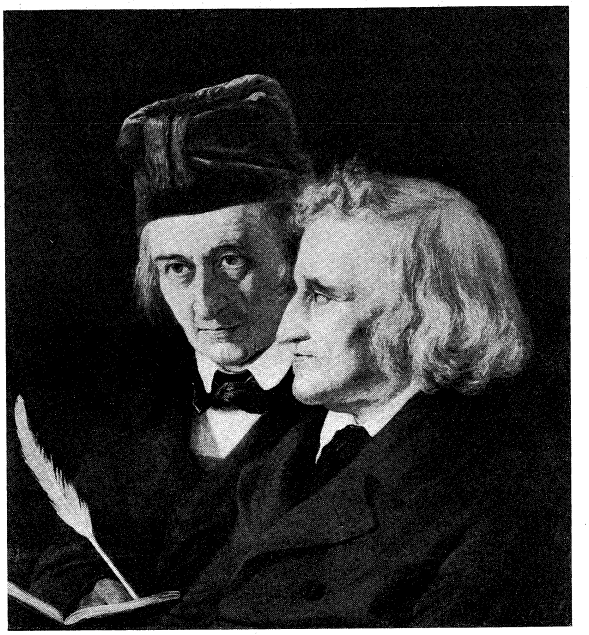
BIBLIOGRAPHY. ROBERT M. HENDERSON, *D.W. Griffith: His Life and Times* (1972), is the only full-length biography; the same author's *D.W. Griffith: The Years at Biograph* (1970), is a detailed study of Griffith's apprentice years as a film director for the Biograph Film Company from 1908 to 1913. The latter book contains an extensive bibliography. IRIS BARRY and EILEEN BOWSER, *D.W. Griffith, American Film Master*, rev. ed. (1965), is a perceptive monograph that gives a brief summary of Griffith's career, credits for the major films, and a critique of the major films after the Biograph period. The following books devote chapters to Griffith and his career in the context of the history of the motion picture: ALBERT R. FULTON, *Motion Pictures: The Development of an Art from Silent Films to the Age of Television* (1960); BENJAMIN B. HAMPTON, *A History of the Movies* (1931, reprinted 1970); LEWIS JACOBS, *The Rise of the American Film* (1939); ARTHUR KNIGHT, *The Liveliest Art* (1957).

(R.M.He.)

Grimm Brothers

The fairy tales of the famed German scholars Jacob and Wilhelm Grimm collectively became a classic of world literature, gave new direction to writing for children, and drew attention to the long-neglected riches of oral traditions. The Grimms' work on their *Kinder- und Hausmärchen* and their comparative studies of the folk literature of many nations led to the birth of the science of folklore. Through his grammar book Jacob became one of the founders of historical linguistics and the science of Germanic philology. His collection of ancient law practices and his *Deutsche Mythologie* were of far-reaching influence everywhere. The *Deutsches Wörterbuch*, begun by the two, set an international example for the historical dictionary of language.

By courtesy of the Staatliche Museen zu Berlin



Jacob (right) and Wilhelm Grimm, oil portrait by Elisabeth Jerichau-Baumann, 1855. In the National-Galerie, Berlin.

Beginnings and Kassel period. Jacob Ludwig Carl Grimm was born on January 4, 1785, and Wilhelm Carl Grimm on February 24, 1786, in Hanau, near Frankfurt am Main, in Hessen-Kassel. They were the oldest in a family of five brothers and one sister. Their father, Philipp Wilhelm, a lawyer, was town clerk in Hanau and later justiciary in Steinau, another small Hessian town, where his father and grandfather had been ministers of the Calvinistic Reformed Church. The father's death in 1796 brought social hardships to the family; the death of the mother in 1808 left 23-year-old Jacob with the responsibility of four brothers and one sister. Jacob, a scholarly type, was small and slender with sharply cut features, while Wilhelm was taller, had a softer face, and was sociable and fond of all the arts. After attending the high school in Kassel, the brothers followed their father's footsteps and studied law at the University of Marburg (1802–06) with the intention of entering civil service. At Marburg they came under the influence of Clemens Brentano, who awakened in both a love of folk poetry, and Friedrich Karl von Savigny, cofounder of the historical school of jurisprudence, who taught them a method of antiquarian investigation that formed the real basis of all their later work. Others, too, strongly influenced the Grimms, particularly the philosopher Johann Gottfried Herder (1744–1803), with his ideas on folk poetry. Essentially, they remained individuals, creating their work according to their own principles. In 1805 Jacob accompanied Savigny to Paris to do research on legal manu-

Influence of Savigny and Brentano

scripts of the Middle Ages; the following year he became secretary to the war office in Kassel. Because of his health, Wilhelm remained without regular employment until 1814. After the French entered in 1806, Jacob became private librarian to King Jérôme of Westphalia in 1808 and a year later *auditeur* of the Conseil d 'État but returned to Hessian service in 1813 after Napoleon's defeat. As secretary to the legation, he went twice to Paris (1814–15), to recover precious books and paintings taken by the French from Hesse and Prussia. He also took part in the Congress of Vienna (September 1814–June 1815). Meantime, Wilhelm had become secretary at the Elector's library in Kassel (1814), and Jacob joined him there in 1816. By that time the brothers had definitely given up thoughts of a legal career in favour of purely literary research. In the years to follow they lived frugally and worked steadily, laying the foundations for their lifelong interests. Their whole thinking was rooted in the social and political changes of their time and the challenge these changes held. Jacob and Wilhelm had nothing in common with the fashionable "Gothic" Romanticism of the 18th and 19th centuries, a state of mind that made them more realists than romantics. They investigated the distant past and saw in antiquity the foundation of all social institutions of their days. But their efforts to preserve these foundations did not mean that they wanted to return to the past. From the beginning, the Grimms sought to include material from beyond their own frontiers—from the literary traditions of Scandinavia, Spain, The Netherlands, Ireland, Scotland, England, Serbia, and Finland.

They first collected folk songs and tales for their friends Achim von Arnim and Clemens Brentano, who had collaborated on an influential collection of folk lyrics in 1805, and the brothers examined in some critical essays the essential difference between folk literature and other writing. To them, folk poetry was the only true poetry, expressing the eternal joys and sorrows, the hopes and fears of mankind.

Publication of the fairy tales

Encouraged by Arnim, they published their collected tales as the *Kinder- und Hausmärchen*, implying in the title that the stories were meant for adults and children alike. In contrast to the extravagant fantasy of the Romantic school's poetical fairy tales, the 200 stories of this collection (mostly taken from verbal sources, though a few were from printed sources) aimed at conveying the soul, imagination, and beliefs of people through the centuries—or at a genuine reproduction of the teller's words and ways. The great merit of Wilhelm Grimm is that he gave the fairy tales a readable form without changing their folkloric character. The results were threefold: the collection enjoyed wide distribution in Germany and eventually in all parts of the globe (there are now translations in 70 languages); it became and remains a model for the collecting of folktales everywhere; and the Grimms' notes to the tales, along with other investigations, formed the basis for the science of the folk narrative and even of folklore. To this day the tales remain the earliest "scientific" collection of folktales. The *Kinder- und Hausmärchen* was followed by a collection of historical and local legends of Germany, *Deutsche Sagen* (1816–18), which never gained wide popular appeal, though it influenced both literature and the study of the folk narrative. The brothers then published (in 1826) a translation of Thomas Crofton Croker's *Fairy Legends and Traditions of the South of Ireland*, prefacing the edition with a lengthy introduction of their own on fairy lore. At the same time, the Grimms gave their attention to the written documents of early literature, bringing out new editions of ancient texts, from both the Germanic and other languages. Wilhelm's outstanding contribution was *Die deutsche Heldensage* ("The German Heroic Saga"), a collection of themes and names from heroic legends mentioned in literature and art from the 6th to the 16th centuries, together with essays on the art of the saga.

While collaborating on these subjects for two decades (1806–26), Jacob also turned to the study of linguistics with an extensive work on grammar, the *Deutsche Grammatik* (1819–37). The word *deutsch* in the title does not mean strictly "German," but it rather refers to the etymological meaning of "common," thus being used to apply to all of the Germanic languages, the historical development of which is traced for the first time. He represented the natural laws of sound change (both vowels and consonants) in various languages and thus created bases for a method of scientific etymology; *i.e.*, research into relationships between languages and development of meaning. In what was to become known as Grimm's law, Jacob demonstrated the principle of the regularity of correspondence among consonants in genetically related languages, a principle previously observed by the Dane Rasmus Rask. Jacob's work on grammar exercised an enormous influence on the contemporary study of linguistics, Germanic, Romance, and Slavic, and it remains of value and in use even now. In 1824 Jacob Grimm translated a Serbian grammar by his friend Vuk Stefanović Karadžić, writing an erudite introduction on Slavic languages and literature.

Grimm's law

He extended his investigations into the Germanic folk-culture with a study of ancient law practices and beliefs, *Deutsche Rechtsaltertümer* (1828), providing systematic source material but excluding actual laws. The work stimulated other publications in France, The Netherlands, Russia, and the southern Slavic countries and has not yet been superseded.

The Göttingen years. The quiet contentment of the years at Kassel ended in 1829, when the brothers suffered a snub—perhaps motivated politically—from the Elector of Hessen-Kassel: they were not given advancement following the death of a senior colleague. Consequently, they moved to the nearby University of Göttingen, where they were appointed librarians and professors. Jacob Grimm's *Deutsche Mythologie*, written during this period, was to be of far-reaching influence. From poetry, fairy tales, and folkloristic elements, he traced the pre-Christian faith and superstitions of the Germanic people, contrasting the beliefs to those of classical mythology and Christianity. The *Mythologie* had many successors all over Europe, but often disciples were not as careful in their judgments as Jacob had been. Wilhelm published here his outstanding edition of Freidank's epigrams. But again fate overtook them. When Ernest Augustus, duke of Cumberland, became king of Hanover, he high-handedly repealed the constitution of 1833, which he considered too liberal. Two weeks after the King's declaration, the Grimms, together with five other professors (the "Göttingen Seven"), sent a protest to the King, explaining that they felt themselves bound by oath to the old constitution. As a result they were dismissed, and three professors, including Jacob, were ordered to leave the kingdom of Hanover at once. Through their part in this protest directed against despotic authority, they clearly demonstrated the academic's sense of civil responsibilities, manifesting their own liberal convictions at the same time. During three years of exile in Kassel, institutions in Germany and beyond (Hamburg, Marburg, Rostock, Weimar, Belgium, France, The Netherlands, and Switzerland) tried to obtain the brothers' services.

The "Göttingen Seven"

The Berlin period. In 1840 they accepted an invitation from the king of Prussia, Frederick William IV, to go to Berlin, where as members of the Royal Academy of Sciences they lectured at the university. There they began their most ambitious enterprise, the *Deutsches Wörterbuch*, a large German dictionary intended as a guide for the user of the written and spoken word as well as a scholarly reference work. In the dictionary, all German words found in the literature of the three centuries "from Luther to Goethe" were given with their historical variants, their etymology, and their semantic development; their usage in specialized and everyday language was illustrated by quoting idioms and proverbs. Begun as a source of income in 1838 for the brothers after their dismissal from Göttingen, the work required generations of successors to bring the gigantic task to an end in our day. Jacob lived to see the work proceed to the letter *F*, while Wilhelm only finished the letter *D*. The dictionary became an example for similar publications in other

countries: Britain, France, The Netherlands, Sweden, and Switzerland. Jacob's philological research later led to a history of the German language, *Geschichte der deutschen Sprache*, in which he attempted to combine the historical study of language with the study of early history. Research into names and dialects was stimulated by Jacob Grimm's work, as were ways of writing and spelling—for example, he used roman type and advocated spelling German nouns without capital letters.

For some 20 years they worked in Prussia's capital, respected and free from financial worries. Much of importance can be found in the brothers' lectures and essays, the prefaces and reviews (*Kleinere Schriften*) they wrote in this period. In Berlin they witnessed the Revolution of 1848 and took an active part in the political strife of the succeeding years. In spite of close and even emotional ties to their homeland, the Grimms were not nationalists in the narrow sense. They maintained genuine —even political—friendships with colleagues at home and abroad, among them the jurists Savigny and Eichhorn; the historians F.C. Dahlmann, G.G. Gervinus, and Jules Michelet; and the philologists Karl Lachmann, John Mitchell Kemble, Jan Frans Willems, Vuk Karadžić, Pavel Josef Šafařik. Nearly all academies in Europe were proud to count Jacob and Wilhelm among their members. The more robust Jacob undertook many journeys for scientific investigations, visiting France, The Netherlands, Belgium, Switzerland, Austria, Italy, Denmark, and Sweden. Jacob remained a bachelor; Wilhelm married Dorothea Wild from Kassel, with whom he had three children: Herman (literary and art historian, 1828–1901), Rudolf (jurist, 1830–89), and Auguste (1832–1919). Wilhelm Grimm died on December 16, 1859, and Jacob on September 20, 1863. Their graves are in the Matthäikirchhof in West Berlin.

MAJOR WORKS

JOINT WORKS: *Kinder- und Hausmärchen* (2 vol. 1812–15; 3 vol. 1819–22), of which there are many translations into English, generally as *Grimm's Fairy Tales*, complete edition based on trans. by Margaret Hunt (1944), by Joseph Campbell (1944), by Francis P. Magoun, Jr. and Alexander H. Krappe as *The Grimms' German Folk Tales* (1960; also published in 1963 and 1965); *Altdeutsche Wälder*, 3 vol. (1813–16); *Deutsche Sagen*, 2 vol. (1816–18); *Deutsches Wörterbuch* (1852–1960; new ed. 1965 ff.).

BY JACOB: *Über den altdeutschen Meistergesang* (1811); *Deutsche Grammatik*, 4 vol. (1819–37); *Deutsche Rechtsaltertümer* (1828); *Reinhart Fuchs* (1834); *Deutsche Mythologie* (1835); *Geschichte der deutschen Sprache*, 2 vol. (1848); *Kleinere Schriften*, 8 vol. (1864–90, reprinted 1965).

BY WILHELM: *Altdänische Heldenlieder, Balladen und Märchen* (1811); *Über deutsche Runen* (1821); *Grâve Ruodolf* (1828); *Die deutsche Heldensage* (1829); *Vrîdankes Bescheidenheit* (1834); *Kleinere Schriften*, 4 vol. (1881–87).

BIBLIOGRAPHY. L. DENECKE, *Jacob Grimm und sein Bruder Wilhelm* (1971); and K. SCHULTE KEMMINGHAUSEN and L. DENECKE, *Die Brüder Grimm in Bildern ihrer Zeit* (1963), standard modern biographies; earlier studies include: *Die Selbstbiographien von Jacob und Wilhelm Grimm* (1830; ed. by I. SCHNACK, 1958); W. SCHERER, *Jacob Grimm* (1865, 1885, new ed. 1921), antiquated but not dispensable; K. ZUCKMAYER, *Die Brüder Grimm. Ein deutscher Beitrag zur Humanität* (1948, 1972); H. GERSTNER (ed.), *Die Brüder Grimm: Ihr Leben und Werk in Selbstzeugnissen, Briefen und Aufzeichnungen* (1952), collected documents; W. SCHOOF, *Jacob Grimm* (1961) and *Wilhelm Grimm* (1960), collected essays with documents; G. GINSCHEL, *Der junge Jacob Grimm, 1805–1819* (1967), an outstanding work. English language works: RUTH MICHAELIS-JENA, *The Brothers Grimm* (1970), detailed and fully illustrated; MURIEL E. HAMMOND, *Jacob and Wilhelm Grimm* (1968), a pleasant introductory work; MURRAY B. PEPPARD, *Paths Through the Forest: A Biography of the Brothers Grimm* (1971), learned and readable.

Festschrift publications: *Brüder Grimm Gedenken 1963*, ed. by L. DENECKE and I.M. GREVERUS with G. HEILFURTH (1963), contains 27 contributions from 15 countries, East and West; *Jacob Grimm zur 100. Wiederkehr seines Todestages*, ed. by W. FRAENGER and W. STEINITZ (1963), 12 contributions from 6 East-European countries.

Letters: For a complete list of edited correspondence, see L. DENECKE (above); for a choice of important editions, RUTH MICHAELIS-JENA (above).

Special subjects: R. STEIG, *Goethe und die Brüder Grimm* (1892, reprinted 1972), contains surprising aspects of this relationship; H. KUECK, *Die Göttinger Sieben* (1934), with political tracts and documents; W. EBEL, *Jacob Grimm und die deutsche Rechtswissenschaft* (1963), a concise study.

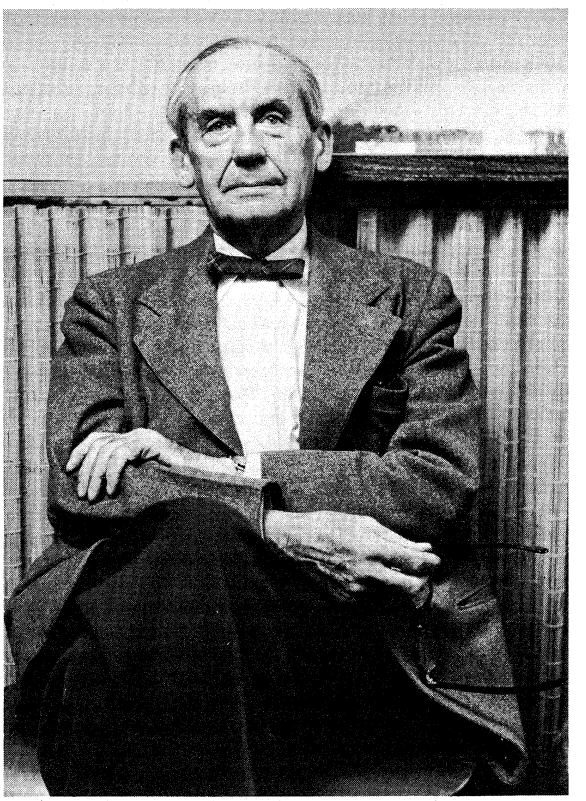
On the nursery and household tales: H. HAMANN, *Die literarischen Vorlagen der Kinder- und Hausmärchen und ihre Bearbeitung durch die Brüder Grimm* (1906); T.F. CRANE, "The External History of the *Kinder- und Hausmärchen* of the Brothers Grimm," *Modern Philology*, 14:577–610, 15:65–77, 355–383 (1917); K. SCHMIDT, *Die Entwicklung der Grimmschen Kinder- und Hausmärchen seit der Urhandschrift* (1932); W. SCHOOF, *Zur Entwicklungsgeschichte der Grimmschen Märchen* (1959); Q. GERSTL, *Die Brüder Grimm als Erzieher: Pädagogische Analyse des Märchens* (1964); C. BUEHLER, *Das Märchen und die Phantasie des Kindes*, 6th ed. (1970).

(L.De.)

Gropius, Walter

Walter Gropius, German-born architect and educator, was an enthusiastic and influential proponent of modern design who furthered his ideas through the Bauhaus school of design (which he founded in Germany in 1919), through his own architectural works, and through his long years of teaching at Harvard University. Among his most important ideas was his belief that all design—whether of a chair, a building, or a city—should be approached in essentially the same way: through a systematic study of needs and problems, taking into account modern construction materials and techniques, without reference to previous forms or styles.

Erich Hartmann—Magnum



Walter Gropius, photographed by Erich Hartmann.

Youth and early training. Walter Adolph Gropius was born in Berlin on May 18, 1883, of an architect father, Walter, and Manon (Scharnweber) Gropius. He studied architecture at the technical institutes in Munich (1903–04) and in Berlin–Charlottenburg (1905–07). He worked briefly in an architectural office in Berlin (1904) and saw military service (1904–05). Before completing school he built his first buildings, farm labourers' cottages in Pomerania (1906). He travelled for a year in Italy, Spain, and England and, in 1907, joined the office of the architect Peter Behrens in Berlin.

Work with Peter Behrens

Gropius acknowledged that his work with Behrens and the design problems he undertook for a German electricity company did much to shape his lifelong interest in progressive architecture and the interrelationship of the arts. From the time he left Behrens in 1910 until 1914, Gropius developed a clear commitment and talent for organization and for promoting his ideas on the arts. In 1911 he became a member of the Deutscher Werkbund (German Work Union), which had been founded in 1907 to ally creative designers with machine production. Gropius argued for such building techniques as prefabrication of parts and assembly on the site. However much he accepted the inevitability and restrictions of mechanization, he felt it was up to the artistically trained designer to "breathe a soul into the dead product of the machine." He was against imitation, snobbery, and dogma in the arts and cautioned against such oversimplification as the notion that the function of a product should determine its appearance.

Gropius' growing intellectual leadership was complemented by his design of two significant buildings, both done in collaboration with Adolph Meyer: the Fagus Works at Alfeld-an-der-Leine (1911) and the model office and factory buildings in Cologne (1914) done for the Werkbund Exposition. The Fagus Works, bolder than any of Behrens' works, is marked by large areas of glass wall broken by visible steel supports, the whole done with little affectation. The Cologne buildings were more formal, some say influenced by the U.S. architect Frank Lloyd Wright. Together these two buildings testify to Gropius' design maturity prior to World War I.

Military service and marriage

During that war Gropius served as a cavalry officer on the western front, was wounded, and received the Iron Cross for bravery. In 1915 he married a widow, Alma (Schindler) Mahler, whom he had met in 1910 when she was still married to the Austrian composer Gustav Mahler. Their wartime marriage, dependent on furloughs, was complicated by her affair with the German author Franz Werfel, and they were divorced in 1919. Their only child, Alma Manon, died in 1935.

Bauhaus period. Even before the end of the war, the city of Weimar approached Gropius for his ideas on art education. In April 1919 he became director of the Grand Ducal Saxon School of Arts and Crafts, the Grand Ducal Saxon Academy of Arts, and the Grand Ducal Saxon School of Arts, which were immediately united as Staatliches Bauhaus Weimar. Acceptance of this appointment was the most decisive step in his career. With his temperament for the practical world of art, politics, and administration, Gropius succeeded in establishing a viable new approach to design education, one that became an international prototype and eventually supplanted the 200-year-old supremacy of the French École des Beaux-Arts.

Gropius' Bauhaus program

A key tenet of Gropius' Bauhaus teaching was the requirement that architect and designer undergo a practical crafts training to acquaint himself with materials and processes. Although the program was to have been a comprehensive one, budget limitations permitted only a portion of the crafts shops to open. No formal study of architecture was offered at Weimar. Despite the early Werkbund principle of joining art with industry, much activity centred around handicrafts, such as ceramics, weaving, and stained-glass design. Many painters and sculptors joined the staff: Paul Klee, Lyonel Feininger, Wassily Kandinsky, Gerhard Marcks, and, later, László Moholy-Nagy and Josef Albers—altogether an astonishing roster of artists.

Somehow it did not seem incongruous for artists to be teaching applied design. As an introduction to design principles, a beginning course, *Vorkurs*, was developed by the Swiss painter and sculptor Johannes Itten, which itself became the most widely copied aspect of the Bauhaus curriculum. Students explored two- and three-dimensional design using a variety of simple materials, such as wire, wood, and paper. The psychological effects of form, colour, and texture were studied as well. Although his instructors were gifted, it was Gropius' own persistence that made this educational experiment work.

Historians disagree on the character of the early Bauhaus years. Certainly in 1919–22 Bauhaus students were allowed to express subjective feelings in their art; individuality and expressionism were not uncommon. The prewar Gropius belief that art must conform to and express the economic character and rational order of modern society seemed to be submerged in a new belief that the greatness of art stood above utilitarian considerations. A reverse shift came in 1922, not without controversy; Itten left, and a more rational and objective approach returned. The individually made products were intended as prototypes for machine production, and some designs were produced commercially. They emphasized geometrical forms, smooth surfaces, regular outlines, primary colours, and modern materials—all of which, to many eyes, epitomized impersonality in art. It is this last phase of Bauhaus output that is publicly accepted as characteristic of Bauhaus "style," although Gropius himself disdained the use of the word concept.

Gropius saw architecture and design as ever changing, always related to the contemporary world. He spoke of the architect's duty to encompass the total visual environment. He himself designed furniture, a railroad car, and an automobile. He emphasized housing and city planning, the usefulness of sociology, and the necessity of using teams of specialists.

Move to Dessau

In 1925 the Bauhaus moved to Dessau with the promise of better financial support and to escape the growing antagonism of the Weimar community. In Dessau, Gropius built the school building and faculty housing (1925–26). The school itself is a key monument of modern architecture and Gropius' best-known building. Its dynamic composition, asymmetrical plan, smooth white walls set with horizontal windows, and flat roof are features associated with the so-called International Style of the 1920s. Gropius resigned as director in 1928 to return to private practice in Berlin. During 1929–30 he designed a portion of a housing colony in Berlin–Siemensstadt. Gropius' regular facades of enormous length, together with a rigid orientation, illustrate an excessively intellectual solution with a "curse of uniformity," which Gropius himself decried in later years.

Harvard years. Unsympathetic to the Nazi regime, he and his second wife, Ise Frank, whom he had married in 1923, left Germany secretly via Italy for exile in England in 1934. Hitler's government closed the Bauhaus in 1933. Gropius' brief time in England was marked by collaboration with the architect Maxwell Fry that resulted in their important work, Village College at Impington, Cambridgeshire (1936).

In February 1937 Gropius arrived in Cambridge, Massachusetts, to become professor of architecture at Harvard University. The following year he was made chairman of the department, a post he held until his retirement in 1952. He became a naturalized U.S. citizen in 1944. At Harvard he introduced the Bauhaus philosophy of design into the curriculum, although he was unable to implement workshop training. He was also unsuccessful in abolishing the history of architecture as a course. His crusade for modern design, however, was immediately popular among the students. His innovations at Harvard soon provoked similar educational reform in other architectural schools in the United States and marked the beginning of the end of a historically imitative architecture.

In addition to his teaching, Gropius collaborated with Marcel Breuer, a former Bauhaus pupil and later fellow teacher, from 1937 until 1940. Among their designs was Gropius' own house in Lincoln, Massachusetts, which, with its use of white-painted wood and fieldstone, restated New England traditionalism in modern terms. This house and others designed by them were controversial, but the architects lived to see acceptance of their ideas. In 1942 Gropius renewed his interest in the production of architecture by industry when he became the vice president of General Panel Corporation, a company that made prefabricated housing. He retired in 1952.

The Architects Collaborative

In 1946, with six of his former Harvard pupils as partners, Gropius formed The Architects Collaborative (TAC), based in Cambridge. Among its varied U.S. and interna-

tional commissions, TAC received one to do the Harvard University Graduate Center (1949–50), a grouping of dormitory buildings and dining commons. The design is reminiscent of but less forceful than the Dessau Bauhaus buildings. Other TAC designs include the United States Embassy in Athens (1960) and the University of Baghdad, Iraq (design accepted 1960, still under construction). Gropius remained an active member of TAC until he died at the age of 86. In accord with his request made in 1933 that his funeral not be a mournful affair but marked in a festive manner, 70 friends in Cambridge drank champagne in his remembrance two days after his death in Boston on July 5, 1969. He was survived by his wife and adopted daughter, Beate, who was Ise Gropius' niece.

Assessment. Most assessments of Gropius' career centre upon his achievements as educator and author rather than as architect. Dedicated to the reform of art and to administration, as a man of visionary zeal and practical ambition, Gropius can be said to have fully achieved his goal. In his own assignments he turned away from personal and subjective aspects in favour of reaching for intellectual solutions of larger and socially urgent problems. His architecture does not have the aesthetic fascination of Wright's or Le Corbusier's but reflects a sober and programmatic concern that marked his whole life. Yet always, in conversation and criticism, he reminded his pupils of the vitality of the individual spirit, of the spontaneity of life itself. His habit of wearing a beret with a business suit was perhaps symbolic of the two worlds he hoped to bridge, "the gap between the rigid mentality of the businessman and technologist and the imagination of the creative artist."

MAJOR WORKS

Fagus Works, Alfeld-an-der-Leine, Germany (1911); with A. Meyer, Pavilion for Deutz Machinery Factory, Werkbund Exhibition, Cologne (1914); Bauhaus, Dessau, Germany (1925–26); Toerten Housing Development, Dessau (1926–28); Municipal Employment Office, Dessau (1927–28); Siemensstadt Housing Estate, Berlin (1929–30); Gropius House, Lincoln, Massachusetts (1937); Housing Development, New Kensington, Pennsylvania (1938–41); Harvard University Graduate Center, Cambridge, Massachusetts (1949–50); U.S. Embassy, Athens (commissioned 1960).

BIBLIOGRAPHY. HERBERT BAYER, WALTER GROPIUS, and ISE GROPIUS (eds.), *Bauhaus, 1919–1928* (1938, reprinted 1959); WALTER GROPIUS, *The Scope of Total Architecture* (1955), a summary of Gropius' comprehensive approach to design; WALTER GROPIUS *et al.* (eds.), *The Architects Collaborative, 1945–65* (1966); HANS M. WINGLER, *Das Bauhaus 1919–1933: Weimar, Dessau, Berlin und die Nachfolge in Chicago seit 1937*, 2nd rev. ed. (1968; Eng. trans., *The Bauhaus: Weimar, Dessau, Berlin, Chicago*, 1969), a detailed documentary record, with many illustrations; JAMES M. FITCH, *Walter Gropius* (1960), a brief, well-illustrated, general account.

(H.F.K.)

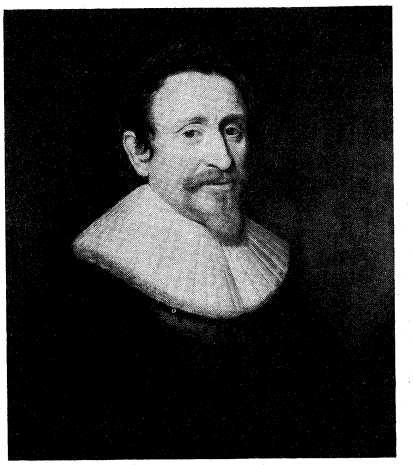
Grotius, Hugo

Hugo Grotius (Huigh or Hugeianus de Groot) was a Dutch statesman, jurist, philologist, theologian, and Humanist whose juridical works are of fundamental importance in international law.

Childhood and education

He was born at Delft on April 10, 1583. His father, a learned man, had been burgomaster of Delft and curator of Leiden University. After initial schooling in Delft, his father entrusted him to the Hague preacher and theologian Johannes Uyttenbogaert, who was to play a leading role in a crisis over Arminianism in 1618–19 that resulted in years of imprisonment for Grotius. An extremely gifted child, Grotius wrote Latin elegies at the age of eight and became a student in the faculty of letters at Leiden University at the age of 11. Very soon Grotius was the best pupil of the famed Latinist Joseph Scaliger, who undoubtedly contributed greatly to his development as a poet.

When at the age of 15 he accompanied the leading statesman Johan van Oldenbarnevelt on an embassy to Henry IV of France, he was received there with great honour and decided to remain to study law at Orléans. That same year his *Pontifex Romanus* appeared, six monologues offering a synthesis of the political situation



Grotius, portrait by M.J. van Mierevelt (1567–1641). In the Rijksmuseum, Amsterdam.
By courtesy of the Rijksmuseum, Amsterdam

in 1598. In later years he was to regret having chosen a career in law rather than letters. In 1599 he returned to Holland and settled as an advocate in The Hague, lodging for a time with his former teacher Uyttenbogaert. In 1600 his "Mirabilia" appeared, a poem about what had taken place on land and sea in the first half of that year. In 1601 the states of Holland appointed the able young lawyer their official Latin historiographer and specifically requested from him a description of the Dutch republic's revolt against Spain. Always interested in the history of his native country, Grotius began the same year, entitling his work *Annales et Historiae de Rebus Belgicis* in the manner of the Roman historian Tacitus. It consists of two parts: the "Annales," covering the period from 1559 to 1588, and the "Historiae" beginning in 1588 and concluding with the Twelve Years' Truce (1609–21).

Grotius was also very prolific in philology and poetry. He edited, with a commentary, Martianus Capella's handbook of the seven liberal arts (1598) and at about the same time the "Phaenomena," an astronomical poem by the Greek Aratus of Soli; he also published, in collaboration with the Humanist Daniel Heinsius, a Latin translation of the Greek bucolic poet Theocritus. In 1601 his *Sacra*, a volume of Latin poetry, appeared, made up of sacred poems, together with the drama *Adamus Exul* ("Adam in Exile"). The latter work, widely read and imitated, was greatly admired by the English poet John Milton. In 1614 he again edited a Roman author, Lucanus.

Argument for freedom of the seas

Involvement in politics. Increasingly, however, he became involved in Dutch politics. Although the republic was then at peace with the united kingdom of Spain and Portugal, the latter claimed a monopoly of trade with the East Indies. When a Dutch admiral seized the Portuguese vessel "Santa Catarina," the Dutch East India Company asked Grotius in 1604 to write a juridical treatise, "De Jure Praedae" ("On the Law of Prize and Booty"), defending the action on the ground that Spain–Portugal had deprived the Dutch of their trading rights. In 1609 one chapter of it, in which Grotius defends free access to the ocean for all nations, appeared under the title "Mare Liberum." The work circulated widely and was often reprinted.

In 1607 Grotius was appointed *advocaat fiscaal* (attorney general) of the province of Holland. In 1608 he married Maria van Reigersberch, the daughter of the burgomaster of Veere, an intelligent and courageous woman who stood by him unwaveringly in difficult years. They had seven children. In the same year he published *Christus Patiens*, a drama that was to be widely imitated.

In 1613 Grotius led an embassy to James I of England. Its official purpose was the settlement of trade differences, but he took advantage of the opportunity to discuss religious matters with the King as well, especially the reunion of all Christian churches, a problem that

concerned him deeply. The same year he became deeply involved in the religious and political controversy that was dividing the republic. Originally this had been a theological argument about predestination between two Leiden professors, Jacobus Arminius and Franciscus Gomarus; it developed into a dispute between the province of Holland and the orthodox Calvinist majority of the States General of the Netherlands under the leadership of Prince Maurice. Grotius, though a gentle and moderate man who always strove for peace and unity among Christians, was profoundly influenced by Oldenbarnevelt and the provinces. Under orders from the estates (assembly) of Holland he published in 1613 the *Ordinum Hollandiae et Westfrisae Pietas*, an ardent plea for the ecclesiastical policy of the estates.

Imprisonment and exile

In 1618 Prince Maurice ordered the arrest of the leaders of the opposition, including Grotius and the statesman Johan van Oldenbarnevelt. The latter was executed for high treason; Grotius was sentenced to life imprisonment and incarcerated in the castle of Loevestein. His wife and children were permitted to join him there. In prison he wrote a poem in behalf of Dutch sailors, whom he saw as peaceful propagators of the Christian faith, "Bewijs van den waren Godsdienst." He later translated it into Latin prose as "De Veritate Religionis Christianae" (1627). It established Grotius' fame and was translated into 13 languages, including Arabic and Urdu. He also began an introduction to the jurisprudence of Holland, *Inleydinghe tot de Hollandsche Rechts-geleerdheyt*, a very important work published in 1631. In the Republic of South Africa, the *Inleydinghe* was law from 1859 to 1901.

Hidden in a chest of books, he made a celebrated escape from the castle of Loevestein on March 22, 1621. He fled to Antwerp and to Paris, where he was received with great honour by Louis XIII and numerous statesmen and scholars. His wife and children were permitted to join him in Paris. There the family lived precariously on what Grotius was able to earn with his pen. Although Louis granted him a pension, it was paid rather irregularly; as a Calvinist he was unable to obtain a professorship.

Life in Paris. In 1625, still in exile, he published his legal masterpiece *De Jure Belli ac Pacis* (*On the Law of War and Peace*), in which he laid the foundations of international law. He defended the position taken by Holland and himself in the religious conflict in *Apologeticus eorum qui Hollandiae Westfrisiaeque et vicinis quibusdam nationibus ex legibus praefuerunt*. He also worked on a Latin translation of Euripides' *Phoenissae* (1630) and on a commentary on the Bible, *Annotationes in Libros Evangeliorum*.

In 1625 Prince Maurice died, and in 1631 Grotius returned to Holland. After hot debate in the assembly and despite the intervention of Prince Frederick Henry of Orange, he was again threatened with arrest. In 1632 he went to Hamburg, then the centre of Franco-Swedish diplomatic relations. So great was his international prestige that the Swedish chancellor, Count A.G. Oxenstierna, offered him in 1634 the important position of Swedish ambassador in Paris. Grotius accepted; he wrote in his benefactor's honour a drama, *Sophompaneas* (1635), in which he relates the fortunes of the biblical Joseph as minister at the Egyptian court, which to him was a reflection of his own situation. The great Dutch poet Joost van den Vondel translated this drama as *Joseph in 't Hof* ("Joseph at the Court"). Grotius settled again in Paris but soon realized that he lacked the talents of a diplomat.

During the years 1636–37 he worked on the *Historia Gotthorum, Vandalorum et Langobardorum* ("History of the Goths, Vandals, and Lombards"), written in honour of Sweden. He also edited the works of Tacitus (1640). In 1644, when Queen Christina invited him to Sweden, he was received with great honour but nevertheless relieved of his post of ambassador. Although he was offered membership in the Swedish Council of State, he refused to settle in Sweden. On his way back to Paris he was shipwrecked on the coast of Pomerania and died of exhaustion at Rostock two days later, on August 28, 1645.

Appraisal of his work

Grotius' enormous gifts were those of a scholar rather than of an original thinker or creative genius. His best poetry was written in Latin; it was complex in both structure and content. His Dutch poetry, of which he wrote a great deal, was didactic in nature. As an historian he worked in the style of Tacitus; his *Annales et Historiae* is a masterpiece of Humanistic historiography. As a theologian his ideal was the early Christian community, and Grotius was not only in close contact with Protestant theologians but also counted Catholic priests among his many friends. A considerable body of his correspondence with eminent contemporaries has survived.

His enduring fame stems from his *De Jure Praedae* and *De Jure Belli ac Pacis*, which form the basis of modern international law. Their principal merit lay in their synthesis of the ideas of older writers and thinkers. His chief innovation was his insistence that nations are bound by natural law, which he considered to be independent of God and based on man's own nature.

BIBLIOGRAPHY. For a bibliography of the works of Grotius, see J. TER MEULEN and P.J.J. DIERMANSE, *Bibliographie des écrits imprimés de Hugo Grotius* (1950) and *Bibliographie des écrits sur Hugo Grotius, imprimés au XVIIe siècle* (1961). A recent collection of Grotius' correspondence is *Briefwisseling van Hugo Grotius*, 6 vol., ed. by P.C. MOLHUIJSEN, vol. 1–2 (1928–36) and by B.L. MEULENBROEK, vol. 3–6 (1961–67). An English translation of Grotius' *Adamus Exul* is in W. KIRKCONNELL, *The Celestial Cycle* (1952).

Editions of Grotius' works on jurisprudence include: R. FRUIN (ed.), *De iure praedae commentarius* (1868). This appeared in an English translation as: *De jure praedae commentarius. Commentary on the Law of Prize and Booty*, 2 vol. (1950). For the *De iure*, see P.C. MOLHUIJSEN (ed.), *Hugonis Grotii de Jure Belli ac Pacis libri tres* (1919). This was translated into English and appeared in "The Classics of International Law" with an introduction by J.B. SCOTT (1925). See also: C. VAN VOLLENHOVEN, *Verspreide geschriften*, pp. 349–602 (1934), and *The Framework of Grotius' Book De Jure Belli ac Pacis, 1625* (1931); G.N. CLARK and W.J.M. VAN EYSINGA (eds.), *The Colonial Conferences Between England and the Netherlands in 1613 and 1615*, 2 vol. (1940–51); and SIR HERSCH LAUTERPACHT, "The Grotian Tradition in International Law," *British Yearbook of International Law*, 23:1–53 (1946).

A major work of scholarship on Grotius is the edition of his Latin poems, *De Dichtwerken van H.G.* (1970–), which includes the Latin text, a Dutch translation, and a thorough commentary.

Biographies of Grotius are: W.S.M. KNIGHT, *The Life and Works of Hugo Grotius* (1925); and A. LYSEN (ed.), *Hugo Grotius: Essays on His Life and Works*, 2 vol. (1925).

(J.A.M.K.I.)

Groundwater

Vertical distribution and definitions

Groundwater is the water that occurs below the surface of the Earth, where it occupies all or part of the void spaces in a geological layer or layers. It is also called subsurface water to distinguish it from surface water, which flows overland and in rivers. Both surface and subsurface water are related through the hydrologic cycle, which is the path taken by the water on Earth: from oceans to atmosphere by evaporation, from atmosphere to the ground by precipitation, and ultimately back to the sea by runoff or streamflow.

The subsurface occurrence of groundwater classically has been divided into a zone of saturation, in which all interstices are filled with water, and a zone of aeration, in which the interstices are occupied partially by water and partially by air. The nomenclature associated with the vertical distribution of groundwater is shown in Figure 1.

The saturated zone often is separated from overlying soil water by an intermediate zone. The water in the intermediate and soil zones is known as suspended, or vadose, water. In arid regions the intermediate zone may be more than 300 metres (1,000 feet) thick, whereas in moist regions it may be absent. The intermediate zone terminates at its base in a capillary fringe, which may best be described as water in many minute, hairlike channels. In coarse-grained sediments the transition from the intermediate zone to the capillary fringe is abrupt,

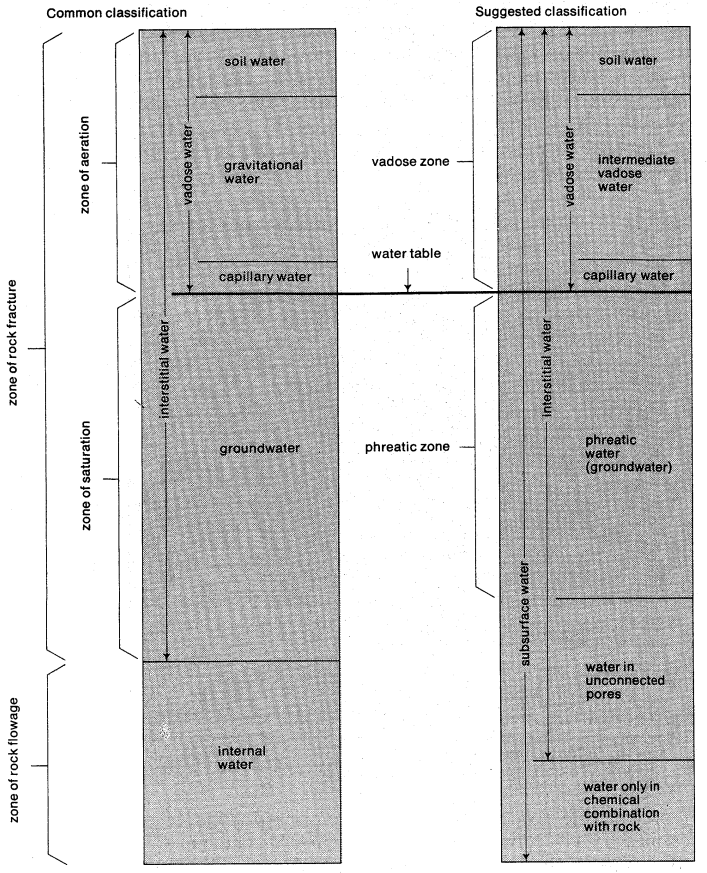


Figure 1: Classifications of subsurface water.
From S.N. Davis and R.J.M. DeWiest, *Hydrogeology* (Copyright 1966); used by permission of John Wiley & Sons, Inc.

but in silts and clays it is gradual. There may be very little difference in moisture content between the intermediate zone and the capillary fringe when recharge (percolation through the ground to the zone of saturation) occurs in fine-grained soils. The irregular surface of the capillary fringe constantly changes its position depending on change in water level and amount of recharge. Water movement in the upper part of the capillary fringe is slowed because of the presence of numerous pockets of air. In the lower part of the fringe, however, the soil is fully saturated, as it is below the water table—the theoretical surface approximated by the water levels in wells that penetrate the saturated zone. The water table separates the zone of groundwater, or phreatic water, from the capillary fringe. If groundwater flow is horizontal, water levels in wells will correspond closely to the water table. The groundwater zone merges at some depth into a zone of dense rock. Water is contained in the pores of such rock, but the pores are not connected and water will not migrate. The depth at which these zones will merge varies with the geological environment. In areas of crystalline intrusive and metamorphic rocks, which are essentially impermeable, the depth of the zone may start at 3,000 metres (10,000 feet), whereas in sedimentary basins the depth may be nearly 15,000 metres (50,000 feet). Temperature and pressure may become so great at depths of more than 30,000 metres (100,000 feet) that the pores are closed and the water will be present only in chemical combination with other material.

HISTORIC AND MODERN USE OF GROUNDWATER

The early development of wells and infiltration galleries in Asia arose because of the scarcity of water in dry regions, locally dense populations, and a dominance of agriculture. Mention of well water and well construction occurs in ancient literature and biblical accounts.

Man and animal power, aided by hoists and primitive hand tools, were the basis for well construction in the Near East. Despite great difficulties, these people built a number of large-diameter wells, some of which could accommodate donkey paths, although they rarely exceeded a depth of 50 metres (165 feet). Little evidence exists of technological advance in well drilling during historic times, although Egyptians perfected core drilling in stone quarries as early as 3000 BC. The ancient Chinese developed a churn drill for water wells that, although made of wood and powered by human hands, in principle was almost identical to modern machines. Wells of amazing depths were produced by using a slow drilling rate sustained for years and even decades. Depths as great as 1,500 metres (5,000 feet) have been reported. Brine and gas, rather than potable water, were obtained from the deepest wells. A slight modification of these methods is still used in rural areas of Laos, Cambodia, Thailand, Burma, and China.

Drilling for water in ancient times

The construction of long infiltration galleries, or *qanāt*s, was the greatest ancient achievement in groundwater utilization. *Qanāt*s collected water from alluvial fans; *i.e.*, sedimentary deposits laid down by streams at the bases of mountains and highlands. These fan-shaped structures, commonly several kilometres long (one kilometre equals 0.6 miles), collected water for agricultural and domestic purposes. The use of *qanāt*s, probably began about 2,500 years ago in Iran, from which the techniques of construction spread eastward to Afghanistan and westward to Egypt. It has been reported that one extensive *qanāt* system irrigated large tracts of fertile land west of the Nile in about 500 BC. Many *qanāt*s are currently used in Iran and Afghanistan, the best known of which are located on the alluvial fans of the Elburz Mountains in Iran (see further ALLUVIAL FANS).

A lack of early cultural contact between western Europe and China caused modern percussion methods of well drilling to develop independently in western Europe. Development stemmed from the discovery of flowing wells in Flanders in about AD 1100 and a few decades later in eastern England and northern Italy. In AD 1126 Carthusian monks from a convent near the village of Lillers, France, dug one of the first wells. In Gonnehem, Flanders, near Béthune, a water mill was driven by four wells cased (provided with interior piping or lining) 3.5 metres (11⅓ feet) above ground level and drilled several hundred feet deep to tap water under pressure from a fractured chalk bed that cropped out in the higher plateaus of the province of Artois. Wells from the region of Artois became so famous that all flowing wells are called artesian wells after them.

Many improvements have been made in the methods of drilling for water in the last 100 years because of understanding borrowed from oil- and gas-drilling experience. The development of hydraulic rotary methods has had the greatest significance in the advance of drilling techniques. Early rotary drilling was accomplished with the aid of an outer casing, which was dispensed with in about 1890 when thick mud was found to be sufficient for holding up the walls of the hole.

In the years between 1910 and 1930 well drilling was further advanced by the perfection of the deep-well turbine pump. Previously, deep wells had to be fitted with low-capacity piston pumps of poor efficiency. The new turbine pumps made irrigation by wells possible in areas previously underdeveloped for agriculture. As a result of the large production of these wells, bigger and more permanent wells are now in demand from the well-drilling industry.

Streams and lakes supply about four-fifths of the water used for all purposes in the United States (exclusive of hydroelectric power and navigation). Nonetheless, subsurface water plays an important economic role as well. Groundwater is more desirable than surface water for at least seven reasons: (1) it is commonly free of pathogenic organisms, and purification for domestic or industrial use is not necessary; (2) its temperature is nearly constant, which is advantageous if the water is used for heat exchange; (3) it is generally free of turbidity and colour; (4) its chemical composition usually is constant; (5) groundwater supplies are not seriously affected by short

Advantages and disadvantages of groundwater use

droughts; (6) most of it has not been affected by radiochemical and biological contamination; (7) it is available in many areas that do not have dependable surface water supplies because the groundwater has been stored by nature through many years of recharge.

Groundwater development is discouraged in some areas for at least three reasons: (1) many regions are underlain by rocks with insufficient porosity or permeability to yield much water to wells; (2) the content of dissolved solids is generally higher in groundwater than surface water in the same region; (3) well development is usually more costly than stream development, especially in regions of moderate to high precipitation.

SPECIAL IMPORTANCE OF GROUNDWATER IN ARID AREAS

Groundwater plays a vital role in the development of arid and semi-arid zones. Withdrawal eventually will deplete even large groundwater basins, but it is of great economic and social value if agriculture and industry thrive even for several decades. In some favourably situated areas within arid and semi-arid zones, groundwater and imported surface water already supply agricultural and domestic needs, although in vast areas further development in groundwater use is desirable.

Aquifers (water-bearing layers of rock) that antedate the formation of deserts will remain generally unaffected by an increase in aridity with the passage of time. In deserts the high temperature and general absence of organic matter suggest that small amounts of infiltrating water will not dissolve much carbon dioxide. In humid temperate regions the reverse is true and much carbon dioxide is dissolved, thus enhancing the chemical weathering (breakdown) of rocks. For this reason, and because of a general lack of moisture in deserts, chemical weathering will be slow, and the formation of solution openings in carbonate and other rocks will be retarded. The existence in some deserts of extensive carbonate aquifers may be attributed to former climates, considerably more moist than those presently prevailing. Other aquifers, such as those of the south flanks of the Atlas Mountains, may originate from the action of large amounts of water infiltrating into intake areas in well-watered uplands adjacent to the deserts.

Desert sedimentation and aquifers

The type of desert sedimentation is greatly influenced by desert conditions and in turn controls the type of aquifers found in unconsolidated (not yet cemented and transformed to rock) Cenozoic deposits (those formed in the last 65,000,000 years). Volcanic or tectonic obstructions will often alter stream courses and result in closed basins that fill with fine lake deposits mixed with saline residues in their central parts. Along the margins of such basins, or in deeper aquifers not affected by present arid conditions, are found the only freshwater or useful aquifers. The most important desert aquifers are formed from stream channel deposits. Desert channels carry such large amounts of material in suspension and traction that flows may become viscous mixtures of mud and debris, which is not true of streams in humid regions. Resulting deposits are usually poorly sorted and of low permeability. Permeable zones develop where streamflow persists long enough to sort the coarser debris and usually occurs during the declining stages of streamflow, when the stream is confined to narrow parts of the channel bottom. Consequently, permeable zones represent only a fraction of the alluvial deposits and are not easy to locate by drilling.

Most of the important desert aquifers of the southwestern United States, northwestern Mexico, northern Chile, and parts of Central Asia are composed of nonindurated Cenozoic deposits formed by volcanism and block faulting (vertical movements along fractures that bound block-shaped mountains). In most other desert regions, however, erosion has been the dominant feature. As a result, bedrock is near the surface, and nonindurated aquifers occur only in shallow deposits along ephemeral streams (those flowing only occasionally), basal parts of very large sand dune tracts, and wind-eroded basins. Extensive consolidated and semiconsolidated aquifers are found in some desert basins that lack thick alluvial deposits. Two examples of this are the Mesozoic sandstones (ranging from 65,000,000 to 225,000,000 years in age) of the Great Artesian Basin of Australia and the Nubian sandstone of Egypt and parts of adjoining countries (Figure 2). Water contained in these systems ranges

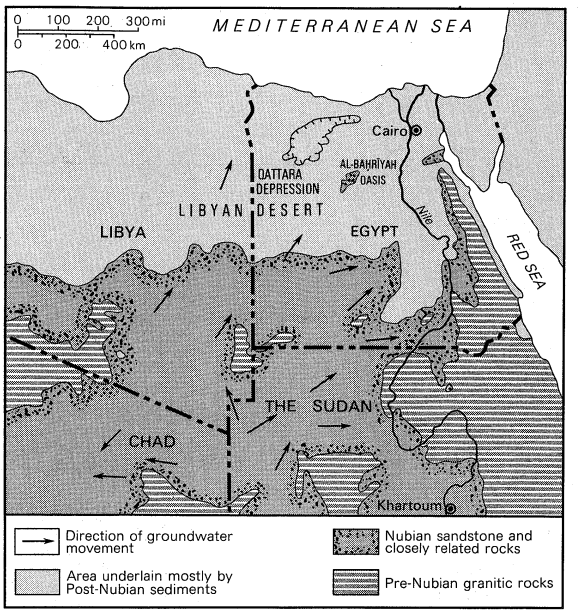


Figure 2: Outcrop area and general direction of groundwater movement in sandstones, northeastern Africa.

from potable, in parts within a few hundred miles of recharge areas, to brackish, in the deeper buried parts of the aquifers. Oases from the Nubian aquifers have existed for countless centuries at Wāḥāt ad-Dākhilah, Wāḥat al-Farafirah, and al-Wāhāt al-Baḥrīyah. Despite the fact that modern wells have increased water discharge and thus caused some reduction of the hydraulic head (due to pressures that result from the height of water in a well), the total amount of water available from the Nubian aquifers is so great that an increase in the number of wells may be justified if the water can be utilized profitably.

Soil moisture, infiltration, and recharge

The quantity and quality of available groundwater in arid regions affect the types of hydrogeologic features that are found. When rainfall is less than 125 to 250 millimetres (five to ten inches), passage of recharge water will be blocked by most soils, depending on the permeability of the soil, its retention of water, and the distribution of rainfall in relation to the temperature. During the summer heat, a loss of soil moisture to a depth of two feet in soils with a specific retention of 15 percent will require nine centimetres (3.6 inches) of rain merely to compensate for the soil moisture deficiency. More than nine centimetres will be needed for recharge if the rain occurs at several different times during the year, causing loss of water from the intervening periods of dry weather. On the other hand, if the sand in a sand dune has a specific retention of less than 5 percent, rainfall of five centimetres will penetrate more than three feet. In this case, penetration probably will go beyond the zone of seasonal drying in some parts of the dune.

Recent work in the western part of the Central Valley of California has proved (at least in this semiarid region) that rainfall does not infiltrate as a saturated line or front. The original porosity of the clay-rich and mudflow deposits found in a number of small alluvial fans in this region is so high that the sediment is unstable under the slightest overburden load. Experiments with flooded test plots and studies of the moisture content of numerous cores have shown that saturation of the porous deposits has not occurred since burial. The dry state extends to more than 30 metres (100 feet) below the surface in some places. Even if the rate of deposition was fairly rapid, thousands of years were required to deposit

these sediments. The subsurface zones of these small alluvial fans, therefore, have not been infiltrated by rain during this time span. The mean annual rainfall in this region today is about seven inches, which is greater than the rainfall in extremely arid regions. It may be concluded, therefore, that in desert soils with an appreciable amount of clay, considerable infiltration does not occur.

Typical barren rock or hard clay surfaces in deserts will shed rains of only a few tenths of an inch. Large amounts of soil and surface debris usually are eroded by runoff, and the eroded material varies from very turbid water to a viscous mudflow. In rare instances, the runoff may be relatively free of suspended material. If turbidity is low and permeability of the channel bottom is deep enough, water will flow into the gravel and sands of the channel bottom. Evaporation will restore much of this water to the atmosphere. Vigorous infiltration may recharge underlying aquifers, predominantly in constricted portions of the channel rather than in portions in which the flow covers large sections of the desert floor. This is caused by the greater concentration of permeable material at the constrictions, which are also most likely to receive the last part of the flood flow; the latter is made up of water draining from the earlier deposited saturated floor debris and is relatively free from suspended sediments.

The great depths to water in upland areas and the exceptionally flat hydraulic gradients (pressure changes with distance and direction) that usually are found in deserts reflect the small amounts of recharge that take place. When the subsurface layers are undisturbed after deposition, the gradients may be so small that it is not possible to distinguish between the important hydrogeologic barriers. The amount of water moving through the subsurface increases, and all gradients are steepened by groundwater development, which accentuates the zones of lower permeability.

Dealt with in this article are the factors that influence the origin, occurrence, and distribution of groundwater, the fluctuation of groundwater levels and abundance, and its movement in the subsurface. For further information on the role of groundwater in the hydrologic cycle, see HYDROLOGIC CYCLE. See also CLIMATE; CLIMATIC CHANGE; RIVERS AND RIVER SYSTEMS; ESTUARIES; CAVES AND CAVE SYSTEMS for information on the relations between these topics and groundwater.

ORIGIN AND OCCURRENCE OF GROUNDWATER

Hydrologic conditions and their influence. Groundwater and surface water are related through the hydrologic cycle, the path traveled by the waters of the Earth, of which they are interdependent components. Water evaporates from the oceans and forms clouds, which move inland, condense, dropping the water to the Earth as precipitation. From the land, through river channels and underground, the water runs off to the oceans. There is no evidence that water decreases in quantity at a global level; it is neither destroyed nor generated.

Evaporation, precipitation, and runoff

A large amount of the water that falls upon the Earth is returned to the atmosphere as vapour through the combined actions of evaporation, transpiration (evaporation via plant life), and sublimation (change of state from solid to vapour). On a global basis the total amount of precipitation on the continents P_c and on the oceans P_o is equal to the total amount of evapotranspiration from the continents E_c and from the oceans E_o. On a continental basis the total precipitation P_c is equal to the evapotranspiration from the continents E_c and the runoff R, a relationship that may be expressed: $P_c = E_c + R$. For all continents the average values are 674 millimetres precipitation = 420 millimetres evapotranspiration + 254 millimetres runoff.

The term runoff is usually taken to be synonymous with streamflow and is the sum of surface runoff and groundwater flow that reaches stream channels. Because groundwater contributes the base flow, or dry-weather flow of rivers, groundwater studies must be related to the runoff cycle. A study of the interrelationship between surface water and groundwater is always indicated, and in many projects, measurements of surface discharge provide the only indications of amounts of groundwater available. Surface runoff equals precipitation minus surface retention and infiltration. Infiltration is the passage or movement of water through the surface of the soil and is to be distinguished from groundwater flow.

Factors influencing surface runoff include precipitation intensity, permeability of the ground surface, duration of precipitation, type of vegetation, area of drainage basin, distribution of precipitation, stream-channel geometry, depth to water table, and the slope of the land surface. Surface runoff is commonly represented in the form of a hydrograph, which is a time record of stream-surface elevation or stream discharge at a particular channel cross section. In general, a hydrograph is a plot of the discharge from a hydraulic or hydrologic unit or system, such as a river or drainage basin, versus time, or it may depict the time variation of the runoff component of a storm. In the case of perennial streams, in periods of drought when no direct overland flow reaches the river, the hydrograph is a line that slopes gently down. The streamflow at this time is provided entirely by groundwater flow.

Infiltration and its effects

When a storm begins, much precipitation is caught and stored by trees and vegetation as interception, but because water stored in this way is usually exposed to wind and open to evaporation, storms of light intensity and short duration may be entirely depleted by interception and by the small amount of water that infiltrates through the surface and fills puddles and surface depressions. For water to infiltrate, a proper soil surface is necessary. Overland flow starts when the available interception and depression storage is exhausted and when the rainfall intensity of the storm at the soil surface exceeds the infiltration capacity. A thin sheet of water, known as the surface detention, then covers the soil surface. When the overland flow reaches a stream channel, it is called surface runoff.

Part of the water that infiltrates the soil will reach the stream channel by continuing a lateral flow (interflow) at shallow depths because of the presence of impervious horizons just below the soil surface. Another part will percolate downward to the water table and provide the base flow of the stream when it reaches the stream channel. The last part will remain above the water table in the zone of unsaturated flow.

When a large storm occurs in the drainage basin of a perennial stream and conditions are such that the rainfall intensity exceeds the rate of infiltration and the volume of infiltrated water exceeds the soil-moisture deficiency, then the hydrograph of the river discharge becomes quite peaked. There is above-normal streamflow because of the contributions from channel precipitation, surface runoff, interflow, and groundwater flow. The contribution from groundwater flow may be negative, however, when continued high runoff causes the river to become influent—surface water flowing from the channel down gradient, to contribute to the groundwater.

It is very difficult to separate the base flow from surface runoff because streamflow measurements do not reveal the times when a stream becomes influent; that is, contributes water to the adjacent aquifers. It is possible to make an approximate separation that is acceptable for most purposes, however. This is based on a horizontal line drawn through the point where the rising limb of the hydrograph starts and its intersection point with the recession curve. The groundwater-flow (base-flow) component is the volume of water given by the area below the horizontal line.

Geological factors. Geological factors, principally the porosity (relative void space) and permeability (a measure of the ease of flow) of water-bearing rock layers, are of great importance in determining groundwater occurrences.

The porous rock and soil media, in which water is collected by water-bearing strata and through which it flows under the influence of various forces, have a solid matrix or skeleton that is an assemblage of solid mineral

grains separated and surrounded by voids, pores, or interstices. These may be filled with water, gases, or organic matter. The term rock commonly applies to a cemented or consolidated porous medium, such as a sandstone or limestone, whereas soil or sediment refers to unconsolidated, uncemented material, such as sand, gravel, or clay.

Porosity is defined as the ratio of the volume of voids to the total volume of the material under consideration. Porosity of unconsolidated materials is determined by the packing of the grains, their shape, arrangement, and size distribution; porosity of consolidated materials is determined by the degree of cementation, the state of solution, and fracturing of the rock. Openings left between grains will be filled by smaller grains, and thus the porosity associated with nonuniform size distribution is smaller than with well-sorted grains.

Porosity is a measure of the water-bearing capacity of a formation and is also a factor in the ability of a formation to transmit water. This ability is known as the hydraulic conductivity of a formation. The relationship between porosity and hydraulic conductivity is complex, and other factors are also involved. For example, large rounded or angular sand grains may have a smaller porosity than small platelike clay particles. This is because of the larger specific surface of the plate shape, which causes high molecular forces of attraction between the clay and water particles.

Porosity is the sum of effective porosity and specific retention. The specific retention is that percentage of porosity that is occupied by water that cannot be drained by gravity from the rock. The hydraulic conductivity K of an aquifer is equal to the permeability of the medium times the unit weight of the water divided by its dynamic viscosity. This may be represented: $K = k\frac{\gamma}{\mu}$, in which γ is the specific weight of water, μ the dynamic viscosity of water, and k the intrinsic permeability of the medium. The intrinsic permeability is characteristic of the solid matrix (*e.g.*, the material that binds together the particles in a sandstone) of the medium alone. It has the dimensions of a surface (that formed by the interstices between particles) and takes into account effects of stratification, packing, arrangement of grains, size distribution, and porosity.

The hydraulic conductivity is expressed as a velocity of flow. In the metric system it may be expressed in centimetres per second; in English units it may be expressed in feet per second or in gallons per day per square foot of surface.

The so-called darcy unit, which is used in petroleum engineering as well as in hydrogeology, is used as a measure of the intrinsic permeability k. It equals 0.987 $\times 10^{-8}$ square centimetres (1.062×10^{-11} square feet).

Aquifers, aquicludes, and aquitards

Geological rock formations of relevance in groundwater studies are defined according to their water-bearing and water-transmitting properties. Thus, an aquifer is a geological formation containing, or capable of containing, water in its voids or pores that may be removed economically and used as a source of water supply. Unconsolidated alluvial deposits of sand and gravel and porous sandstones are examples of water-bearing formations.

A geological formation so impervious that for all practical purposes it completely obstructs the flow of groundwater (although it may be saturated with water itself) and completely confines other strata with which it alternates in deposition is an aquiclude. A shale is an example.

If a geological formation is impervious and semiconfining and transmits water at a very slow rate compared with the aquifer, it is called an aquitard. Over a large area of contact, however, such a formation may permit the passage of large amounts of water between adjacent aquifers, which it separates from each other. Clay lenses interbedded with sands, if thin enough, may form aquitards. If a rock formation neither transmits nor stores water, it is called an aquifuge.

When the surface of the saturated groundwater zone is in direct contact with the atmosphere through open spaces in permeable material, it is called free, and the groundwater is unconfined. Aquifers containing water in these conditions are called unconfined *sensu lato*. Confined water is separated from the atmosphere by impermeable material. Aquifers containing such water are called confined *sensu lato*. The division between confined and unconfined water is entirely a matter of gradation, and the term semiconfined is used for intermediate conditions.

Confined water is also called artesian water; this term, however, was first applied to water under sufficient pressure to produce flowing wells. In recent years artesian has been used more or less as a synonym for confined water.

World distribution of groundwater. A vast amount of groundwater is distributed throughout the world, and a large number of groundwater reservoirs are still underdeveloped or uninvestigated. In September 1968 the Water Resources Section of the United Nations invited a group of experts to study groundwater storage at the global level. The panel of experts, together with the technical advisers of the UN Resources and Transport Division, drafted a manual on groundwater storage and presented 70 case studies, constituting a sampling of occurrence and world distribution of groundwater. This sampling does not reflect the overall world situation, however, because the cases presented apply principally to areas that have been carefully investigated in response to a shortage of water resources or to a high demand for water. The number of cases cited do suffice to give some general impression of the distribution of groundwater according to climate, physiography, and geology (see Table 1).

Table 1: Distribution of Groundwater

	number of cases
Climatic distribution	
Type of climate	
Arid or semi-arid	16
Tropical humid	9
Cold or temperate, humid	7
Desertic	5
Temperate (oceanic, Mediterranean, etc.)	25
Continental	8
Physiographic distribution	
Physiographic characteristics	
Alluvial valleys	19
Coastal plains	18
Deltas	2
Islands	3
Inland plains (fluvoglacial and morainic)	5
Other types of inland plains and deserts	14
Plateaus	4
Mountains (mainly carbonate karstic rock)	5
Geographic distribution	
Country or continent	
Americas and Caribbean	9
Africa and Madagascar	12
Europe	21
Middle East and Central Asia	13
Asia, Far East, and the Pacific	15
Geological distribution	
Type of rock	
Unconsolidated sand and gravel	51
Carbonate (dolomite, limestone)	26
Sandstone	14
Volcanic	4

The most frequently investigated or exploited groundwater reservoirs are of the unconsolidated clastic (mainly sand and gravel) or carbonate hard-rock type of alluvial valleys and coastal plains under temperate or arid conditions.

FLUCTUATION OF GROUNDWATER LEVELS

Phreatophytes and daily fluctuations. Phreatophytes are desert plants that grow along streams and in areas with relatively shallow water tables. They have deeply penetrating roots that reach the groundwater and thus permit plant survival. Some phreatophytes are valuable guides to the presence of potable water in arid and semiarid regions because they have a low tolerance for salt. The ash, alder, willow, cottonwood, and aspen have been useful in this way and generally grow where the water table is less than ten metres (30 feet) deep.

Water consumption by phreatophytes

Most of the phreatophytes have little or no economic value and waste a large quantity of water in arid regions. Alfalfa, *Medicago sativa*, which is a widespread phreatophyte of great economic value, is an exception to most other phreatophytes. In the western United States alone, amounts of about 30,000,000,000 cubic metres (about 25,000,000 acre-feet) of water are wasted by phreatophytes each year. The loss of water via phreatophytes in the arid and semiarid regions of the world is many times this quantity.

Various factors influence the amount of water used by phreatophytes, such as plant species, density of growth, climate, and hydrology. Sunlight, temperature, and humidity are the most important climatic factors. If there is no sunlight or if the temperature drops below about 4° C (40° F), transpiration in some plants will almost stop. In other plants, transpiration will be slight during the night but will stop when the mean temperature drops below 18° C (65° F). Plants will tend to transpire less if the relative humidity is high.

The most important factors are depth to water and water quality. For many phreatophytes, water consumption and depth to the water table are inversely related. Thus, if the water table would rise from a depth of 40 centimetres (16 inches) to a depth of 70 centimetres (28 inches), water use might be doubled. The effect of water quality varies in different plants. Most plants such as salt grass will have an optimum growth with water that has dissolved-solids concentration of several thousand parts per million.

Water-level fluctuations in shallow wells are related to plant transpiration. The nightly rises in water levels signify recovery of water at times of no transpiration. In view of the fact that the water fluctuations are small compared with the total energy moving the water and the thickness of the saturated material, the maximum velocity of water rise in the hydrograph is a close approximation to the velocity of water moving continuously into the area of water withdrawal. Water use, therefore, is approximately equal to this maximum velocity, multiplied by the total area of open space within the saturated material.

Water use by dense plant growth ranges from less than one foot per year in subarctic environments to more than 2.3 metres (7.5 feet) per year in hot, dry environments. Salt cedar, *Tamarix gallica*, which is native to Europe and Asia, uses the highest quantities of water. Tests conducted in Arizona showed that this plant consumed 270 centimetres (110 inches) of water from a water table at a depth of 123 centimetres (48 inches). Representative transpiration values for various phreatophytes are given in Table 2.

Table 2: Water Use by Dense Growths of Phreatophytes

common name	scientific name	climate	depth to water		annual water consumption	
			cm	in.	cm	in.
Salt cedar	*Tamarix gallica*	hot, dry	123	48	270	110
		hot, dry	213	84	224	88
Greasewood	*Sacrobatus vermiculatus*	cool, dry	50	20	66	26
Willow	*Salix*	hot, dry	61	24	134	53
Cottonwood	*Populus*	hot, dry	220	87	238	94
Alfalfa	*Medicago sativa*	cool, dry	91	36	80	31
		hot, dry	138	54	113	44
Alder	*Alnus*	hot, dry	...	...	162	64

Current wastage of groundwater by useless vegetation can be avoided by various means. When the water is very shallow, the water table can be lowered by drainage wells, which will kill most grasses and small shrubs that are phreatophytes. The water that is drained can then be used for other practical purposes. Weed killers, burning, or mechanical cleaning will remove other useless phreatophytes, and the area can be planted with useful phreatophytes, such as alfalfa. Removal can be quite difficult, however, because of the great growth rate of some of the noxious plants. One example is the salt cedar, which can produce more than 500,000 seeds per plant per year, many of which will germinate rapidly and produce a jungle-like growth in less than five years.

Climatic variations. Short-term climatic variations of groundwater levels are those extending over periods of several years or more. Alternating series of wet and dry years, in which the rainfall is above or below the mean value, will produce long-period fluctuations of levels. The long periods of rainfall and groundwater levels in San Bernardino Valley, California, shown in Figure 3, illustrate this point. Rainfall is not an accurate indicator of groundwater level changes. Correlation of the monthly precipitation on two small drainage basins, Trent River and Swift Creek in North Carolina, with monthly discharge of these rivers was undertaken for the 15-year period 1952–66. In this application of the interrelationship between surface water and groundwater, measurements of surface-water discharge were made to indicate the quantity of available groundwater. It is assumed that all base flow of a river is contributed by its groundwater basin and that this water, released from the basin, can be replenished by recharge through rainfall. If there were not enough recharge, the base flow would decrease over a number of years, and this would be reflected in the annual trend of the streamflow records. If no such trend can be observed, it may be assumed that the base flow is replaced by recharge.

From D.K. Todd, *Ground-Water Hydrology* (copyright 1959); used by permission of John Wiley and Sons, Inc.

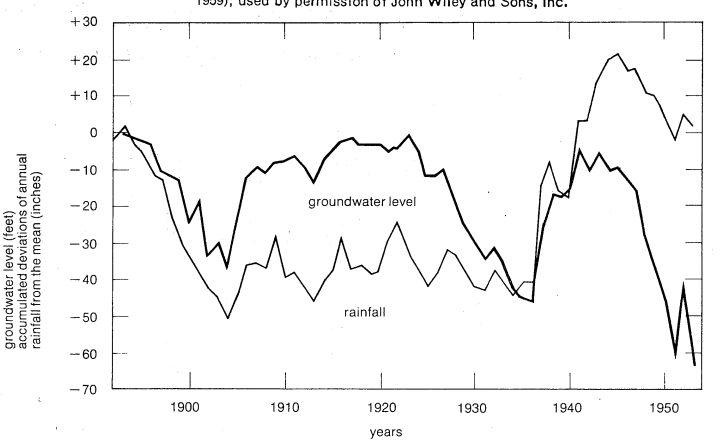


Figure 3: Variation of annual groundwater level and annual rainfall in San Bernardino Valley, California.

In the case of the Trent River, base flow rose from about 14 percent of the average precipitation for a five-year period to 20 percent of the average precipitation for a ten-year period, when the average precipitation increased 7 percent from one period to the other. In the case of Swift Creek, base flow increased from 10 to 19 percent of the average precipitation when the precipitation increased by 6 percent. Thus, groundwater levels reflect short-term climatic variations.

MOVEMENT OF GROUNDWATER

Darcy's law. Groundwater moves from levels of higher energy to levels of lower energy, and its energy is essentially the result of elevation and pressure. Groundwater flow is amenable to a classical mathematical treatment only if the flow is laminar (nonturbulent), as in unconsolidated sediments and in some sedi-

mentary rocks. For water movement in karstic terrain (limestone media, with caves and sinks), in which the flow is often turbulent, only approximative mathematical formulations or empirical expressions based upon hydraulic experiments are currently available. For a mathematical expression of groundwater motion in the subsurface, the interested reader should consult the literature cited in the bibliography of this article.

In laminar flow, kinetic energy of groundwater is neglected because groundwater velocities are very small. During flow, groundwater experiences a frictional loss in energy caused by contact with the walls of the granular medium along its seepage path. The water lost per unit length of distance travelled, called the hydraulic gradient, is proportional to the velocity of groundwater for laminar flow in sandy aquifers or seepage through earth embankments. The proportionality of hydraulic gradient and groundwater can be expressed by a linear law of flow called Darcy's law.

Similarity to flow in pipes

The similarity between groundwater flow and laminar pipe flow was recognized in 1856 by Henri Darcy, a French engineer, who conducted an experiment on a vertical cross-sectional area A filled with sand under conditions simulated by Figure 4. Darcy concluded from

From M.K. Hubbert, *Journal of Geology* (copyright 1940); University of Chicago Press

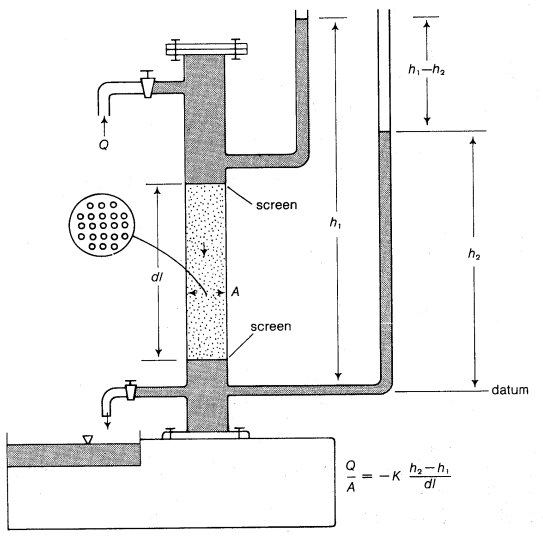


Figure 4: Apparatus demonstrating Darcy's law (see text).

his study of the flow through horizontally stratified beds of sand that the flow rate Q was directly proportional to the energy loss, inversely proportional to the length of the flow path, and proportional to a coefficient K that depends on the type of sand. The coefficient K is also subject to the properties of the fluid, however. Darcy's law may be expressed as $Q = -KA(dh/dl)$, in which dh/dl = head lost over a length of porous medium dl, *and* $K = k(\gamma/\mu)$, in which K is the hydraulic conductivity as before; γ is the specific weight of the fluid, μ is the dynamic viscosity of the fluid, and k is the intrinsic permeability of the medium. If the granular skeleton of the porous medium were a simple geometrical assembly of prismatic, unconnected tubes, groundwater flow could be dealt with microscopically by the laws of hydrodynamics. In fact, however, the seepage path is a tortuous one, branching into a multitude of tributaries. With the introduction of a twofold averaging macroscopic concept, Darcy's law avoids the insurmountable difficulties of the hydrodynamic microscopic picture.

The law is based on consideration of a fictitious flow velocity, the Darcy velocity or specific discharge, through a determined cross section of a porous medium rather than the true velocity between the grains. It also considers average hydraulic values rather than local values of this velocity. These concepts were simplified because of the nature of Darcy's experiment, which permitted the measurement of only average hydraulic values in the sand-filled cylindrical pipe. Because Darcy's law reflects average conditions, it is essentially a statistical result.

Flow nets. A widely used graphical method consists of fitting a flow net to the boundary conditions that define a given groundwater problem. A flow net is a two-dimensional graph consisting of two families of curves of a distinctive nature: flow lines or streamlines that show how water travels, and equipotential lines that connect points of the same potential. Use of the graph is thus limited to the investigation of two-dimensional cross sections that represent the main flow and to the analysis of three-dimensional problems with either axial or radial symmetry.

Streamlines are everywhere at right angles to equipotential lines, and flow nets are particularly useful when observations of static water levels are made in closely spaced wells. If the heads of water in the wells are equal to the elevation of the water level in the wells, it is possible to construct groundwater contours from an observation of these water levels. This permits estimation of the direction of groundwater flow, which is orthogonal (at right angles) to the family of constructed contour lines.

Oil and gas associations. The anticlinal theory was the first attempt to establish laws of oil and gas accumulation in subterranean media in the early days after the discovery of oil at Titusville, Pennsylvania, in 1859. It conceived of a static system of gas, oil, and water with horizontal interfaces, resting on the crests of anticlines. This theory was generally accepted in the first half of the 20th century, except for some sporadic thinking about the requirements of fluid dynamics in the 1920s and 1930s, which made the flow of water an essential requirement for oil and gas movement. In 1953 there was renewed emphasis on the hydrodynamic aspect of the problem, which is basically the correct approach. The most prominent features in this approach are the concepts that force potentials act on the fluids in question and that the fluids involved are immiscible. This leads to the occurrence of distinct fluid–fluid interfaces, along which surface-tension effects are important.

A water-saturated environment is the normal site for petroleum and gas, formed by the decomposition of organic matter deposited in sedimentary rocks. For convenience because of this, the potential energy of these fluids is expressed in terms of that of the surrounding groundwater. In the study of oil and gas movement in porous rocks, a distinction is made between primary migration from the source rocks, in which the fluids originate in a dispersed state, to the reservoir rocks and the movement that occurs in them.

The force potential per unit mass for liquid petroleum at a given point includes the work done against gravity and pressure and that involving the energy between petroleum, water, and rock, whereas these substances meet to form an interface or surface (Figure 5). In general, considering oil and water, rocks are preferentially wet by water, and the pressure inside the oil exceeds the pressure inside the water. This is important in the primary migration of oil from source rocks to reservoir rocks, which is related to the geometry of the porous medium (overall configuration of the rock body) and the pressure gradient across the boundary of the different rocks. Discontinuities in permeability are the paramount driving force in migration.

The force acting on the liquid petroleum is oriented along the direction in which the grain size of the rock increases most rapidly. This force is inversely proportional to the grain size of the rock. Because this increase in grain size normally occurs perpendicularly to the planes of sedimentary deposition, the oil globules have a tendency to move in that direction under the influence of capillarity. Once a globule reaches the boundary between rocks of different texture, part of its mass is subjected to a capillary force. It follows that an oil globule, in a liquid or gaseous state in a water-saturated environment, can pass only from fine- to coarse-textured rocks and not in the opposite direction. Capillary action would not be present if oil were not surrounded by water; hence, the importance of groundwater in oil and gas migration.

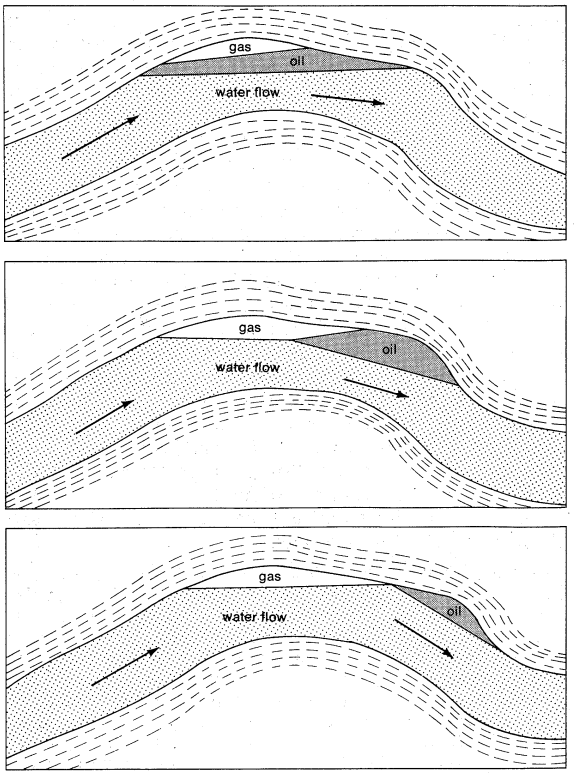


Figure 5: Hydrodynamic oil and gas accumulation in gently folded thick sand, showing changes in oil–gas–water interfaces with water flow through time (top is earliest).
From M.K. Hubbert, *Journal of Geology* (copyright 1940); University of Chicago Press

Movement in reservoir rocks

In the flow of oil within reservoir rocks, the role of capillary forces may be neglected because of the large grain size. The capillary pressure of oil in sandstone is only of the order of tenths of an atmosphere (one atmosphere is atmospheric pressure at sea level, or 14.70 pounds per square inch). The impelling force for a particular flow of the ambient (surrounding) water is the variation in density of the hydrocarbons present. Density variation also is responsible for the direction of the total force and for the separation of liquid petroleum and gas under particular geological conditions. Depending upon the angle of dip or inclination of the reservoir strata, gas may be deflected upward along the inclined strata, whereas liquid petroleum will move downward. A change in the magnitude of the impelling or driving force could change this situation and cause both fluids to migrate in the same direction. It is clear that elements of petroleum generally will move from regions of higher energy to regions of lower energy and will come to rest when the energy of their surroundings is higher than their own energy or when they are trapped between regions of higher energy and impermeable barriers. Oil and gas accumulations in such traps will form a static horizontal interface; the oil–water interface will be sloping if water is flowing underneath the oil.

Saline water in coastal areas. Saltwater encroachment or intrusion is the shoreward movement of sea or ocean water into confined or unconfined coastal aquifers and results in the displacement of freshwater from these aquifers. If freshwater and salt water are considered as two immiscible fluids, then they are separated by a sharp interface with a slope somewhat like that of an oil–water interface. Because the interface slopes, the front of the salt water can be compared to a tongue, progressing into the land as a result of overdraft of the overlying freshwater and pushed back seaward when precipitation replenishes the amount of freshwater. Actually, freshwater and seawater mix to form a region of dispersed water, and saltwater dispersion should be taken into account in order to obtain a more accurate concept of the nature of groundwater flow in a coastal aquifer. When the water supply of islands and coastal regions depends upon groundwater, the study of saltwater encroachment, which may reach several thousand feet inland, is of great interest and importance (Figure 6). Wells that become contaminated with salt water may either be abandoned or

From M.K. Hubbert, *Journal of Geology* (copyright 1940); University of Chicago Press

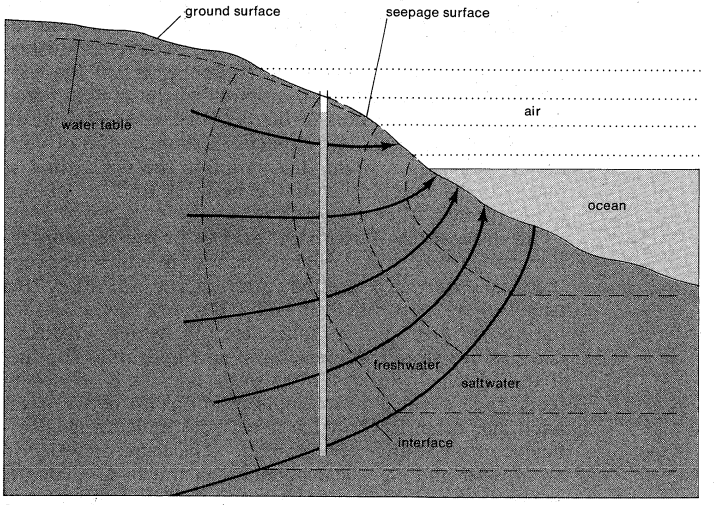


Figure 6: Saltwater encroachment (dashed seepage surfaces) and freshwater flows (arrows) as related to wells in coastal regions.

have to be injected with freshwater to stop the inland movement of the salt water and to establish a freshwater barrier. An idea of the freshwater losses to the ocean through discharge under sea level might be obtained by a study of the location of an idealized interface.

Physical analogue models

Analogue models of groundwater systems. One of the physical analogues currently used in groundwater studies is the sandbox. It represents a true model because an aquifer and the model both involve flow through porous media. A sand model is a scale model of an aquifer with the boundaries scaled down and the permeability modified. Unconfined aquifers can be modelled with the water table naturally reproducing the upper boundary; confined aquifers are reproduced by providing an impervious layer so that pressure can be applied. The effect of capillary rise is greatly exaggerated in this kind of analogue, however, and it lacks versatility. For this reason, other models, such as the viscous-flow analogue and the electric analogue, are favoured. The sandbox is still in use in soil mechanics and agricultural laboratories.

In 1899 it was shown that pressure and gravity forces can cause a viscous liquid to flow between closely spaced parallel plates. The first apparatus used to demonstrate this flow was designed in 1897, and since then the analogue has been used extensively to simulate the two-dimensional laminar flow of water through porous soil. Theoretically, such a viscous-flow analogue is a perfect model because it conforms to the equations of flow that describe groundwater movement. One requirement, however, is that the density of the fluid remain constant, and this is difficult to satisfy because of the problem of keeping the temperature of the fluid constant and because the viscosity of most oils and glycerine is very much temperature dependent. Analogues utilizing such fluids are usually in temperature-controlled rooms, and, if the fluid is pumped on some occasions it is necessary to cool the fluid after it leaves the pump.

The versatility of the parallel-plate, viscous-flow analogue as a tool for the investigation of groundwater flow lies in the easy simulation of different values of the hydraulic conductivity of the soil through variation of the spacing of the plates and the fluid properties. Regional variations of hydraulic conductivity may be accounted for by the insertion of thin strips to reduce the spacing in the desired region.

The model also has been used with the plates oriented horizontally to study the effects on the groundwater table of artificial replenishment in the well fields of the water works company of Zealand Flanders. For the analogue to

be suitable for the study of unconfined aquifers, it must be assumed that groundwater-level fluctuations are relatively small compared with the thickness of the aquifer.

Electric analogue models

Electrical models involve an R–C (resistor–capacitor) analogue network in which solutions to the equations describing groundwater flow may be obtained by varying current and voltage; the solutions can be visualized directly on an oscilloscope. The use of electric-circuit elements to simulate the distributed properties of the subsurface strata is only for the points of the medium that correspond to the points of the network, consisting of an assembly of resistors and capacitors. Application of voltage and current sources that simulate the excitations is made to the boundaries of the network and at interior nodes (points) when required.

The dissipation of electrical energy in the resistor–capacitance network is similar to the way in which a porous medium consumes groundwater energy to allow the travel of water through its voids. Because of the similar nature of electrical conductance, in which electrical charges are stored in capacitors, and hydraulic conductivity, in which water storage in an aquifer is related to its storage coefficient, a simple relationship between capacitance of an R–C analogue network and storage coefficient of an aquifer can be made. The head of water in an aquifer and electrical potential in the R–C analogue network are analogous. Hence, the elevation of water in a well can be obtained by application of Ohm's law (current equals voltage divided by resistance); consequently, electrical R–C analogue models are of great utility, and their use in studies of current or future groundwater conditions is widespread.

Other electrical analogues also have been used. The electrical resistance network, employing a solid sheet made of a conductive material in a tank filled with an electrolyte (conducting fluid solution), has been a favourite tool of agricultural engineers and soil physicists in studies of drainage and infiltration. The voltage distributions in the sheet and in the electrolyte simulate the conditions of groundwater flow.

The electrolytic tank is usually a watertight container of a nonconductive material filled to a shallow depth with a few centimetres of electrolyte. The following model conditions must prevail: the conducting liquid should have no electrical reactance, its resistivity must be uniform, there must be a linear relation between voltage and current, chemical reactions between liquid and electrodes must not occur, and the rate of evaporation of the liquid must be slow to prevent change of the resistivity in time.

The principle governing the conductive-sheet analogue and the one underlying the electrolytic tank are the same. Sheets of electrically conducting material of uniform properties in all directions and high resistance are suitable. This material may consist of filter paper soaked in a suspension of colloidal graphite, of woven grids of metal wire and silk thread, of conductive rubber, or of metallized paper, among other things.

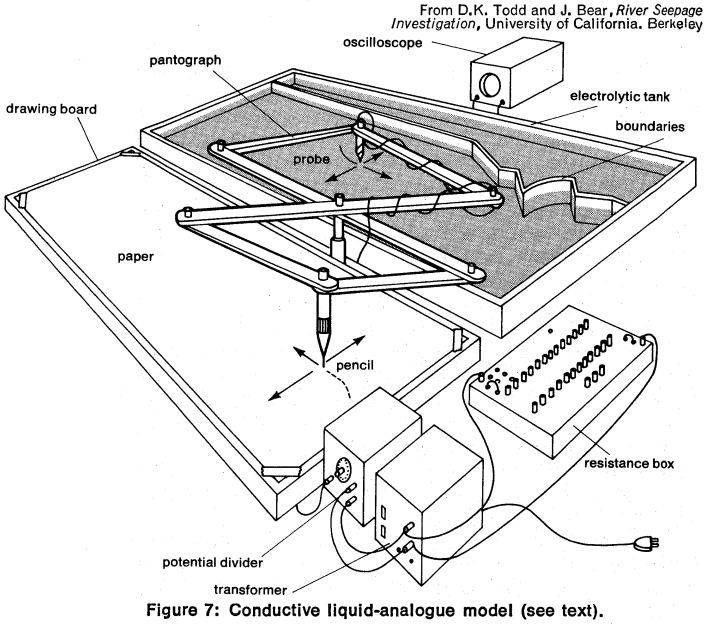


Figure 7: Conductive liquid-analogue model (see text).

Shown in Figure 7 is a typical liquid-analogue model consisting of the electrolytic tank, the electrical equipment, and the drawing board with pantograph. Voltage is reduced, as in the case of the resistance-network analogue, to avoid electrocution and heating of the electrolyte. The voltage is divided into a large number of subdivisions by a potential divider. The identification of the potential at the probe position is made with the help of an oscilloscope through which no current flows when the probe is of the desired potential and the pencil of the pantograph is depressed to leave a mark on the drawing paper. By connecting points of the same potential, equipotential lines are sketched. The application of the theory of the electrolytic tank, which has been known for a long time, to groundwater flow is based on the similarity between the differential equations that describe groundwater flow and those that govern electrical current flow through conducting materials.

BIBLIOGRAPHY. The following books provide an overview of the several aspects of groundwater covered in this article: D.K. TODD, *Ground Water Hydrology* (1959); and R.K. LINSLEY, M.A. KOHLER, and J.L.H. PAULHUS, *Applied Hydrology* (1949). More modern treatments of the subject, including the mathematical equations that describe groundwater motion, well hydraulics, and similar topics, are presented in: R.J.M. DE WIEST, *Geohydrology* (1965); S.N. DAVIS and R.J.M. DE WIEST, *Hydrogeology* (1966); and P.S. EAGLESON, *Dynamic Hydrology* (1970). The basis for much of the modern theory of groundwater flow is contained in the classic paper: M.K. HUBBERT, "The Theory of Groundwater Motion," *J. Geol.*, 48:785–944 (1940). See also R.J.M. DE WIEST (ed.), *Flow Through Porous Media* (1969).

(R.J.M.De W.)

Growth

Although the essence of growth is increase, it is useful to distinguish between growth of living things and that of inanimate objects. The growth of inanimate objects, such as occurs during the formation of crystals, is limited. The crystals can increase in size but are unable to reproduce themselves; living things, on the other hand, not only increase in size but also reproduce themselves. Growth of living things always occurs by an increase in the number or size of the basic units of organisms, cells, and always eventually includes reproduction. To appreciate fully how living things grow and how this growth is regulated requires a consideration of the way in which cells increase in size and number.

The increases in cell size and number that take place during the life history of an organism are seldom random; rather, they occur according to a plan that eventually determines the size and shape of the individual. Growth may be restricted to special regions of the organism, such as the layers of cells that divide and increase in size near the tip of the plant shoot. Or the cells engaged in growth may be widely distributed throughout the body of the organism, as in the human embryo. In the latter case, the rates of cell division and of the increase in cell size differ in different parts. That the pattern of growth is predetermined and regular in plants and animals can be seen in the forms of adults. In some organisms, however, notably the slime molds, no regular pattern of growth occurs, and a formless cytoplasmic mass is the result.

The rate of growth of various components of an organism may have important consequences in its ability to adapt to the environment and hence may play a role in evolution. For instance, an increase in the rate of growth of fleshy parts of the fish fin would provide an opportunity for the fish to adapt more easily to terrestrial locomotory life than could a fish without this modified fin. Without disproportionate growth of the fin—ultimately resulting from random changes in the genetic material (mutations)—the evolution of limbs through natural selection might have been impossible.

Growth is often used synonymously with development, and, indeed, increase in size is a striking feature of development. But development also includes changes in the types of cell specialization (differentiation) and extensive movements of cells. Movements of cells and tissues during embryonic development are partially responsible for the form of the body and various organs.

The discussion below focusses on the types of growth and how growth rates may be regulated by factors in the external environment and from within the organism (see DEVELOPMENT, ANIMAL; DEVELOPMENT, PLANT for detailed discussions of development).

TYPES OF GROWTH

In cells. The increase in size and changes in shape of a developing organism depend on the increase in the number and size of cells that make up the individual. Increase in cell number occurs by a precise cellular reproductive mechanism called mitosis. During mitosis the chromosomes bearing the genetic material are reproduced in the nucleus, and then the doubled chromosomes are precisely distributed to the two daughter cells, one of each chromosomal type going to each daughter cell. Each end of the dividing cell receives a complete set of chromosomes before the ends separate. In animal cells this is a pinching off (cytokinesis) of the cell membrane; in plant cells a new cellulose wall forms between the new cells.

During the period of cell life preceding the actual distribution of chromosomes, the mother cell often grows to twice its original size. Hence, a cycle consisting of cell growth and cell division is established. Cell growth—an increase in cytoplasmic mass, chromosome number, and cell surface—is followed by cell division, in which the cytoplasmic mass and chromosomes are distributed to the daughter cells. An increase in cytoplasmic mass does not always occur during cell-division cycles, however. During the early development of an embryo, for example, the original egg cell, usually a very large cell, undergoes repeated series of cell divisions without any intervening growth periods; as a result, the original egg cell divides into thousands of small cells. Only after the embryo can obtain food from its environment does the usual pattern of growth and mitosis occur.

In plants. The fact that most plant cells undergo extensive size increase unaccompanied by cell division is an important distinction between growth in plants and in animals. Daughter cells arising from cell division behind the tip of the plant root or shoot may undergo great increases in volume. This is accomplished through uptake of water by the cells; the water is stored in a central cavity called a vacuole. The intake of water produces a pressure that, in combination with other factors, pushes on the cellulose walls of the plant cells and is responsible for the increase in length and girth of the cells and of the plant. In plants, much of the size increase occurs after cell division and results primarily from an increase in water content of the cells without much increase in dry weight.

The role of water in plant-cell growth

The very young developing plant embryo has many cells distributed throughout its mass that undergo the cycle of growth and cell division. As soon as the positions of the root tip, shoot tip, and embryonic leaves become established, however, the potential for cell division becomes restricted to cells in certain regions called meristems. One meristematic centre lies just below the surface of the growing root; all increases in the number of cells of the primary root occur at this point. Some of the daughter cells remain at the elongating tip and continue to divide. Other daughter cells, which are left behind in the root, undergo the increase in length that enables the new root to push deeper into the soil. The same general plan is evident in the growing shoot of higher plants, in which a restricted meristematic region at the tip is responsible for the formation of the cells of the leaves and stem; cell elongation occurs behind this meristematic centre. The young seedling secondarily develops cells associated with the vascular strands of phloem and xylem—tissues that carry water to the leaves from the soil and sugar from the leaves to the rest of the plant. These cells can divide again, providing new cell material for development of a woody covering and for more elaborate vascular strands. Hence, the growth of higher plants —*i.e.*, those aspects involving both the pattern of stems, leaves, and roots and the increase in bulk—results primarily from cell division at the meristem followed by a secondary increase in size because of water uptake. These activities occur throughout the period of plant growth.

In animals. *General features.* The growth of animals is more restricted in time than is that of plants, but cell division is more generally distributed throughout the body of the organism. Although the rate of cell division differs in different regions, the capacity for cell division is widely distributed in the developing embryo. Increase in size is rapid during the embryonic period, continues at a reduced rate in juveniles, and thereafter is absent. Cell division and size increase continue, however, even after increase in total body size no longer occurs. Because these events are balanced by cell death, post-juvenile increase in cell number is primarily a replacement phenomenon. Height increase in mammals is limited by cessation of cell division and bone deposition in the long bones. The long juvenile period of growth in humans is unusual, most higher animals attaining mature size soon after the end of embryonic development. Some organ systems undergo little cell division and growth after birth; for instance, all of the germ cells (precursors of egg cells) of the female are formed by the time of birth. Similarly, all of the nerve cells of the brain are formed by the end of the embryonic period. Further increase in the size of the nervous system occurs by outgrowth of nerve fibres and deposition of a fatty insulation material along them. Although the greatest increase in size of nerve cells occurs, as in plant cells, after the cessation of cell division, the nerve fibre outgrowth in animals represents a true increase in the amount of cytoplasm and cell surface and not just an uptake of water.

Some organs retain the potential for growth and cell division throughout the life-span of the animal. The liver, for example, continues to form new cells to replace senescent and dying ones. Although cell division and growth occur throughout the liver, other organs have a special population of cells, called stem cells, that retain the capacity for cell division. The cells that produce the circulating red cells of mammalian blood are found only in the marrow of the long bones. They form a permanent population of dividing cells, replacing the red cells that continuously die and disappear from the circulation.

The rates of both growth and cell division can vary widely in different body parts. This differential increase in size is a prime factor in defining the shape of an organism.

Tumours. When growth is not properly regulated, anomalies and tumours may result. If the increase in the number of liver cells is abnormal, for example, tumours of the liver, or hepatomas, may result. In fact, one feature of malignant tumours, or cancers, is the absence of the usual growth patterns and rates. The cells of malignant tumours, in addition to having abnormal growth rates, have altered adhesive properties, which enable them to detach easily from the tumour; in this way the cells may spread to other parts of the body (metastasize) and grow in unusual locations. It is the growth of tumours in places other than the organ of origin that usually causes the death of an organism. Tumours may vary widely in their growth rates. The accompanying Table presents a sample of some rates of cell division in tumour tissues. Tumours may grow very rapidly or so slowly that the rate approaches that of normal cell division in adult tissues. Tumours are not only characterized by an increase in the rate of cell division but also by abnormal patterns of growth. The new cells formed in the tumour are not organized and incorporated into the structure of the organ and may form large nodules. These abnormal growths may present no medical problems (*e.g.*, moles) or may cause disastrous effects, as is the case of the pressure on the brain caused by a tumourous mass of the meningeal covering of the brain.

Features associated with tumour growth

Regeneration. Not all abnormal growths are tumours. If a tree is partially burned, cells below the bark produce

Rates of Growth of Selected Tumours

tumour type	host animal	time required for tumour to double in size
L 1210 leukemia	mouse	13 hours
Ehrlich ascites	mouse	8 hours
Human leukemia	man	60 hours
Human leukemia	tissue culture	18 hours
Lewis lung tumour	mouse	19 hours
Breast tumour	mouse	33 days
Human breast tumour	man	43 days
Burkitt's lymphoma	man	1.5 days
Stomach cancer	man	6.8 days
Cancer of the colon	man	14.2 days

Source: H.E. Skipper, "The Cell Cycle and the Chemotherapy of Cancer," *The Cell Cycle and Cancer*, R. Baserga (edit.), 1971.

a new covering for the exposed vascular strands. Growth may not be normal, and an obvious scar or growth of the new bark is apparent. Similarly, if the skin of a mammal is severely injured, the repair, although abnormal and imperfect, causes the organism no physiological difficulty. Many organisms possess the ability to regrow, or regenerate, with varying degrees of perfection, parts of the body that are lost or injured. Salamanders possess remarkable powers of regeneration, being able to form new eyes or a new limb if the original is lost. Lizards can regenerate a new tail; even humans can regenerate parts of the liver. The reasons for the differences in regenerative powers in different animals remains a fascinating mystery of great practical importance. When regeneration does occur, some specialized cells usually lose their specialized characteristics and enter a period of an increased rate of cell division; subsequently, the new cells respecialize into the tissues of the original body part. Plants whose tops are lost as in pruning can also sometimes form new meristematic centres from dormant tissues and produce new shoots (see also REGENERATION, BIOLOGICAL).

Compensatory growth. Many organs of animals occur in pairs, and if one is lost the remaining member increases in size, as if responding to the demands of increased use. If one of the two kidneys of a human is removed, for example, the other increases in size. This is called a compensatory reaction and may occur either by some increase in cell size (hypertrophy), by an increase in the rate of cell division (hyperplasia), or both. Although an increase in cell number is primarily responsible for the compensatory reaction of the kidney, the number of individual filtration units (glomeruli) does not increase. Hence, cell division increases the size of glomeruli but not the total number. Some of the most striking examples of increases in cell size in animals take place during stimulation of endocrine organs, which secrete regulatory substances called hormones; when the thyroid gland is stimulated, for example, the individual cells of the gland may increase dramatically in size.

Compensatory growth in the kidney

FACTORS THAT REGULATE GROWTH

The environment. *Temperature.* The environment in which an organism lives plays an important role in modifying the rate and extent of growth. Environmental factors may be either physical (*e.g.*, temperature, radiant energy, and atmospheric pressure) or chemical. Organisms and the cells of which they are composed are extremely sensitive to temperature changes; as the temperature decreases, the biochemical reactions necessary for life occur more slowly. A lowering of the temperature by 10° C (18° F) slows metabolism at least twofold and often more.

The width of trees increases partly by cell division and enlargement of secondary meristematic tissue below the bark. During the cold of winter, cell division and enlargement may cease completely; but during the spring renewed growth occurs. This intermittent growth is influenced by temperature, light, and water. The amount of growth may decrease considerably if the spring is cold, if day length is changed by obstructions blocking the sunlight, or if a drought occurs. In fact, the width of the growth rings visible on the surface of the cut tree trunk provides a partial history of climatic conditions, the spacing of the growth rings of different size having been correlated with known periods of drought and cold to provide reliable archaeological dating of various structures, as in the timbers used in Indian pueblos in the southwestern United States.

Temperature also affects both warm- and cold-blooded animals. Many warm-blooded (*e.g.*, bears) and cold-blooded (*e.g.*, frogs) vertebrates cease growing during the cold winter and simply enter an inactive or dormant state, which is characterized by a very low rate of metabolism. In animals that do not become dormant, increased demands for food consumption occur during cold periods to provide energy to maintain body temperature; this utilization of food energy may limit the energy available for size increase if food is in short supply.

Pressure. Because atmospheric pressure is relatively constant except in the mountains, it probably is of little importance in growth regulation. Increases in pressure in the ocean's depths may be significant, however, since it is known that increases in hydrostatic pressure interfere with cell division. Tissues of deep-sea fishes must have become adapted to such pressure effects, which have been little studied thus far. Movements of the terrestrial atmosphere—winds—may affect growth patterns in trees and shrubs, as is evident in the exotic shapes of certain conifers that grow along coastlines exposed to strong prevailing winds (see also PRESSURE AND TEMPERATURE, BIOLOGICAL EFFECTS OF).

Light. Of all the physical factors, light plays the best understood and most dramatic role. Many of the effects of light on plant growth are obvious and direct. Light energy is the driving force for photosynthesis, the series of chemical reactions in green plants in which carbon dioxide and water form carbohydrates and upon which all life ultimately depends. Insufficient light causes death or retardation of growth in green plants. But light also has indirect effects of great importance. Green plants possess small amounts of a pigment called phytochrome that can exist in two forms. One form absorbs red light (660 millimicrons, or mμ; 1mμ = 3.937×10^{-8} inch). When plants containing this pigment absorb red light, the pigment is converted to another form, which absorbs far-red light (730 mμ); the latter form can be converted back again to the original red absorbing form. These conversions have dramatic consequences; for example, red light inhibits stem elongation and lateral root formation but stimulates leaf expansion, chloroplast development, red flower coloration, and spore germination. Cycles of red and far-red light also can affect flower formation.

Indirect effects of light on plants

The effects of light on animals, although less obvious, may be important, as, for example, the effect of light on growth of the reproductive system of some animals. Increase in day length, hence in the amount of light, seems to initiate growth and development of the sex organs (gonads) in some birds during the spring. Curiously, the eyes are not the receptors for the light signal that activates the endocrine system to initiate growth of gonads; rather, cells deep in the brain are sensitive to the small amounts of light that pass directly through the thin skull of the bird.

Most animals show cyclic activity, or rhythms, in various important physical (*e.g.*, movement) and chemical (*e.g.*, respiration) events that are essential to the individual. These rhythms are often regulated by short exposure to light.

Chemical factors. Chemical factors of importance in the environment include the gases in the atmosphere and the water, mineral, and nutritional content of food. Plants require carbon dioxide, water, and sunlight for photosynthesis; drought slows plant growth and may even kill the plant. The effects of atmospheric contaminants—*e.g.*, oxides of nitrogen, hydrocarbons, and carbon monoxide—are known to have deleterious effects on the growth and reproduction of both plants and animals.

Plants and animals require minerals and small amounts of elements such as zinc, magnesium, and boron. Nitrogen and phosphorus are provided to plants as nitrates and

phosphates in the soil. Inadequate quantities of any nutritional factor in the soil result in poor plant growth and poor crop yields. Animals require oxygen, water, and elements from the environment. Because they are unable to synthesize sugars from carbon dioxide, animals must acquire these nutrients through the diet, either directly, by the consumption of plants, or indirectly, by the consumption of other animals that in turn have utilized plants as food. If the quality or quantity of this food is poor, either growth is retarded or death occurs (see NUTRITION; NUTRITION AND DIET, HUMAN; NUTRITIONAL DISEASES AND DISORDERS).

Vitamins, a class of compounds with a variety of chemical structures, are needed by animals in small amounts. Animals cannot synthesize all vitamins they require; those that cannot be synthesized must therefore be acquired in the diet, either from plants or from other animals that can synthesize the vitamin. Because certain vitamins are necessary in certain important metabolic reactions, vitamin deficiency during growth may have a variety of effects—stunting, malformation, disease, or death (see VITAMIN).

Internal factors. The organism is dependent on the environment for the raw materials for growth, but growth is also regulated internally. Because the size and form of plants and animals are under genetic control, events such as the rate and site of cell division and the extent of cell enlargement can be affected by mutations. It is not yet known, however, precisely how these factors, which are the ultimate determinants of growth, are controlled in individual cells.

One very important class of intrinsic growth regulators is that of the hormones. The principle plant hormone, auxin, is produced in the leaves; it moves by precise mechanisms, as yet poorly understood, to the other parts of the plant, controlling such processes as elongation of plant cells. Auxin somehow changes the characteristics of the rigid cell wall of the plant cell so that it becomes more flexible; the internal pressure within the cell then forces it to become larger. Other plant hormones may also play a role in the process; hormones such as cytokinins and gibberellins influence the rate of cell division in the meristems. Some dwarf plants can be stimulated to grow to normal size simply by applying gibberellin.

Hormones also play a decisive role in animal growth. One hormone from the pituitary gland at the base of the brain is called growth hormone because of its extensive and widespread effects on growth. A deficiency of growth hormone in pre-adolescents results in dwarfism, and oversupply of the hormone (often caused by a tumour) results in gigantism. If an excess of growth hormone is produced after the long bones can no longer grow—*i.e.*, post-adolescence—a disease called acromegaly, which is characterized by increases in the size of the hands and feet and broadening of facial features, results. A deficiency of thyroid hormone in children also causes growth retardation.

The sex hormones secreted from the pituitary gland interact in a complex way to regulate the growth of the gonads. The gonads in turn produce estrogen and progesterone in females and testosterone in males; these hormones control the development of human secondary sexual characteristics—body hair, enlargement of mammary glands in females, and growth of the vocal cords in males. Although the growth hormones and sex hormones play a vital role in growth, the exact mechanism by which they function has not been established with certainty (see HORMONE).

In addition to having the ability to synthesize the factors that regulate growth, plants and animals evidently possess exquisite mechanisms for integrating and regulating the production of hormones; *i.e.*, the appropriate amounts of the right hormones are produced at the right time and the right place for normal growth.

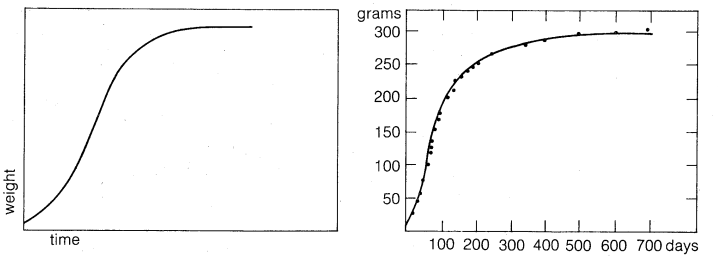
Limits of growth

Although many plants, including trees, grow throughout their lives, growth of parts of the organism is not perpetual; *e.g.*, leaves of a given species attain a specific size and can grow no larger. In animals, growth stops entirely, except for replacement, after the juvenile period. The limits for both total body size and organ size are probably established by genetic mechanisms. The factors involved in limiting the growth of an organism are not yet definitely known, but evidence indicates that the liver releases into the bloodstream protein molecules that can limit growth of the organ. Thus, one theoretical view is that an organ may produce substances that serve to limit its own growth, thereby establishing a feedback mechanism. A protein called nerve-growth factor is important for the growth of some parts of the mammalian nervous system. If too much of the nerve-growth factor is present, growth of sympathetic nerve fibres is extensive and aberrant. If the nerve-growth factor is eliminated from the body—by injection of an antibody against the factor—the sympathetic nerves wither and disappear. Other subtle growth regulatory substances specific for various organ systems may eventually be discovered.

THE DYNAMICS OF GROWTH

The growth curve and the measurement of growth. The mathematical analysis of the rate of growth has been a subject of interest for many years. It is based on the rule of cell division: one cell gives rise to two daughter cells. Hence, the theoretical increase in cell number would be a geometric series, in which one cell produces two cells, then four, eight, 16, and so on. In reality, however, the rate of growth is not constant but declines after a period of time, usually because of influences in the environment or because of inherent genetic limitations. Thus the curve showing the growth of cell populations and of organisms is usually S-shaped, or sigmoid, when growth is plotted against time on a graph (see Figure). The increase in cell number resulting from cell division accounts for the rising part of the curve; the rate of cell division decreases at the plateau in the curve. The S-shaped growth curve is generally applicable to the growth of organisms. If growth is plotted against time on a logarithmic scale, the early intense growth (called log growth) in the rising phase of the growth curve falls on a straight line.

From B.I. Balinsky, *An Introduction to Embryology* (1970); W.B. Saunders Company



Sigmoid growth curves.
(Left) Ideal sigmoid growth curve. (Right) An actual postnatal growth curve of the white rat.

The rate of growth may be defined by the differential equation $v = dW/dt\ (1/W)$, in which v is the growth rate and W is the weight at any given time, t. The solution of this equation provides a value for relative increase—the increase in weight related to the initial mass of the growing substance. The animal that most closely approaches a constant rate of growth is an insect larva. In most animals the rate of growth declines as the organism becomes larger and older.

Although the S-shaped growth curve describes with fair accuracy the growth of populations of single cells, such as bacteria or cells of higher organisms in tissue culture—the growth in a sterile nutrient environment of cells of tissues from organisms—the growth rates of different parts of whole organisms vary. The relationship of the growth of one part of an organism to that in another part is called allometry. An equation expressing the fundamental relationship of allometric growth is $y = b\,x^k$, in which y is the size of one organ; x is the size of another; b is a constant; and k is known as the growth ratio. Although such mathematical tools have allowed a very thorough description of the differential growth of different parts of an organism, they have unfortunately

not provided insight into the physical and chemical control of the growth rate.

The study of growth. Even though the chemical, physical, and genetic bases of growth are elusive, much has been learned about the process by growing tissues in a sterile nutrient environment. Even if the source of the tissue is an organ that has completely stopped growing, such as the nervous system of an animal or the phloem of a plant, the cells will begin to grow again in culture, often at a logarithmic rate of increase. It may therefore be concluded that the organism as a whole places constraints upon the ability of individual cells to reproduce and that, when these constraints are removed, the growth potential of the cells is no longer restrained. Even in tissue culture, however, the rate of cell growth eventually slows, hence the sigmoid-shaped growth curve. During the rapid growth phase of cells in tissue culture, they usually lose the ability to carry out the specialized function characteristic of their organ of origin; for example, if cartilage cells divide rapidly, they no longer synthesize cartilaginous matrix. This phenomenon of apparent despecialization has been a topic of great theoretical interest: are rapid growth and specialization mutually exclusive activities? Evidence shows that some types of specialized cells may be maintained in tissue culture for very long periods of time and still retain the ability to carry out specialized biosyntheses, so that the apparent loss of specialized function in tissue culture cells may not fundamentally result from a mutual exclusivity of growth and differentiation.

Contact inhibition of growth

When the growth of tissue-culture cells begins to slow, one factor responsible is exhaustion of critical components from the medium. But even if the medium is frequently replaced, when the bottom of the culture dish becomes densely packed with a layer of cells, the growth rate drops—a phenomenon called contact inhibition of growth. It is believed that cells so close that they are always touching provide a signal that retards the rate of cell division. Apparently identical cells in tissue culture also show great variation in growth rate. Some cells from the skin, for instance, when placed in culture, may divide every eight hours; other similar cells may divide only every 36 hours. The growth of cells in a controlled environment such as tissue culture offers many possibilities for studying the fundamental mechanisms controlling cell growth and, consequently, the growth of organisms and populations.

BIBLIOGRAPHY. L.B. AREY, *Developmental Anatomy*, 7th ed. (1965), a standard and authoritative treatment of the growth and development of the human embryo; B.I. BALINSKY, *An Introduction to Embryology*, 3rd ed. (1970), a definitive modern textbook of growth and development of animals; R.J. GOSS, *Adaptive Growth* (1964), a modern treatment of the mechanisms of compensatory growth in animals; JULIAN HUXLEY, *Problems of Relative Growth* (1932), a successor to D'Arcy Thompson's book, with formulation of the problems of allometric growth; D'ARCY WENTWORTH THOMPSON, *On Growth and Form*, abridged edition ed. by J.T. BONNER (1961), the classic mathematical treatment of the dynamics of growth, reissued in an abridged form; J.G. TORREY, *Development in Flowering Plants* (1967), a modern authoritative treatment (in paperback) of the occurrence and mechanisms of growth in higher plants.

(F.H.W.)

Gruiformes

The bird order Gruiformes comprises a rather loose assemblage of 12 families that are generally agreed to be related but that differ widely in many aspects. They are an ancient group with a rich fossil history, but many are now restricted in range and few in number. Members of the order occur on every continent, but the only family with worldwide distribution is the Rallidae (rails, gallinules, and coots), with 132 living species. Cranes (Gruidae) are found on every continent except South America, but many of the 14 species have small populations, some on the verge of extinction. The bustards (Otidae), with 23 species, have a wide distribution, limited to the Old World, but hunting pressures and modern agricultural methods have greatly reduced their numbers. The mesites (Mesitornithidae), however, are confined to Madagascar, and the kagu (*Rhynochetus jubatus*) to the island of New Caledonia. Other small families in the order contain the hemipodes, or button quails (Turnicidae), plains-wanderer (Pedionomidae), limpkin (Aramidae), trumpeters (Psophiidae), finfoots (Heliornithidae), sun bittern (Eurypygidae), and cariamas (Cariamidae).

Although man's impact on them is very great, gruiform birds, because of their scarcity, have a negligible impact on man. With the possible exceptions of the sandhill crane (*Grus canadensis*), which descends on the grainfields of the Canadian prairies during the autumn migration, causing some crop damage, and the brolga, or Australian crane (*Grus rubicundus*), which causes similar damage in Queensland, no gruiform can be considered harmful to man's interests; some of the larger species, in fact, are hunted for food or sport.

GENERAL FEATURES

Size and weight

Gruiform birds range in size from the tiny button quails (*Turnix*) and miniature rails, such as the North American black rail (*Laterallus jamaicensis*) barely 15 centimetres (six inches) long, to the stately sarus crane (*Grus antigone*) of India, standing nearly 1.6 metres (five feet) high. The enormous kori bustard (*Otis kori*) and the Eurasian great bustard (*O. tarda*) may weigh up to 18 kilograms (about 40 pounds) and are the heaviest modern flying birds. Gruiforms vary widely in structure; some are adapted for life in or near water, others for life on land. Some forms fly well; but a number of species are flightless.

As a group, the gruiforms are probably best known for their impressive and graceful courtship displays, the most famous of which, the dances of the cranes, are imitated and adapted by many native peoples. The Ainu of Japan have a crane dance in honour of the Japanese crane (*Grus japonensis*), and many African tribes imitate the dance of the crowned crane (*Balearica pavonina*). Less known, but no less spectacular, are the striking wing display of the sun bittern (*Eurypyga helias*) and the strutting and booming of the larger bustards.

NATURAL HISTORY

Ecology. *Habitat.* Gruiform birds live in a variety of habitats, from water and marshes to arid plains. The most aquatic are the finfoots and coots (*Fulica*). The former live along slow-flowing streams where heavy overhanging vegetation affords them cover, the latter on more open water. Most rails live in saltwater or freshwater marshes. The limpkin (*Aramus guarauna*) is essentially a marsh bird; in Florida it inhabits the sawgrass marshes and cypress swamps of the Everglades. Cranes bridge the gap between marsh and dry-land birds, nesting in marshes but occurring in open plains and cultivated fields on migration and in winter. The sun bittern prefers muddy, wooded riverbanks but also occurs in woods well away from water. Trumpeters, mesites, the kagu, and some rails live in forest and dense brush. The remaining gruiform families inhabit more open country: the cariamas of South America favour grassland or hot, dusty plains with scattered bushes; the Old World bustards, button quails, and the plains wanderer prefer open, grassy plains, although they will accept old pastures and cultivated fields.

Food habits. Corresponding to the wide variety of habitats utilized by the gruiforms is the great diversity of food taken by them. In general, more animal than vegetable food is taken, but as a group the gruiforms are omnivorous. Coots and gallinules consume much aquatic vegetation. Finfoots live largely on mollusks, frogs, and small fish. The limpkin has a more specialized diet, consisting chiefly of certain large snails, which are broken open at certain favourite feeding stations that contain telltale piles of broken shells. The limpkins' near relatives, the cranes, eat almost any animal food, including rats, mice, moles, lemmings, lizards, snakes, frogs, tadpoles, snails, and a variety of insects. On its wintering grounds in Texas, the whooping crane (*Grus americana*) lives largely on crustaceans. The sandhill crane includes

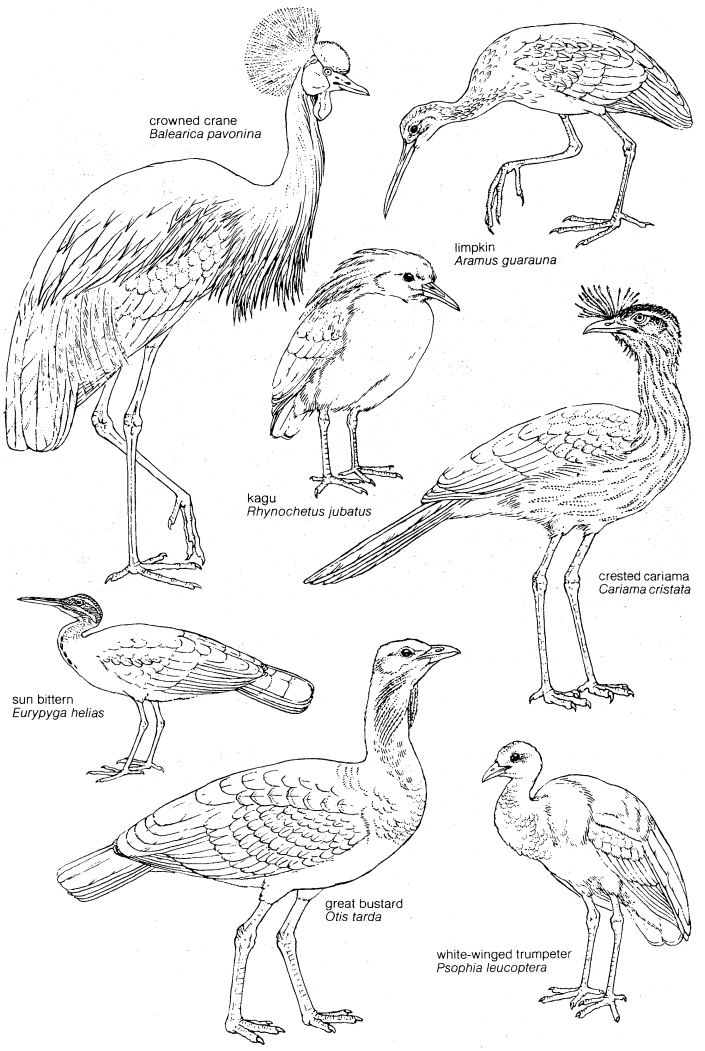


Figure 1: Body plans of some larger gruiforms.
Drawing by R. Keane

berries and grass in its otherwise animal diet on the northern tundra and gleans some plant material from old potato and grain fields on its southern wintering grounds. The Asiatic cranes that winter in Japan (such as the hooded crane, *Grus monacha*, and the white-naped crane, *G. vipio*) glean grain in rice paddies, and Japanese cranes in Hokkaido are fed corn (maize) by the local farmers. Cranes use their powerful bills for digging in the ground to get at bulbs and roots below the surface, and a similar foraging action has been noted for the gray-necked wood rail (*Aramides cajanea*) of tropical America, which probes in ground debris, flicking it aside with the bill. Rails as a group, like cranes, are omnivorous, though the bulk of their diet consists of small marsh animals, such as snails, crustaceans, frogs, and water insects. The purple swamphen (*Porphyrio porphyrio*), a vegetarian, often has a feeding platform on which it stands and pulls up the surrounding water plants; it will also climb up reeds and eat the flower heads.

The other members of the order have a similar mixed animal and vegetable diet, though with the emphasis always on the animal side. Captive cariamas readily eat meat, and a captive kagu feasted on ground steak provided by its owner. Bustards are fond of grasshoppers, and their varied diet also includes dung beetles, termites, centipedes, grass, clover, vegetable crops, and even, in Africa, the gum from the trunks of *Acacia* trees. The sun bittern stalks insects as a heron stalks fish, stealthily approaching its prey with neck drawn in, then grabbing it with a sudden, stabbing thrust of the bill.

Reproductive behaviour. *Courtship.* The breeding cycle of many gruiform birds begins with elaborate courtship rituals and displays. Cranes pair for life, and the strong pair bond necessary to maintain this partnership is initiated and continued by a series of displays that, since they often consist of two birds facing each other and leaping into the air, are generally known as dances.

Dancing by cranes

The ceremony frequently begins as two birds circle each other with a curious, formal step, the legs stiff and the head and neck held high. The next action is bowing or "head bobbing," in which the head is held horizontally, the neck curved down in a U, and in this position the head and neck are bobbed or pumped up and down. One bird may do this while the other bird looks on, or both may do it together. One bird (or both) then turns in a circle on the spot while continuing to bow. The momentum increases until suddenly both birds leap into the air and follow this by dancing. While dancing, the birds often pick up some object from the ground and toss it in the air as they leap. The object is usually something used in nest building, such as a stick, leaf, or tuft of grass, and the action may be related to nest-building drives. Again, while leaping in the air one bird may turn its back on the other. Another beautiful display involving a pair of cranes is the "duet." The birds first circle each other with the same formal step that initiates the dance, but instead of going into the dance they droop their wings, throw back their heads, and call in unison.

Although dancing intensifies at the beginning of the breeding season and is doubtless primarily connected with courtship and pair formation, it can occur at any time of year and may have other functions as well. Dancing often seems to be a method of releasing pent-up energy, as when a bird dances on its own or when a dancing pair sets off dancing in nearby individuals of a flock. So strong is the instinctive urge to dance that a five-day-old chick is recorded as leaping up and down and going through other motions of the dance, even though it had never seen another crane.

Vocalizations of rails

Pair formation in many rails is effected chiefly by voice; males establish territories and defend them vigorously with distinctive songs and calls while at the same time attracting females. Rails produce an amazing variety of noises. In North America, the clapper rail (*Rallus longirostris*) has a loud cackling call, the sora (*Porzana carolina*) an explosive whinny, and the yellow rail (*Coturnicops noveboracensis*) makes a noise like two small stones being clicked together. In Europe, the water rail (*Rallus*

Drawing by R. Keane

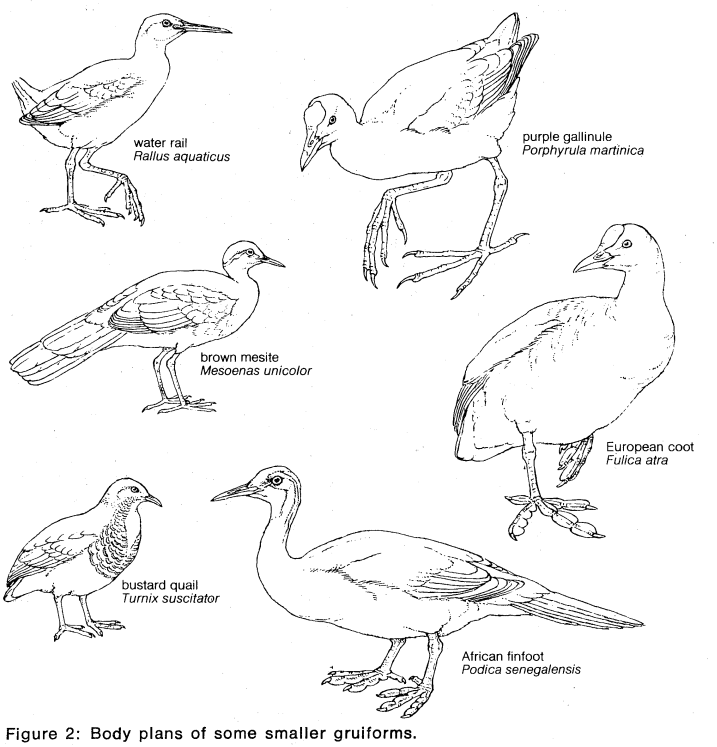


Figure 2: Body plans of some smaller gruiforms.

aquaticus) squeals like a stuck pig, and the corncrake (*Crex crex*) produces a rasp like a heavy comb being drawn over a piece of wood. The gray-necked wood rail greets the dawn with a ringing "Pop-tilly, pop-tilly, pop-tilly, ko-ko-ro-ko-ko," which has given it the local name of "cocaleca" in Panama. In Africa, a large rail, *Himantornis haematopus*, of the Congo forests makes a pumping sound, and the black crake (*Limnocorax flavirostra*) in the papyrus swamps makes a curious gurgling sound. The tiny crakes in the genus *Sarothrura* have a variety of melodious calls, the buff-spotted crake (*S. elegans*) making a low moaning noise like the sound of a tuning fork. More familiar are the plebeian squawks and grunts of the wide-ranging coots and gallinules.

Rails also have courtship displays involving lowering the head, raising the wings, and fanning the tail, often while uttering some special call. Coots and gallinules display their colourful frontal shields and fan their tails to show off the white under tail coverts. Other species circle their partners in various postures. The display often ends with the male chasing the female, and the chase may end in copulation. In button quails the role of the sexes is reversed. The female is the most brightly coloured and courts the male, erecting her tail, puffing out her neck, and running around him while uttering a low crooning note. The male sun bittern selects an open spot, often in a patch of sunlight, where he spreads and raises his tail and wings until they meet in front of his head, exposing elegant patterns of red-brown, olive, gray, and black. In this posture he runs in a circle and may jump up in the air or bob his head.

In its aggressive display, the kagu stands erect, with the long feathers of the crest raised, the wings held out from the body, and the tail drooped. In this posture, it bounces at the adversary. In a playful mood the kagu will toss sticks and stones around with its bill, in a manner reminiscent of cranes. A captive cariama has been seen to run at a tree with lowered head, jerking its tail and giving a short cluck, then striking the tree with both feet. In spring wild birds have a bustard-like display (see below) accompanied by loud yelping calls.

Vocal displays

Trumpeters, named for their loud, resonant cries, have a cranelike dance involving strutting around on the ground and leaping into the air. Males of the larger bustards, the great and kori bustards, for example, develop a special pouch in the neck during the breeding season with which they produce loud, booming calls. The Australian bustard (*Ardeotis australis*) "booms" with closed beak, producing a sound somewhat like the distant roar of a lion. Special display feathers that grow on the head, neck, and breast are molted after the breeding season. A large bustard in full breeding regalia, with all feathers puffed out, is scarcely recognizable; the head is hardly visible, being immersed in an immense ruff, and the wings and tail are raised until they meet over the back, the total effect being that of an enormous feather ball. The ball then struts around in front of a female, who feigns indifference. Smaller bustards have somewhat different displays. The crested bustard (*Lophotis ruficrista*) of Africa has an aerial display flight in which it rises about 100 feet into the air and then planes steeply back to earth.

Nesting. Gruiform birds nest both on the ground and in bushes and trees. Of the ground nesters, button-quails and the plains wanderer nest in a grass-lined hollow, often building a domed roof and side entrance. Cranes raise up a pile of vegetation in open shallow water, and the limpkin builds a concealed nest in dense marsh grasses. Marsh-dwelling rails build simple nests of grass and aquatic plants, often in a thick tuft of grass, the blades of which they pull down over the nest to conceal it. The kagu builds a nest of sticks and leaves in a depression in the ground. Bustards scarcely have a nest at all, the eggs being laid on bare ground, sometimes beneath a bush or clump of grass for concealment. Finfoots and the sun bittern prefer a nest site on a branch of a tree, and finfoot nests generally overhang water. Of the two species of cariamas, one nests on the ground and the other in bushes. Trumpeters are variously reported as nesting on the ground and in holes in trees. Mesites place their simple stick nests a metre or two up in a tree. Being flightless, they must always find a site where a connecting series of branches leads from the ground to the nest, enabling them to hop up to it.

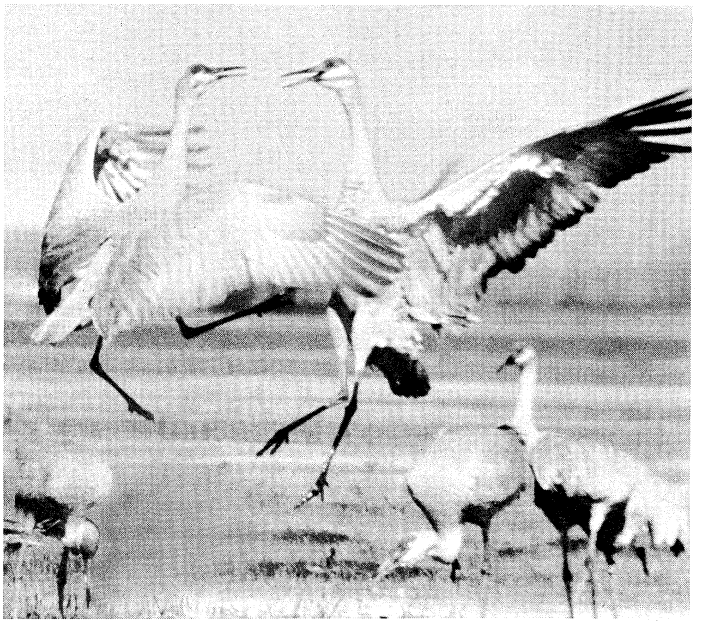


Figure 3: Sandhill cranes (*Grus canadensis*) leaping into the air as part of the courtship dance characteristic of this family of birds.
By courtesy of the U.S. Bureau of Sport, Fisheries, and Wildlife; photograph, David Marshall

Clutch sizes vary widely within the order, from the single egg of the kagu to over ten in some rails. Exceptional clutches of 15–20, recorded for some coots and gallinules, may be due to more than one female laying in the same nest. The eggs are usually white or buff, sometimes pale gray or pale green, immaculate in a few species, but usually with brown spots or blotches at one end. The incubation period is typically about three weeks, extending to four weeks in the cranes and larger bustards. Both sexes incubate the eggs and care for the young when hatched in all families except bustards, trumpeters, and button quails. Button quails are polyandrous (an individual mating with several members of the opposite sex), and the tasks of incubation and care of the young are performed entirely by the males. In bustards and trumpeters, the female does all of the incubating and caring for the young. Young gruiforms are downy, plain black in rails and dark brown in limpkins, variously patterned in most other groups. They leave the nest immediately or very soon after hatching, except for those of the sun bittern, cariamas, and bustards, which are cared for at the nest for a short time. There is some evidence that in the trumpeters, which habitually travel in flocks in the adult state, several pairs may pool their young and look after them communally.

Incubation of the eggs

Molt. The sequence of molts and plumages is very poorly known except in cranes, the young of which have a brown or gray juvenile plumage, with white-tipped or blackish feathers in some species. The juvenile feathers are gradually replaced at each successive molt with the white or gray feathers of the adult, but the last brown-tipped feathers do not completely disappear until about the third summer, when the bird is a little over two years old. Adult cranes, at least those nesting in northern temperate zones, molt in two stages; many of the wing and tail feathers are molted in early summer, at which time the birds may be flightless for a while; the rest of the plumage is gradually molted between August and October.

Migration and locomotion. Most gruiforms are nonmigratory. Bustards, button quails, and the plains wanderer migrate locally, following the rains to feed and nest. Only the birds nesting in the North Temperate zone

are true migrants; this group includes many cranes, some rails, and the Eurasian bustards. The spectacular migrations of cranes have excited man's interest since earliest times. The peoples of eastern Asia welcome the return of the cranes as symbolic of the coming spring, and in fall the farmers of Japan welcome the birds back to the rice paddies where they spend the winter. In Japan, the cranes return every year to the same traditional wintering grounds, where they are given strict protection. The last remaining whooping cranes in North America are likewise carefully looked after and are counted at the Aransas National Wildlife Refuge in Texas on their return from the north. Most cranes cover great distances on migration. Sandhill cranes travel 4,000 miles from their nesting areas in Alaska and eastern Siberia to southern United States, and common cranes (*Grus grus*) and demoiselle cranes (*Anthropoides virgo*) cover similar distances in the Old World.

When the northern marshes freeze over in winter, rails are forced to head south. Amazingly, these birds that fly so weakly across a marsh, with floppy wings and dangling legs, are able to travel thousands of miles each year on migration. The corncrake, nesting in Scotland, may winter in South Africa, and the sora rail of North America regularly crosses 800 miles of ocean to reach Bermuda.

The Eurasian bustards travel shorter distances, going only far enough to escape bitterly cold weather. Some species form large flocks at migration time. Bustards are great walkers, and local migrations in Africa are for the most part performed slowly on foot.

Travel on foot

Gruiform birds have a predilection for travel on foot. Many only fly when pressed, and some, like the mesites, have become flightless or nearly so. Many oceanic islands have been colonized by rails, which then evolved flightlessness in the absence of predators. The subsequent advent of rats, cats, pigs, or goats on such islands, usually with accidental or intentional assistance from man, has resulted in the extinction of a number of such rails. Rails typically sneak away on foot in thick vegetation, and button quails are equally loath to fly, preferring to walk away with their quick, nervous gait, stopping every so often to raise their heads and look around for danger. Trumpeters run fast and can even swim. Finfoots spend most of their life in water and prefer to hide in thick riverside bushes when disturbed, rather than fly. On the ground, gruiform birds move efficiently and even elegantly. The sun bittern walks gracefully with slow, precise steps, its neck outstretched. Cariamas run swiftly over the plains. Rails have a very characteristic walk in which the tail is flicked up with each step, and both the limpkin and the kagu share this tail-flicking action. Even such fine fliers as the cranes prefer walking, and there is no more elegant sight in the avian world than a tall and stately crane walking with deliberate and dignified gait across the prairie. In flight, cranes and the limpkin have a characteristic wing action—a slow downstroke followed by a quick, flicking upstroke.

FORM AND FUNCTION

Gruiform birds vary greatly in shape and size and exhibit a broad range of morphological characteristics. Their plumage is predominantly brown or gray. Some have brightly coloured soft parts such as the bare, red skin on the head and neck of some cranes, used in displays, and the bright red and yellow bills and frontal (forehead) shields of gallinules. The crowned crane has a curious crest of stiff, golden feathers. The sexes are alike in most groups, except among the button quails, in which the female is more brightly coloured, and the bustards, where the males are more colourful.

The wings are rounded and often long, although in the nearly flightless mesites they are greatly reduced. The length of the tail varies, being proportionately short in button quails, rails, and trumpeters and rather long in mesites, finfoots, and the sun bittern. Cranes have very long inner secondary feathers (those of the inner wing or "forearm"), which extend beyond the end of the tail, giving the impression of a long-tailed bird. The bill is generally long and slender, particularly so in cranes, many rails, and the limpkin, although the cariamas have hooked bills which are doubtless used in tearing up mammalian prey. The legs are rather long, reflecting a preference for walking. The toes vary greatly—in the finfoots and coots they are lobed for swimming, in rails and the limpkin they are long and slender for walking on lily pads and other aquatic vegetation, but bustards and cariamas have short toes for running on hard surfaces. The hind toe, when present, is usually elevated.

Special anatomical features

Some groups have anatomical features peculiar to themselves. The mesites possess five pairs of powder down patches, far more than any other group, and the function of these is uncertain. Cranes, and the adult male limpkin have an extremely long trachea, or windpipe, that is coiled in several convolutions. These convolutions of the trachea probably give added power and resonance to the voice, which can carry for distances of a mile or more. Rails have a laterally compressed body, which gives rise to the expression "thin as a rail," enabling them to sneak between reeds and blades of grass without telltale movements of the vegetation. Most young rails have a claw at the tip of the alula (the "bastard wing" or "thumb") that enables them to clamber around on marsh vegetation. Finfoots have a sharp spur of uncertain function at the bend of the wing. The kagu, like the mesites and some rails, is flightless, a condition that may lead to its extinction by the dogs, cats, pigs, and rats that have been introduced on New Caledonia. Males of great and kori bustards have a gular (throat) pouch during the breeding season that opens into the mouth under the tongue and can be inflated at will. It is used by the birds to produce booming calls during courtship. The Australian bustard has no gular pouch, producing its calls by filling the esophagus with air. The esophagus is similarly used in sound production by the button quails and by rails of the genus *Sarothrura*.

EVOLUTION AND PALEONTOLOGY

Gruiform birds have the best fossil record of any avian order, going back to the Late Cretaceous, nearly 100,000,000 years ago. Fourteen fossil families are known, divided into 86 genera and 146 fossil species. In addition, 39 extant species are also known as fossils.

The oldest gruiform bird is the crane-sized *Laornis* from the Upper Cretaceous of New Jersey, recently shown to be closely related to the rails. *Telmatornis*, from the same epoch, formerly considered a gruiform, is now placed closer to the thick-knees (Burhinidae) in the order Charadriiformes.

Next in age are two families from the Paleocene (about 60,000,000 years ago), the Gastornithidae and the Diatrymidae, the latter containing some fascinating giant species—*Diatryma steini* stood about 7 feet tall and had a massive head and bill—but it is in the Eocene, starting about 54,000,000 years ago, that gruiform birds first became abundant and the first representatives of modern families appeared: rails, cranes, and a bustard. In the Oligocene the limpkins, and the suborder Cariamae had their beginnings. The Cariamae are represented today by only two living species, but their fossil history shows that in earlier epochs they were a more widespread and successful group. They include a number of flightless giants, the best known of which are several species of *Phororhacos* from the lower Miocene of Patagonia. These were powerful birds, the largest in excess of 2.2 metres (seven feet) in height. The best-known form, *P. longissimus*, must have been a formidable predator, having a massive skull 65 centimetres (two feet) long and 25 centimetres (10 inches) high, with a hook at the tip of the beak.

With the exception of the above, and a single *Turnix* from the Pleistocene of China, the smaller modern gruiform families have no fossil records. Nevertheless, the abundance of fossils shows what an amazingly successful group the gruiforms were. They are known from all six continents, and for every fossil species discovered there must have been tens or even hundreds that are unknown. Today, however, the situation is quite different; as a group, gruiform birds are on the decline. Of the twelve gruiform families, fully eight are represented by three

species or less, and four have only a single species. The only one that is at all numerous is that of the rails, with about 130 species. Many gruiform birds appear in the "Red Data Book" of endangered species, published by the International Union for the Conservation of Nature. Five of the fourteen extant cranes are listed as in danger, the best known being the whooping crane. The Red Data Book lists seven rails as in danger and fifteen more that have become extinct since about 1600. Many of the extinct rails lived on small islands and a number had become flightless, making them vulnerable to man and his animals. Nevertheless, the Gruiformes had been declining long before man entered the scene, apparently because the ecological conditions that favoured them in the past are less available to them today.

Endangered species

CLASSIFICATION

Distinguishing taxonomic features. The order Gruiformes is a heterogeneous group bound by taxonomic characters that are largely anatomical, and hence not readily evident on the live bird. Few are common to all members of the group, and no single character serves to separate the Gruiformes from all other orders. Some of the features that have been used in classifying the order are: the condition of the feet and toes; the type of palate, osteologically; the type of pelvic musculature; the number of carotid arteries; the presence (eutaxy) or absence (diastataxy) of the fifth secondary flight feather in the wing; the number of wing and tail feathers; the number of cervical vertebrae; the type of nostrils; the presence or absence of an aftershaft (a small "second feather" attached to the shaft of a body feather); and the behaviour of the young.

Annotated classification. The classification presented here is based on work by the paleontologists Alexander Wetmore, Pierce Brodkorb, and Joel Cracraft. Groups indicated by a dagger (†) are known only as fossils.

ORDER GRUIFORMES

Primarily marsh-dwelling birds of medium to large size. Toes not webbed (lobed in a few genera). Hallux (hind toe) usually elevated, sometimes absent. Two carotid arteries usually present. Aftershaft usually present. Crop lacking. Young usually nidifugous (precocious).

Suborder Mesitornithides

Distinguished from other suborders by reduced clavicles (collarbones) and presence of five powder-down patches.

Family Mesitornithidae (mesites). No fossil record. Superficially dovelike, but with characters of order and suborder. Hallux well-developed and functional. Three species; terrestrial, in forest and dry brush; confined to Madagascar; length 25–27 cm.

Suborder Turnices

Principal difference from other suborders is the possession of a well-developed basipterygoid process (a projection at the base of the skull).

Family Turnicidae (button quails). Pleistocene (one specimen) to present. Small and quaillike; short legs, no hallux; beak short, slightly down-curved. Unique in the order in having only one carotid artery. Female larger and more brightly coloured than male. Eggs roundly oval. Eutaxic. Fifteen species; in grasslands and brush; southern Europe, Asia, Africa, and Australia; body length 11–19 cm.

Family Pedionomidae (plains wanderer). No fossil record. Similar to button quails, but differs as follows: hallux present; two carotid arteries; wing diastataxic; eggs pointed. One species, confined to dry plains of Australia; body length about 16 cm.

†Suborder Gastornithes

Fossil only.

†*Family Gastornithidae.* Upper Paleocene of Europe; five species; large, flightless.

†*Family Diatrymidae.* Upper Paleocene to middle Eocene; North America and Europe; four species; large flightless predators.

Suborder Grues

Adults with down feathers in both pterylae (feather tracts) and apteria (areas between tracts); hallux present and variable; oil gland present and tufted.

Superfamily Gruoidea (cranes, limpkin, and trumpeters)

Only gruiforms with supraorbital furrows (*i.e.*, above the eye); distinguished from others by pelvic muscle formula. Ten to twelve tail feathers.

†*Family Geranoididae.* Fossil only; lower to middle Eocene; North America; seven species known.

†*Family Bathornithidae.* Fossil only; lower Oligocene to lower Miocene; North America; eight species known.

†*Family Eogruidae.* Fossil only; upper Eocene to upper Miocene; eastern Asia; two species known.

Family Gruidae (cranes). Lower Eocene to present. Bill long and straight; neck and legs long; hallux small and elevated. Head usually partly naked in adult. Wing diastataxic. Caudo-femoral muscle present (except in *Balearica*). Twenty-one fossil and 14 Recent species; plains and marshes of the world, except South America; length about 79 to 150 cm.

†*Family Ergilornithidae.* Fossil only; lower Oligocene to lower Pliocene; eastern Europe and Asia; three known species.

Family Aramidae (limpkins). Lower Oligocene to present. Bill long; neck long and slender; head feathered. Hallux large and functional. Caudofemoral muscle absent. Swamps and marshes of New World tropics and subtropics; five fossil and one Recent species; length 58 to 71 cm (in Recent forms).

Family Psophiidae (trumpeters). No fossil record. Bill short. Separated from Gruidae and Aramidae by nostril shape, eutaxic wing, absence of occipital foramina (perforations at the base of the skull). Three species, in lowland forest of South America; length 43 to 53 cm.

Superfamily Ralloidea

Lack occipital foramina and supraoccipital furrows. Hallux well-developed and functional; cervical vertebrae fewer than in Gruoidea.

Family Rallidae (rails, gallinules, and coots). Upper Cretaceous to present. About 56 fossil and 132 Recent species; principally in marshes; worldwide; length about 14 to 51 cm.

†*Family Idiornithidae.* Fossil only. Upper Eocene to lower Oligocene of Europe and North America; eight species known.

Suborder Heliornithes

Distinguishes from all other gruiforms, except coots (*Fulica*), by possession of lobed feet; by pelvic muscle formula, type of flexor tendons, and possession of 18 tail feathers.

Family Heliornithidae (finfoots). No fossil record. Three species; slow-flowing streams in tropical areas of Central and South America, Africa, and India and Southeast Asia; length 30 to 62 cm.

Suborder Rhynocheti

Large aftershaft; powder downs occurring as scattered groups of feathers. Large operculum (covering flap) over nostrils. Young nidoculous (dependent).

Family Rhynochetidae (kagu). No fossil record. One species, confined to forested highlands of New Caledonia; length about 56 cm.

Suborder Eurypygae

One pair of powder-down patches. Oil gland nude. Eighteen cervical vertebrae. Young nidoculous.

Family Eurypygidae (sun bittern). Beak medium length, straight, sharp; neck slender. Tail of medium length. One species; margins of woodland streams in Central and South America; length about 46 cm.

Suborder Cariamae

Distinguishing features in palate structure; talon-like nail on second toe; and type of flexor tendons.

†*Family Cunampaiidae.* Fossil only. Lower Oligocene of Argentina; one known specimen.

†*Family Brontornithidae.* Fossil only. Lower Oligocene to middle Miocene of Argentina and Uruguay; two species known.

†*Family Palaeociconiidae.* Fossil only. Lower Oligocene to middle Pliocene of Argentina; four species known.

†*Family Prophororhacidae.* Fossil only. Lower Pliocene to lower Pleistocene of Argentina; three species known.

†*Family Phororhacidae.* (Phorusrhacidae of some authors). Fossil only. Lower Miocene to lower Pleistocene; ten species, nine from Argentina, one from Florida. Medium to large flightless predators.

†*Family Psilopteridae.* Fossil only. Lower Oligocene to middle Pliocene of Argentina; nine species known.

Family Cariamidae (cariamas or seriemas). No fossil record. Moderate-sized cursorial birds; legs long, feet small; tail long; beak broad, moderately long, slightly decurved. Fore-

head and back of neck crested. Two species; in grassland and brush, respectively, of east-central South America; length 76 to 92 cm.

Suborder Otides

Oil gland absent. Scales on tarsus (lower leg) hexagonal; hallux absent. Sternum with two pairs of notches. Wing with 11 primary flight feathers. Egg white-protein structure (of Otidae) unlike those of other gruiforms.

Family Otidae (bustards). Cursorial, but strong flying birds of open plains of Eurasia, Africa and Australia; three fossil and 23 Recent species; length 37 to 132 cm.

†*Family Gryzajidae.* Fossil only. Lower Pliocene of the Ukranian S.S.R.; one species known.

Critical appraisal. The great diversity in the order Gruiformes is reflected in the uncertainty of taxonomists about gruiform relationships. The most widely accepted classification is that of the American paleontologist Alexander Wetmore, used above, with 12 Recent families in one order. Some authorities place the plains wanderer (*Pedionomus*) in Turnicidae. A German ornithologist, Erwin Stresemann, underlining the differences between the gruiform families, divided them into ten orders, most of them monotypic (*i.e.*, with only a single form). A British anatomist, P.R. Lowe, dismayed by the wide variety of characters in the Gruiformes, eliminated the order altogether, distributing the families among other orders, a decision later repeated by a French worker, R. Verheyen, who divided the families among five orders, some of which are not recognized today. Verheyen's order of Gruiformes contained only the Psophiidae, Aramidae, Gruidae, and Otidae. In an attempt to resolve this confusion, an American biochemist, H.T. Hendrickson, studied the egg-white proteins of 10 of the 12 gruiform families (he had no material for Pedionomidae or Mesitornithidae) and found that a close group was formed by five families—Eurypygidae, Heliornithidae, Rallidae, Turnicidae, and Psophiidae. The Aramidae bridge the gap between this group and the Gruidae, but the remaining families (Rhynochetidae, Cariamidae and Otidae) seemed to be very different, the last so different that they may be of independent origin from the rest of the order.

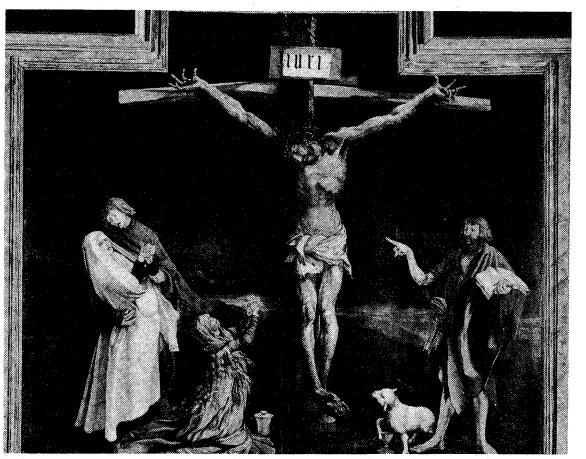
BIBLIOGRAPHY. A.C. BENT, "Life Histories of North American Marsh Birds," *Bull. U.S. Natn. Mus.* no. 135 (1926), contains long and comprehensive accounts of North American cranes and rails and the limpkin; H.F. WITHERBY *et al.* (eds.), *The Handbook of British Birds*, vol. 4 and 5 (1940–41), provides excellent accounts of European cranes, rails and bustards, with coloured plates; R.P. ALLEN, *The Whooping Crane* (1952), is the classic monograph on this endangered species, with a brief descriptive and distributional treatment of the cranes of the world; L.H. WALKINSHAW, "The Sandhill Cranes," *Bull. Cranbrook Inst. Sci.* no. 29 (1949), contains detailed information on this species; H.E. HOWARD, *A Waterhen's Worlds* (1940), is a monograph of the common gallinule or moorhen, with emphasis on behaviour, one of the early classics in ethology; S. KEITH, C.W. BENSON, and M.P. STUART IRWIN, "The Genus *Sarothrura* (Aves, Rallidae)," *Bull. Am. Mus. Nat. Hist.*, vol. 143, art. 1 (1970), is a study of this genus of African rails, with colour plates of the birds, diagrams of vocalizations, and a general treatment of rail voices; H.T. HENDRICKSON, "A Comparative Study of the Egg White Proteins of Some Species of the Avian Order Gruiformes," *Ibis*, 3:80–91 (1969), contains an exhaustive listing of the taxonomic characters for the order. See also the general works cited in the bibliography for BIRD, many of which have extensive text sections and fine illustrations of gruiforms.

(G.S.Ke.)

Grünewald, Matthias

One of the most fascinating and, in many ways, enigmatic German artists of the 16th century, Matthias Grünewald was also, perhaps, the greatest painter of his time. Since his death, Grünewald's painterly and expressive achievement remains one of the most striking in the history of art. The name by which the artist is known today is a 17th-century fabrication, but the ten or so paintings (some of which are composed of several panels) and approximately 35 drawings that survive have been jealously guarded and carefully scrutinized in modern times.

Although it is commonly agreed that "Master Mathis" was born in Würzburg, Germany, the date of his birth remains problematic; estimates vary from 1455 to 1480. In all probability he is not identical with another painter, Mathis, active in Aschaffenburg from about 1480 to 1490. His first securely dated work, the "Mocking of Christ," seems to be that of a young man just become a master, not of an older painter more than 50 years of age. Thus, Grünewald, whose real surname was Gothardt, appears first in documents of about 1500 either in the town of Seligenstadt am Main or Aschaffenburg.



"Crucifixion," centre panel (wings closed) of the "Isenheim Altarpiece" by Matthias Grünewald, before 1516. In the Musée d'Unterlinden, Colmar, France. 2.69 m × 3.07 m.
Giraudon

Court painter at Mainz

By about 1509 Grünewald had become court painter and later the leading art official (his title was supervisor or clerk of the works) to the archbishop of Mainz, Uriel von Gemmingen. The archbishop was in residence at Aschaffenburg, just a three-hour walk down the river from Seligenstadt. Grünewald supervised the rebuilding of the cleric's castle at Aschaffenburg, which, among other projects, included an elaborate chimney.

In about 1510 Grünewald received a commission from the Frankfurt merchant Jacob Heller to add two fixed wings to the altarpiece of the "Assumption of the Virgin" recently completed by Albrecht Dürer. These wings depicting four saints (today divided between the Städelsches Kunstinstitut, Frankfurt, and the Staatliche Kunsthalle, Karlsruhe) are painted in grisaille (gray on gray) and already show the artist at the height of his powers. Like Grünewald's drawings, which are done primarily in black chalk with some yellow or white highlighting, the Heller wings convey coloristic effects achieved without the use of colour.

In about 1515 Grünewald was entrusted with the largest and most important commission of his career. Guido Guersi, the Italian preceptor, or knight, who led the religious community of the Antonite monastery at Isenheim (southern Alsace), asked the artist to execute a series of painted wings for the shrine of the high altar that had been carved in about 1505 by Niclaus Hagnower of Strasbourg. It will no doubt always remain a matter of conjecture why the Antonite cleric entrusted such a large and important task to an artist who had never before been active beyond the borders of his native province. No doubt Guersi had some say in the choice and representation of subject matter, but the Isenheim wings also provided Grünewald's genius with one of its fullest expressions, here based, to a large extent, on the text of the popular, mystical *Revelations* of St. Bridget of Sweden (*c.* 1370).

Another important clerical commission came from a canon in Aschaffenburg, Heinrich Reitzmann. As early as 1513 he had asked Grünewald to paint an altar for the Mariaschnee Chapel in the Church of SS. Peter and Alexander in Aschaffenburg. The artist turned to this work in the years 1517–19 (the altar is dismembered and fragmentary today, one panel being in Stuppach, Württemberg, another in Freiburg im Breisgau).

Grünewald apparently married in about 1519, when he would have been almost 50 years old. The marriage does not appear to have brought the artist much happiness (at

least, that is the tradition recorded in the 17th century). His wife did bring with her a son, Andreas, whom Grünewald adopted. After his marriage Grünewald also occasionally added his wife's surname, Neithardt, to his own, thereby accounting for several documentary references to Mathis Neithardt or Mathis Gothardt Neithardt, as well as a monogram intertwining the letters M, G, and N.

In 1514 Uriel von Gemmingen had died, and Albrecht von Brandenburg had become the reigning prince (in 1518 he became a cardinal as well). Although Grünewald may have left the court for several years, his last works show that he was in the prelate's service for about ten years. The costly robes and even jewels of a courtier were listed among the artist's possessions at his death, so that he must have participated, to some extent, in the life of one of the most worldly and cultivated courts of Germany.

For Albrecht, Grünewald did execute one of his most luxurious works, portraying "The SS. Erasmus and Mauritius" (Erasmus is actually a portrait of Albrecht) that is now in the Alte Pinakothek in Munich. Apparently, because of his sympathy with the Peasants' Revolt of 1525, Grünewald left the Cardinal's service in 1526. He spent the last two years of his life visiting in Frankfurt and Halle, cities sympathetic to the newly emerging Protestant cause. In Halle he was involved in supervising the town waterworks. Grünewald died in August 1528; among his effects were discovered several Lutheran pamphlets and documents.

Reputation and influence

Despite his artistic genius, failure and confusion no doubt marked much of Grünewald's life. For instance, the chimney at Aschaffenburg and the fountain at Halle never did work properly. He seems not to have had a real pupil nor, perhaps, even an assistant. His avoidance of the graphic media also limited his influence and renown. In this way Grünewald presents an important contrast to his German contemporary Albrecht Dürer. Grünewald's coloristic effects and dramatic approach to subject matter can perhaps best be observed in his numerous portrayals of the Crucifixion (National Gallery of Art, Washington, D.C.; Kunstmuseum, Basel, Switzerland; "Ishenheim Altarpiece," Musée d'Unterlinden, Colmar, France; Kunsthalle, Karlsruhe, West Germany).

Although Grünewald's works continued to be highly prized, the man himself was almost forgotten by the 17th century. The German painter Joachim von Sandrart, the artist's fervent admirer and first biographer (*Teutsche Akademie*, 1675), was thus responsible for preserving some of the scanty information that we have about the artist, as well as naming him, erroneously and from an obscure source, Grünewald. At the lowest ebb of his popularity, in the mid-19th century, Grünewald was labelled by German scholarship "a competent imitator of Dürer"; however, the late 19th-century and early 20th-century artistic revolt against rationalism and naturalism, typified primarily by the German Expressionists, led to a thorough and, finally, scholarly re-evaluation of the artist's career, and Grünewald's art is today recognized as an often painful and confused but always highly personal and inspired response to the turmoil of his times.

MAJOR WORKS

PAINTINGS: "Lindenhardt Altarpiece" (1503; Parish Church, Lindenhardt, West Germany); "Mocking of Christ" (*c.* 1503; Alte Pinakothek, Munich); "The Crucifixion" (before 1508; Kunstmuseum, Basel, Switzerland); "St. Cyriakus and St. Lawrence" (1511–13; Städelsches Kunstinstitut, Frankfurt); "Isenheim Altarpiece" (before 1516; Musée d'Unterlinden, Colmar, France); "Virgin of Stuppach" (1519; Parish Church, Stuppach, West Germany); "Miracle of the Snow" (1519; Augustinermuseum, Freiburg im Breisgau, West Germany); "The Small Crucifixion" (1519–20; National Gallery of Art, Washington, D.C.); "Crucifixion" and "Christ Bearing the Cross" (part of the "Tauberbischofsheim Altarpiece"; both 1522–23; Staatliche Kunsthalle, Karlsruhe, West Germany); "The SS. Erasmus and Mauritius" (1523; Alte Pinakothek, Munich); "Mourning Over the Body of Christ" (before 1525; Church of SS. Peter and Alexander, Aschaffenburg, West Germany).

DRAWINGS: "Christ on the Cross" (*c.* 1503; Staatliche Kunsthalle, Karlsruhe); Studies for the "Isenheim Altarpiece" including "Study for St. Anthony" (Kupferstichkabinett, Berlin); "Study for St. Anthony" (Kupferstichkabinett, Dresden, East Germany); "Study for St. Sebastian" (2 drawings, Kupferstichkabinetts, Dresden and Berlin); "Kneeling Madonna" (2 versions; Kupferstichkabinett, Berlin); "Woman with Folded Hands" (Sammlung Oskar Reinhart, Winterthur, Switzerland); "Woman with Folded Hands" (2 versions; Freiherr Speck von Sternburg Collection, Lützschena, East Germany); "Study for the Virgin of Stuppach" (Kupferstichkabinett, Berlin); "Man with Folded Hands" (study for St. John of Karlsruhe "Crucifixion"; Kupferstichkabinett, Berlin); "Woman with Folded Hands" (Ashmolean Museum, Oxford); "Kneeling Apostles: SS. James and Peter" (Kupferstichkabinett, Dresden); "Drapery of a Seated Figure" (Smith College Museum of Art, Northampton, Massachusetts); "Kneeling Man Gesticulating" (Kupferstichkabinett, Berlin); "Madonna in the Clouds" (Museum Boymans-van Beuningen, Rotterdam); "St. Dorothy" (Kupferstichkabinett, Berlin); "Weeping Child" and "Screaming Child" (Kupferstichkabinett, Berlin); "Head of a Smiling Woman" (Louvre, Paris); "Self-Portrait(?)" (1529; Universitätsbibliothek, Erlangen, West Germany); "Woman's Head" (Kupferstichkabinett, Berlin); "Kneeling Man and Two Angels" (Kupferstichkabinett, Berlin); "Woman with Two Children" (Kupferstichkabinett, Berlin); "Standing Saint" (Albertina, Vienna); "Nude Man with Trumpet" (formerly Franz Koenigs Collection, Haarlem, The Netherlands); "Three Men's Heads" ("Trias Romana"; Kupferstichkabinett, Berlin); "Profile of a Bearded Man" (Schlossmuseum, Weimar, East Germany).

BIBLIOGRAPHY. H.A. SCHMID, *Die Gemälde und Zeichnungen von Matthias Grünewald* (1911), the first, still basic biography of the artist; H. FEURSTEIN, *Matthias Grünewald* (1930), another standard German monograph, especially important for its emphasis on the literary sources of Grünewald's imagery; ARTHUR BURKHARD, *Matthias Grünewald: Personality and Accomplishment* (1936), the most important monograph in English, but limited; W.K. ZULCH, *Der historische Grünewald: Mathis Gothardt-Neithardt* (1938), an important monograph with many new documentary sources; GUIDO SCHOENBERGER (ed.), *The Drawings of Mathis Gothart Nithart, Called Grünewald* (1948), the first important work on Grünewald's drawings in English; LOTTLISA BEHLING, *Die Hanzeichnungen des Mathis Gothart Nithart, genannt Grünewald* (1955), the most complete, modern scholarly treatment of the drawings; NIKOLAUS PEVSNER and MICHAEL MEIER, *Grünewald* (1958), a well-illustrated study in English; J.K. HUYSMANS and EBERHARD RUHMER, *Grünewald: The Paintings* (1958), another recent, well-illustrated monograph on the paintings; EBERHARD RUHMER, *Grünewald: Drawings* (1970), the most recent scholarly treatment of the drawings in English.

(C.S.Ha.)

Guadalquivir River

The Guadalquivir River (Arabic Wādī al-Kabīr, "Great River") is, as its name implies, the historic major river of southern Spain, with a drainage area of 22,160 square miles (57,390 square kilometres), and a length of 408 miles (657 kilometres). Although, judged by size alone, it is only the second river of Spain and the fourth of the Iberian Peninsula, it has several distinctive characteristics that, taken together, make it of outstanding importance: its natural environment encompasses one of the richest and most varied areas of plant and animal life in Europe, and its irrigative capacity, particularly in its wide and fertile plain, has, in addition to supporting the world-famous vineyards and olive groves of Andalusia, provided a basis for an industrializing region of national and continental importance. The significance of the river over the last two thousand or more years is suggested by the detailed geographical descriptions of it found in the works of Classical and medieval Greek, Roman, Arabic, and Christian writers. This tradition is continued today in the scientific research of Comisión Hidrográfica del Guadalquivir, the Sociedad de Ciencias Naturales "Aranzadi," the Real Sociedad Geográfica, and many other public and private bodies.

The course. The Guadalquivir divides into three regions, illustrating its development from a turbulent mountain stream (its "youth" to geographers), through "maturity," to the wide meanderings of its final floodplain ("old age"). It rises almost 5,250 feet (1,600 metres) above sea level, in Jaén Province of southern Spain, to the north of the Sierra Nevada Mountains fronting the Mediterranean Sea.

Geologic structure of the river's course

After rising, it flows northward for about 30 miles (50 kilometres), running through a narrow valley cut into soils and rocks of the Mesozoic Era of as much as 225,000,000 years ago. After emerging from the man-made lake of El Tranco de Beas, it describes a great arc to the west, following a westerly or southwesterly trend all the way to the Atlantic. Between the cities of Andújar and Montoro, the river leaves the Mesozoic rocks to cut a widening plain in more ancient strata. It flows over Mesozoic rocks again near Marmolejo, and, just before passing the city of Cordoba, the beautiful capital of the medieval Arabian caliphate, runs across the easily eroded rocks and soils of the geologically more recent Tertiary and Quaternary periods (laid down less than 65,000,000 years ago), which it then follows to its mouth. After irrigating the fruitful regions of Posadas and Lora del Río, the Guadalquivir reaches Seville, the capital of Andalusia, and one of the most historic and attractive of Iberian cities. It has been made navigable from Seville to the Atlantic (about 50 miles [80 kilometres]), notably by the construction of the channel of the Corte de Tablada, which will take vessels of up to 30,000 tons. Here the river meanders lazily across a hot coastal plain, traversing the swamps of Las Marismas, the largest in Spain, before entering the Atlantic at Sanlúcar de Barrameda, at the eastern end of the Gulf of Cadiz.

Plant life. The drainage basin of the Guadalquivir encompasses one of the greatest floral resources of Europe, containing representatives of half of the continent's species of plant life, together with virtually all those of the North African subtropical region. The surrounding mountains are crowded by forests of pines and oaks, and by brushwoods. More than a third of the total surface is covered by innumerable olive groves, which annually produce more than 15,000 tons of oil and help to make Spain the leading world producer of this vegetable oil. In addition, cereals (wheat, rye, and barley), cultivated in 30 percent of the basin area, support the regional agriculture, and viticulture produces the great Spanish sherries, such as the amontillados (pale dry sherries), and other wines at a rate of 5,600,000 cubic feet (1,600,000 hectolitres) per year.

Commercial plants

Animal life. Fauna is as varied as the plant life, with animals representing a great variety of European and North African species. In the mountains, wild boar, goat, fallow deer, stag, chamois, partridge, among many others, are found, and the area is one of the great European hunting regions. Fish, notably trout and barbels, are found throughout the Guadalquivir, its artificial lakes, and its more than 800 tributaries.

Regional development. Traditionally, agriculture was the basic way of life of the Guadalquivir Basin, but, by the 1970s, the region was experiencing an accelerating industrialization spurred by government plans for development, one of which has introduced industry to the upper basin at a cost of over 5,000,000,000 pesetas ($70,000,000). As a result of this development, the percentage of the labour force employed in agriculture has declined from 60 percent to 40 percent in recent decades. In addition, there has been a spectacular growth in hydroelectric capacity. By the early 1970s, there were over 80 hydroelectric stations in the region with an installed power of 288,000 kilowatts. The six largest dams, with associated lake areas, were El Tranco de Beas, 6.8 square miles (1,772 hectares); Charco Manzano (Guadalén), 5.2 square miles (1,351 hectares); La Lancha (Jándula), 5.2 square miles (1,350 hectares); Bembézar, 4.1 square miles (1,070 hectares); Torre del Aguila, 4.1 square miles (1,066 hectares); and El Pintado, 4.0 square miles (1,042 hectares). (J.B.Ro.)

Guadeloupe

Guadeloupe (La Guadeloupe), a French overseas *département* (large administrative district), is a group of islands situated in the eastern Caribbean Sea. Most of the islands form a part of the Windward Islands, and all are included in the Lesser Antilles island chain. Guadeloupe is about 4,300 miles away from France and about 370 miles north of the coast of Venezuela. The nearest territorial neighbours of the principal group are the British dependencies of Montserrat to the northwest and Dominica to the south. The largest territory in Guadeloupe consists of the twin islands of Basse-Terre to the west and Grande-Terre to the east, the two being separated by a narrow channel, the Rivière Salée; other islands in the group are Marie-Galante to the southeast, La Désirade to the east, and the Îles des Saintes (Terre de Haut and Terre de Bas) to the south. Two more island dependencies—Saint-Barthélemy and Saint-Martin (the southern third of the latter being under The Netherlands administration)—are situated about 150 miles to the northwest, lying to the northwest of the outer arc of the Lesser Antilles.

The total area of Guadeloupe is 687 square miles (1,780 square kilometres). The population numbered more than 300,000 in 1970. Basse-Terre (population 15,000), on the island of the same name, is the seat of government. The largest town, however, is Pointe-à-Pitre on Grande-Terre, a port with a population of 50,000, which is the economic capital. Guadeloupe lies about 74 miles north of Martinique, the other French overseas *département* in the Windward Islands. (See also MARTINIQUE; for an associated physical feature, see CARIBBEAN SEA.)

Relief and drainage. Basse-Terre, which has an area of 364 square miles, has a chain of mountains running north to south and culminating in Soufrière, a dormant volcano 4,868 feet (1,484 metres) high; it erupted in 1797 and in 1836 and now is a source of hot and sulfur springs. Other summits are Mont Sans Toucher, which is 4,442 feet (1,354 metres) high, and Morne de la Grande Découverte, which is 4,143 feet (1,263 metres) high. The mountain chain forms a watershed from which rivers run down to the sea. The Rivière Rouge (Red River) is so called because iron-bearing sediments give it a rust colour; the place where it falls into the Sablon River is called Matouba, or Matouba's Leap. The principal river on the island is the Goyaves; other rivers are the Grande Plaine, the Petite Plaine, the Moustique, the Lézarde, and the Rose. Basse-Terre has a beautiful coastline, indented with bays and fringed with picturesque beaches.

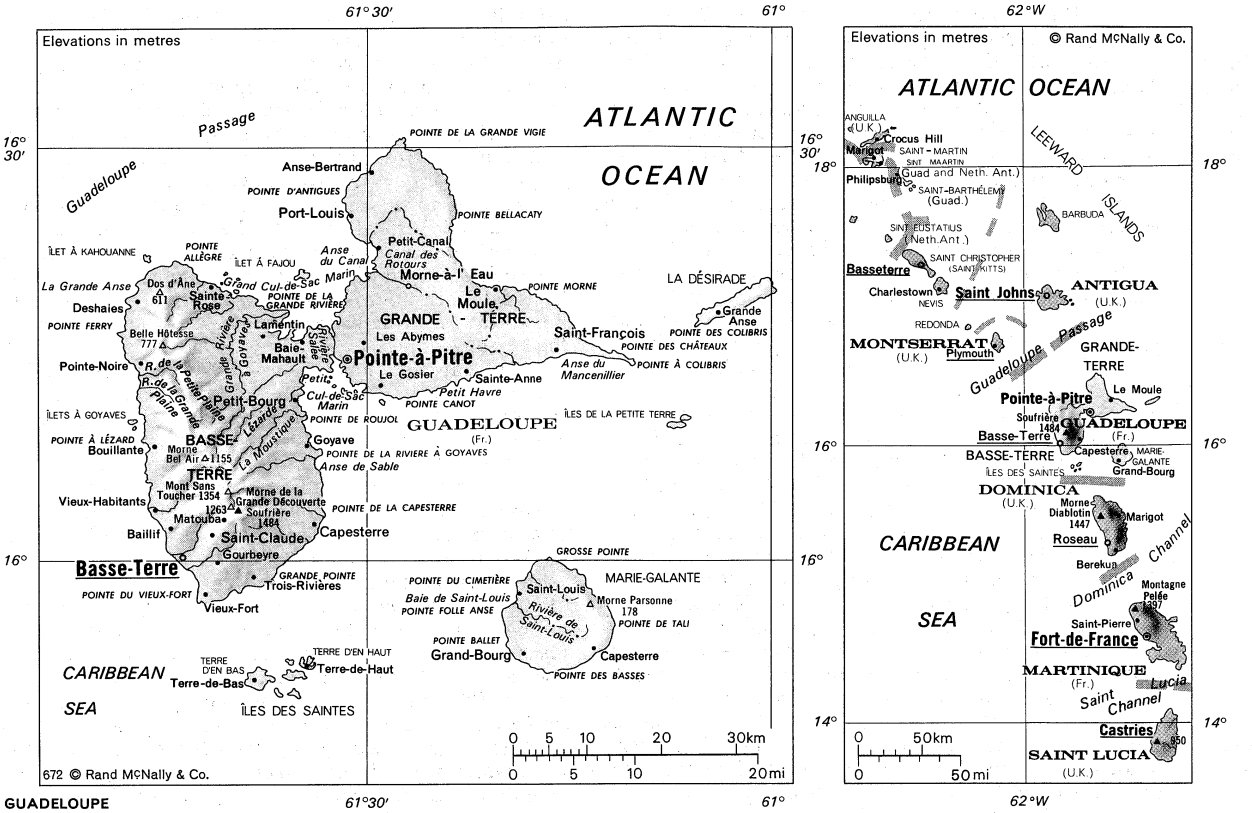
Principal rivers

The island of Grande-Terre has an area of 220 square miles and is generally low lying; it has only a few bluffs that do not exceed 490 feet in height. Saint-Martin and Saint-Barthélemy are rugged and rise to an altitude of 1,391 feet (424 metres) and 921 feet (281 metres), respectively.

Climate. The tropical climate is tempered by the northeast trade winds. The temperature on the coast varies between 77° and 82° F (25° and 28° C), respectively, with extremes reaching 68° and 93° F (20° and 34° C). In the mountains above 1,900 feet, the temperature drops to 61° F (16° C), and at the summit of Soufrière drops to 39° to 41° F (4° to 5° C). There are two distinct seasons—the "Creole Lent," or dry season, which lasts from December to April, and the winter, or rainy season, from July to September–October.

Precipitation varies with altitude and orientation. Grande-Terre receives about 39 inches of rain, for example, while the mountainous parts of Basse-Terre receive more than 100 inches a year. Hurricanes occur occasionally, usually coming from the south.

Vegetation and animal life. The heat, the rainfall, and the fertility of the volcanic soils produce a luxuriant vegetation that is diversified according to altitude. Extensive mangrove swamps occur on the banks of the Rivière Salée near Pointe-à-Pitre. Dense forest occurs in the mountainous regions of Basse-Terre, beginning almost at sea level on the windward slopes and at an altitude of about 750 feet on the leeward side and continuing to altitudes of about 3,000 feet or more. Here, chestnut trees and bracken are found, as well as such hardwoods as mahogany and ironwood. On the highest peaks some flooded basins produce a vegetation of grasses and sedges. Grande-Terre has been cleared of most of its original forests; only a few patches of woodland remain. The smaller islands, such as La Désirade and Saint-Martin, have a different type of vegetation, consisting



primarily of dry forest with groves of latania (a kind of fan palm) and cactus.

Animal life has been modified since the arrival of the Europeans. Raccoons are sought for their fur. The agouti (a short-haired, short-eared, rabbit-like rodent) is still found on the heights of Capesterre, southeast of Basse-Terre. The paucity of game animals is due—apart from excessive hunting—to the presence of mongooses, which abound throughout the islands. In some regions, wild duck, waterfowl, and teal are found.

The warmth of the water around the islands is responsible for a rich variety of fish life, including tarpon, snook (a basslike kind of fish), hogfish, snapper, parrot fish, and many species of ray fish.

History. Discovered on November 4, 1493, by Christopher Columbus, the two main islands, then together known as Karukera (Island of Beautiful Waters), were peopled by Caribs, who had displaced the original Arawak inhabitants. The territory was consecrated to Our Lady of Guadeloupe of Estremadura in Spain, from whom it takes its name.

Early European settlement

Preliminary attempts by the Spanish to establish themselves were repulsed by the Caribs in 1515, 1520, and 1523. In 1626 the Spanish, who had established themselves on the coast, were driven away by Pierre Belain d'Esnambuc, a Frenchman who established a trading company. In 1635 two Frenchmen, Léonard de L'Olive and Jean Duplessis d'Ossonville, landed and established a colony. Until 1640 the colonists fought against the Carib Indians, but thereafter the colony prospered. Made a part of the royal domain of France in 1674, Guadeloupe benefitted from the influence of Jean-Baptiste Labat (1663 to 1738), a strong personality who was the effective founder of the Basse-Terre colony and who, in 1703, armed the black slaves (who had already been brought to the island) in order that they might fight against the English; he also established the first sugar refineries, thereby laying the foundations for the era of prosperity that followed. In 1759 Guadeloupe was occupied by the British for four years but was restored to France in 1763. In 1794 it was again occupied by British troops, allied with French royalists, but was recaptured by Victor Hugues, an official of the French revolutionary government, who proclaimed the abolition of slavery and had several hundred white planters massacred. When slavery was re-established by Napoleon's government in 1802, a revolt of the slaves occurred and culminated in the heroic act of the antislavery forces, who blew themselves up at Matouba when threatened by French forces under the command of Gen. Antoine Richepanse; Richepanse himself had been sent by Napoleon to pacify Guadeloupe, but he died of yellow fever in the same year. The British occupied Guadeloupe in 1810, after which it was—after some temporary changes in status—finally restored to France in 1816.

The abolition of slavery in 1848 was the most significant development of the territory's 19th-century history. Universal suffrage was abolished during the reign of Napoleon III of France, but in 1870 colonial representation in the French Parliament was restored. In 1940 Guadeloupe gave its allegiance to the Vichy government in France, but in 1943 it adhered to Gen. de Gaulle's Free French forces. In 1946 it was given the status of a French *département.*

The contemporary islands. *Population and demography.* The population is composed principally of Creoles (*i.e.*, persons born in the islands), most of whom are black or mulatto, except on Les Saintes where the inhabitants are mainly white. The diminution of the white element during the period of the French Revolution accentuated the African character of the population. The white population on the smaller islands is mostly descended from 17th-century Norman and Breton settlers.

The population numbered 229,000 in 1954 and rose to about 313,000 at the time of the 1967 census, and to 326,000 in 1970, representing a density of 474 persons per square mile. The average rate of increase of the population amounts to about 2 percent a year. About 300,000 people live on the two largest islands, while about 20,000 people live on Marie-Galante, 3,000 on the Îles des Saintes, 1,600 on La Désirade, 2,400 on Saint-Barthélemy, and 5,000 on Saint-Martin.

While French is the official language and is in current use, a local Creole dialect is also widely spoken.

The economy. The economy is marked by a stagnant agricultural sector, an embryonic industrial sector, and a highly developed public-service sector. In effect, the economy is sustained primarily by the salaries of officials

Guadeloupe, Area and Population

	area		population	
	sq mi	sq km	1967 census	1970 estimate
Arrondissements				
Basse-Terre	370	957	137,000	...
Pointe-à-Pitre	289	749	168,000	...
Saint-Martin–Saint-Barthélemy*	28	73	7,000	...
Total Guadeloupe	687	1,780†	313,000‡	326,000

*The *arrondissement* consists of the dependencies of Saint Barthélemy and the northern part of Saint Martin (the southern part belongs to The Netherlands Antilles). †Area figures do not add to total given because of rounding. ‡Figures do not add to total given because of rounding.
Source: Official government figures.

and by French credits, which consist of aid in the form of allocations and grants.

Sugarcane and bananas

Sugarcane and bananas are the two principal cash crops. Sugar exports amount to between 130,000 and 190,000 tons a year. The banana plantations suffered from a series of hurricanes in the 1960s, but the devastation resulted in the replanting of plantations with more productive types of trees. By the late 1960s, banana exports had increased to more than 100,000 tons a year, all of which were sent to France. Rum is also exported.

Industrial development is limited, largely because of the shortage of power. The rivers are too short to permit hydroelectric development. In 1969, however, liquid steam was discovered at a depth of almost 1,000 feet; the steam has a pressure of 104 pounds, a temperature of 482° F (250° C), and a yield of 100 tons an hour. The establishment of a geothermic plant is now considered feasible. Meanwhile, electricity is provided by a semi-public corporation that has four power stations.

There is a severe deficit in the balance of external trade, most of which is with France and the franc zone. Most imports are consumer goods.

Transportation. Guadeloupe maintains regular air and sea links with France and with the North American continent. Traffic at the port of Pointe-à-Pitre is constantly growing; in 1969 alone it was visited by more than 900 ships. The port of Basse-Terre specializes in the banana-export trade. Le Raizet, north of Pointe-à-Pitre, is an international airport used by French, United States, British, and Dutch airlines. On the island of Saint-Martin, the town of Marigot is a free port; there is also an airport at Philipsbourg in the Dutch sector that serves both parts of the island. Local steamers connect Basse-Terre and Grande-Terre with the other island dependencies. The road system on the main islands, which is about 780 miles in length, is kept in excellent condition.

Air services

Administration. The *département* is under the executive authority of a prefect; there is a legislative council consisting of 36 elected members. Guadeloupe is represented in the French National Assembly by three deputies and in the French Senate by two senators.

The territory of Guadeloupe is divided into three *arrondissements* (wards) that are in turn divided into 34 *communes* (the smallest territorial divisions), each administered by an elected municipal council. The judicial system is French. There is a court of appeal at Basse-Terre and a judicial tribunal at Pointe-à-Pitre; justices of the peace are established in each of the 36 cantons (electoral districts).

Social conditions and cultural life. The same social legislation is in effect as in metropolitan France. There is a general hospital at Pointe-à-Pitre, as well as a Pasteur Institute, and other hospitals.

Education

French is the medium of instruction in the schools, and the literacy rate is over 83 percent. In addition to about 300 primary schools, there are two *lycées* (state-supported secondary schools) as well as a teachers' training college and a school of agriculture. A school of humanities and a law school, at Pointe-à-Pitre, constitute in embryonic form what is planned to become the University of the Antilles.

There are about a dozen daily or weekly newspapers. Radio and television programs are broadcast regularly by ORTF (Office de Radiodiffusion Télévision Française) from Pointe-à-Pitre.

Future prospects. Guadeloupe's primary problem is economic and arises from the fact that whereas on the one hand the territory has a rapidly increasing population, on the other its main economic sector is the public service—a situation that results in the *département* consuming more goods than it produces. Tourism is insufficient to bridge the gap, and emigration to France has also proved an insufficient palliative in a situation in which employment opportunities remain limited at a time when the younger generation is becoming increasingly well educated. Problems stemming from these circumstances are expected to grow more acute.

BIBLIOGRAPHY. GUY LASSERRE, *La Guadeloupe: étude géographique*, 2 vol. (1961), is the basic work. See also HENRI BANGOU, *La Guadeloupe, 1492–1848* (1962); and ANTOINE VICTOR JOROND, *La Guadeloupe et ses îles: guide pratique du visiteur* (1965).

(R.Co.)

Guatemala

Guatemala is the third largest of the Central American republics, bordered on the north and west by Mexico, on the east by Honduras and El Salvador, and on the south by the Pacific Ocean. Its northeast border is with British Honduras. The country also has a small coastline on the Caribbean (Gulf of Honduras). Its total area is 42,042 square miles (108,889 square kilometres). The population at the beginning of the 1970s was about 5,600,000. The capital is Guatemala city, the largest city in Central America.

The oldest known inhabitants of Guatemala were a sedentary, agricultural people who made utensils of baked clay. Between the 2nd and 10th centuries AD, these early inhabitants produced the great Mayan civilization, the ruins of which may be seen in the jungles of Petén and the neighbouring Yucatán Peninsula of Mexico. The descendants of the Mayas mixed with other Indian tribes that came to the country after the 10th century and formed small kingdoms, which were later subdued by the Spanish conquerors of the 16th century. (For historical aspects, see CENTRAL AMERICAN STATES, HISTORY OF. For Guatemala's claim to British Honduras, see BRITISH HONDURAS [BELIZE].)

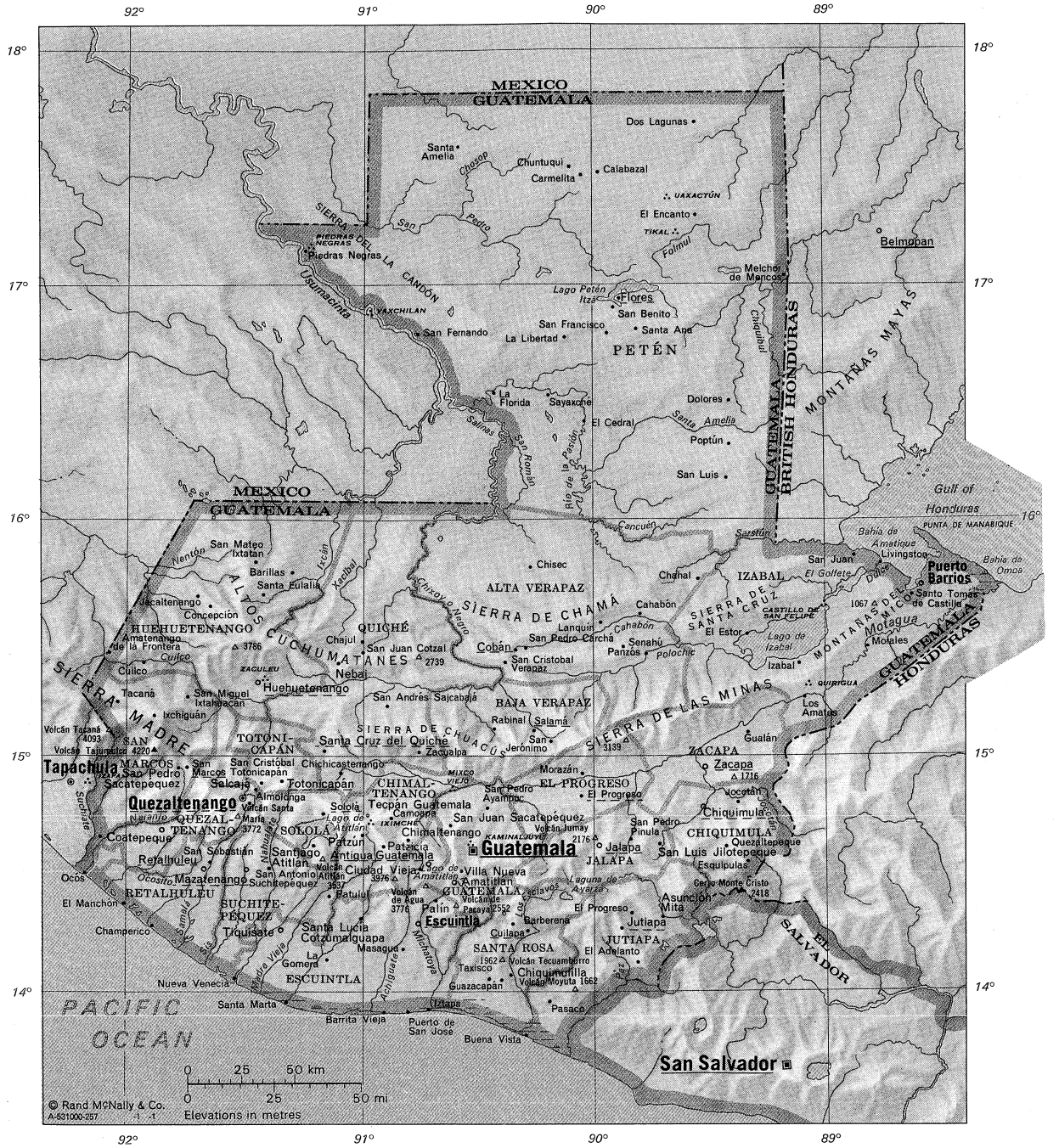
THE LANDSCAPE

Physical geography. The mountain ranges, or cordilleras, of Central America cross Guatemala from west to east, dividing the country into four regions: the Petén, the central highlands, the Atlantic littoral, and the southern coast.

The four regions

The Petén is located between Mexico and British Honduras (Belice) in the northern part of the country and has the same geological characteristics as the rest of the Yucatán Peninsula belonging to Mexico. It is a limestone tableland with altitudes of no more than 650 feet, except in the southern part, extending over about a third of the country. It is relatively unpopulated, covered in large part by jungles and some flat grassland. Petén is dotted with lakes and ponds; there are also many water holes and small ponds that form during the rainy season and drain into underground caverns. Some of the most impressive Mayan ruins (Tikal, Uaxactún, and Piedras Negras) lie sleeping in the jungles of Petén.

The central highlands comprise two branches of the cordilleras (systems of approximately parallel mountain ranges): the Sierra Madre and the Altos Cuchumatanes. The Altos Cuchumatanes, which in the western part of the country reach elevations as high as 9,500 feet, extend northeast (where they are known as the Sierra Chamá, the Sierra de Santa Cruz, the Montañas Mayas, the Sierra de Chuacús, and the Sierra de las Minas), until they drop into the Atlantic Ocean. The Sierra Madre crosses the country from west to east, entering Honduras and El Salvador; it forms the central plateau upon the terraces, valleys, and slopes of which live the greater part of the country's population and where the capital of the republic is located. The southern range of the Sierra Madre is



GUATEMALA

MAP INDEX

Political subdivisions

Alta Verapaz....15·40n 90·00w
Baja Verapaz....15·05n 90·20w
Chimaltenango...14·40n 90·55w
Chiquimula.....14·40n 89·25w
El Progreso.....14·50n 90·00w
Escuintla.....14·10n 91·00w
Guatemala.....14·40n 90·30w
Huehuetenango..15·40n 91·35w
Izabal.....15·30n 89·00w
Jalapa.....14·35n 89·55w
Jutiapa.....14·10n 89·50w
Petén.....16·15n 89·50w
Quezaltenango...14·45n 91·40w
Quiché.....15·30n 90·55w
Retalhuleu.....14·20n 91·50w
San Marcos.....15·00n 91·55w
Santa Rosa.....14·10n 90·18w
Sololá.....14·40n 91·15w
Suchitepéquez...14·25n 91·20w
Totonicapán.....15·00n 91·20w
Zacapa.....15·00n 89·30w

Cities and towns

Almolonga.....14·49n 91·30w
Amatitlán.....14·29n 90·37w
Antigua Guatemala.....14·34n 90·44w
Asunción Mita...14·20n 89·43w
Barberena.....14·18n 90·22w
Barillas.....15·48n 91·18w
Barrita Vieja.....13·55n 95·46w
Buena Vista.....13·49n 90·19w
Cahabón.....15·34n 89·49w
Calabazal.....17·29n 89·59w
Carmelita.....17·21n 90·10w
Chahal.....15·45n 89·34w
Chajul.....15·30n 91·02w
Champerico.....14·18n 91·55w
Chichicastenango.....14·56n 91·07w
Chimaltenango...14·40n 90·49w
Chiquimula.....14·48n 89·33w
Chiquimulilla....14·05n 90·23w
Chisec.....15·49n 90·17w
Chuntuquí.....17·31n 90·09w
Ciudad Tecún Umán.....14·40n 92·09w
Ciudad Vieja.....14·31n 90·46w
Coatepeque.....14·42n 91·52w
Cobán.....15·29n 90·19w
Concepción.....15·37n 91·41w
Cuilapa.....14·17n 90·18w
Cuilco.....15·24n 91·58w
Dolores.....16·31n 89·25w
Dos Lagunas....17·42n 89·36w
El Adelanto.....14·10n 89·50w
El Cedral.....16·26n 90·03w
El Encanto.....17·17n 89·34w
El Estor.....15·32n 89·21w
El Manchón.....14·23n 92·02w
El Progreso.....14·51n 90·04w
El Progreso.....14·21n 89·51w
Escuintla.....14·18n 90·47w
Esquipulas.....14·34n 89·21w
Flores.....16·56n 89·53w
Gualán.....15·08n 89·22w
Guatemala.....14·38n 90·31w
Guazacapán.....14·04n 90·25w
Huehuetenango..15·20n 91·28w
Ixchiguán.....15·12n 91·53w
Izabal.....15·24n 89·08w
Iztapa.....13·56n 90·43w
Jacaltenango....15·40n 91·44w
Jalapa.....14·38n 89·59w
Jocotán.....14·49n 89·23w
Jutiapa.....14·17n 89·54w
La Florida.....16·33n 90·27w
La Gomera.....14·05n 91·03w
La Libertad.....16·47n 90·07w
Lanquín.....15·34n 89·58w
Livingston.....15·50n 88·45w
Los Amates.....15·16n 89·06w
Masagua.....14·12n 90·51w
Matias de Gálvez, see Santo Tomás de Castilla
Mazatenango....14·32n 91·30w
Melchor de Mencos.....17·04n 89·10w
Morales.....15·29n 88·49w
Morazán.....14·56n 90·09w
Nebaj.....15·24n 91·08w
Nueva Venecia...14·03n 91·33w
Ocós.....14·31n 92·11w
Palín.....14·24n 90·42w
Panzós.....15·24n 89·40w
Pasaco.....13·59n 90·12w
Patulul.....14·25n 91·10w
Patzicía.....14·38n 90·56w
Patzún.....14·41n 91·01w
Piedras Negras..17·11n 91·15w
Poptún.....16·21n 89·26w
Puerto Barrios...15·43n 88·36w
Puerto de San José.....13·55n 90·49w
Quezaltenango...14·50n 91·31w
Quezaltepeque...14·38n 89·27w

MAP INDEX (continued)

Rabinal.....15·06n 90·27w
Retalhuleu.....14·32n 91·41w
Salamá.....15·06n 90·16w
Salcajá.....14·53n 91·27w
San Andrés Sajcabajá.....15·13n 90·55w
San Antonio Suchitepéquez..14·32n 91·25w
San Benito.....16·55n 89·54w
San Cristóbal Totonicapán....14·55n 91·26w
San Cristóbal Verapaz.....15·23n 90·24w
San Francisco...16·48n 89·56w
San Jerónimo....15·03n 90·12w
San José, see Puerto de San José
San Juan.....15·52n 88·53w
San Juan Cotzal.....15·26n 91·01w
San Juan Sacatepéquez...14·43n 90·39w
San Luis.....16·14n 89·27w
San Luis Jilotepeque.....14·39n 89·44w
San Marcos.....14·58n 91·48w
San Mateo Ixtatán.....15·50n 91·29w
San Miguel Ixtahuacán.....15·15n 91·45w
San Pedro Ayampuc.....14·47n 90·27w
San Pedro Carchá.....15·29n 90·16w
San Pedro Pinula.....14·40n 89·51w
San Pedro Sacatepéquez...14·58n 91·46w
San Sebastián...14·34n 91·39w
Santa Ana.....16·48n 89·50w
Santa Cruz del Quiché.....15·02n 91·08w
Santa Eulalia....15·45n 91·29w
Santa Lucía Cotzumalguapa.14·20n 91·01w
Santa Marta.....13·58n 91·18w
Santiago Atitlán.14·38n 91·14w
Santo Tomás de Castilla.....15·42n 88·37w
Sayaxché.....16·31n 90·10w
Senahú.....15·24n 89·50w
Sololá.....14·46n 91·11w
Tacaná.....15·14n 92·05w
Taxisco.....14·04n 90·28w
Tecpán Guatemala.....14·46n 91·00w
Tiquisate.....14·17n 91·22w
Totonicapán.....14·55n 91·22w
Villa Nueva.....14·31n 90·35w
Zacapa.....14·58n 89·32w
Zacualpa.....15·05n 90·50w

Physical features and points of interest

Achiguate, *river*..13·55n 90·55w
Agua, Volcánde, *volcano*.....14·28n 90·45w
Amatique, Bahía de, *bay*.....15·55n 88·45w
Amatitlán, Lago de, *lake*.....14·28n 90·33w
Atitlán, Lago de, *lake*.....14·41n 91·12w
Atitlan, Volcán, *volcano*.....14·35n 91·11w
Ayarza, Laguna de, *lake*.....14·25n 90·07w
Cahabón, *river*...15·25n 89·36w
Cancuén, *river*...16·00n 89·58w
Castillo de San Felipe, *ruins*....15·39n 89·00w
Chamá, Sierra de, *mountains*...15·35n 90·20w
Chiquibul, *river*..17·02n 89·10w
Chixoy o Negro, *river*.....16·05n 90·25w
Chocop, *river*....17·18n 90·35w
Chuacús, Sierra de, *mountains*...15·05n 90·45w
Cuchumatanes, Altos, *mountains*.....15·40n 91·25w
Cuilco, *river*.....15·26n 92·08w
Dulce, *river*.....15·49n 88·45w
El Golfete, *lake*...15·45n 88·54w
Folmul, *river*.....17·27n 89·10w
Honduras, Gulf of.....16·10n 88·20w
Ixcán, *river*.....16·07n 91·05w
Iximaché, *ruins*..14·44n 90·59w
Jocotán, *river*....14·52n 89·28w
Jumay, Volcán, *volcano*.....14·42n 90·00w
Kaminaljuyú, *ruins*.....14·39n 90·32w
Lacandón, Sierra del, *mountains*..17·05n 90·55w
Izabal, Lago de, *lake*.....15·30n 89·10w
Los Esclavos, *river*.....13·51n 90·19w
Madre, Sierra, *mountains*.....15·15n 92·00w
Madre Vieja, *river*.....14·00n 91·26w
Manabique, Puntade, *point*..15·56n 88·37w
Michatoya, *river*..14·06n 90·39w
Mico, Montañas del, *mountains*..15·38n 88·50w
Minas, Sierra de las, *mountains*..15·10n 89·40w
Mixco Viejo, *ruins*.....14·53n 90·39w
Motagua, *river*...15·44n 88·14w
Moyuta, Volcán, *volcano*.....14·02n 90·06w
Nahualate, *river*..14·03n 91·32w
Naranjo, *river*....14·30n 92·11w
Nentón, *river*.....15·48n 91·52w
Ocosito, *river*....14·30n 92·11w
Omoa, Bahía de, *bay*.....15·40n 88·08w
Pacaya, Volcán de, *volcano*.....14·23n 90·36w
Pacific Ocean....13·50n 92·00w
Paz, *river*.....13·44n 90·10w
Petén Itzá, Lago, *lake*.....16·58n 89·50w
Piedras Negras, *ruins*.....17·12n 91·15w
Polochic, *river*...15·26n 89·20w
Quiriguá, *ruins*...15·18n 89·07w
Pasión, Rio de la, *river*.....16·28n 90·33w
Samalá, *river*....14·12n 91·48w
San Pedro, *river*..17·15n 91·00w
San Román, *river*.....16·21n 90·22w
Santa Amelia, *river*.....16·13n 90·02w
Santa Cruz, Sierra de, *mountains*.....15·40n 89·20w
Santa María, Volcán, *volcano*.14·46n 91·33w
Sarstún, *river*....15·43n 89·15w
Sis, *river*.....14·09n 91·39w
Suchiate, *river*...14·33n 92·15w
Tacaná, Volcán, *volcano*.....15·08n 92·06w
Tajumulco, Volcán, *volcano*.15·02n 91·54w
Tecuamburro, Volcán, *volcano*.14·09n 90·24w
Tikal, *ruins*.....17·20n 89·39w
Uaxactún, *ruins*..17·24n 89·39w
Usumacinta, *river*.....17·15n 91·27w
Xaclbal, *river*....16·06n 90·58w
Yaxchilán, *ruins*..16·53n 90·59w
Zaculeu, *ruins*...15·21n 91·28w

an imposing chain of 33 volcanoes running parallel with the Pacific coast from Mexico to El Salvador. Among them are Tajumulco (13,845 feet [4,220 metres]), Tacaná (13,428 feet [4,093 metres]), Santa María (12,375 feet [3,772 metres]), Atitlán (11,601 feet [3,537 metres]), Acatenango (13,041 feet [3,976 metres]), Agua (12,385 feet [3,776 metres]), Fuego (12,343 feet [3,763 metres]), and Pacaya (8,371 feet [2,552 metres]). Santa María, Fuego, and Pacaya are in permanent activity; the others are either extinct or dormant. Among these volcanoes lie small lakes of great beauty, such as Atitlán and Amatitlán.

The Atlantic littoral is a region of lowlands, containing the country's largest lake, Izabal (with an area of about 228 square miles), and large rivers such as the Motagua, Polochic, Dulce, and Sarstún (Sarstoon). The climate is warm and the rainfall heavy except in a small arid region. The Bahia de Amatique has Guatemala's principal ports. The southern coastal lowland extends 160 miles from Chiapas in Mexico to El Salvador, averaging about 30 miles in width. The soil is of volcanic origin, well-watered by numerous rivers and intense rains.

Geology. The northern part of the country is formed by rocks of the Cretaceous (from 65,000,000 to 136,000,000 years old) and Tertiary (from 2,500,000 to 65,000,000 years old) periods. In the south and east are still more ancient, principally Paleozoic rocks (from 225,000,000 to 570,000,000 years old). Carbonated and calcareous (chalky) deposits predominate; in south Petén, metamorphic (formed by heat and pressure) and igneous (formed by the solidification of molten magma) rocks are found, and ridges representing an ancient rocky mountain system predominate. The central mountainous region is predominantly of volcanic Tertiary rock, covered by recent volcanic layers of lava and ash.

Hydrography. The mountain system determines the hydrographic basins. Rivers emerging on the northern slopes flow into the Gulf of Mexico (Usumacinta and its tributaries) or the Caribbean (Hondo, Belize, Sarstún, Polochic, Motagua). Rivers on the south of the watershed flow into the Pacific (Suchiate, Ocosito, Samalá, Nahualate, Michatoya, Los Esclavos, Paz). The Rió Usumacinta (688 miles long) forms the accepted frontier with Mexico and is the largest of Central America. The Motagua (249 miles) crosses Guatemala from west to east. Both are partially navigable by boats of low draft. The rivers flowing to the Pacific are of shorter length and carry less water; several are almost dry during the dry season but during the rainy season become dangerous torrents.

The river system

Climate. Guatemala is located in the tropic zone, but temperate seas and an irregular terrain provide a diversity of climates. At sea level, mean annual temperatures range between 77° and 86° F (25° and 30° C). In the temperate zones, located at altitudes of 2,000 to 6,000 feet, temperatures range down to 63° F (17° C) in the higher altitudes. At altitudes over 6,000 feet, temperatures may be as low as 55° F (13° C). The coldest months are December and January, but there is little variation in the weather from day to day. There is a rainy season from May to November and a dry season from November to May. On the Atlantic coast, where the winds blow during the whole year from warm Caribbean waters, there is hardly any dry season. In the central region the annual rainfall varies from 80 inches in the high plateaus to less than 40 in the arid section of the eastern part.

Vegetation and animal life. In the north are tropical forests and some flat grasslands. The forests are rich in fine woods, rubber, a variety of palms, and the evergreen chicozapote tree (*Achras zapota*), which produces chicle. Near the southern part of this region, some pine forest may be found. In the low or hot lands, the vegetation is similar. The coastal vegetation also includes mangrove. In higher altitudes the lowland species give way to pines, firs, cypresses, willows, and oaks.

Tropical jungles and evergreen forests

In addition to domesticated cattle, sheep, pigs, and mules, wild animals such as deer, monkeys, and peccaries (nocturnal mammals resembling pigs) are common, especially in the less settled areas. Jaguars, tapirs, and pumas are rarer. Crocodiles are found in the Rió Polochic and manatees (tropical, aquatic, herbivorous mammals) in Lago de Izabal and elsewhere. The bird life of the country is remarkably rich—wild turkeys and ducks, doves, and pheasants abound; one almost extinct bird of magnificent plumage, the quetzal, has been chosen as the national emblem.

THE PEOPLE

Composition. On the basis of cultural traits such as language and dress, the population is divided into two large groups: Indians and Ladinos. The Ladinos are mostly of mixed Hispanic–Indian origin. Indians who

The Indian influence

have adopted Ladino ways are considered Ladinos. The census of 1950 showed 54 percent of the population as Indians and 46 percent as Ladinos. In 1964 the census reported 43 percent Indians and 57 percent Ladinos. The government statistical office does not explore the racial composition of the people, but well over 80 percent of the population is either Indian or of mixed blood. Whites and Negroes are few. Slavery was abolished in 1824, and the black slaves soon mixed with the Indian and Spanish population. Inhabitants of Asian origin are also very few in number. The most traditional Indian groups live in the highland plains, particularly in the west and north. They engage in agriculture, sheep raising, primitive textile industries, and local trade. They also work as labourers and compose the majority of soldiers in the national army.

The official national language is Spanish, but approximately 20 Indian dialects are also spoken. The majority of the latter are of Mayan origin. The most widely spoken Indian languages are Quiché (the language of the *Popol Vuh*, a valuable book written by the Indians during the first half of the 16th century), Cakchiquel, Mam, and Kekchí. Many Indians also speak Spanish.

There is no established church, but the prevailing form of religion in Guatemala is Roman Catholic. There are a few Protestants of various denominations and some Jews.

Demography. The total population of Guatemala was estimated in 1972 at 5,600,000. It is believed to be increasing at a rate of 2.4 percent a year. This growth is due almost entirely to the excess of births (about 39 per 1,000) over deaths (15). According to a 1971 estimate 69 percent of the population lived in rural areas and 31 percent in the cities and towns. Some 56 percent were under the age of 19. The most densely populated areas are Guatemala city, and its environs in the highland plain (over 819,000), and the western part of the south coast. The majority of the population live in small villages.

THE ECONOMY

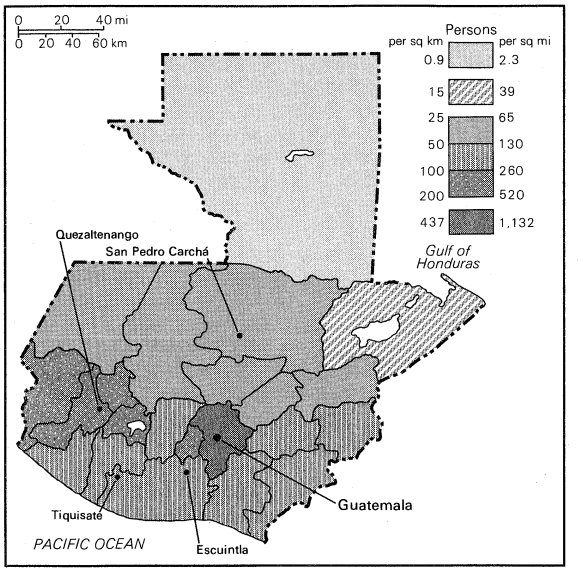
Guatemala is an underdeveloped country, although its natural resources are considerable. Large foreign companies operate in the economy through capital investments made in apparently national companies.

Natural resources. *Mineral resources.* Official surveys of Guatemala's mineral wealth show deposits of coal, gold, silver, antimony, chromium, copper, iron, lead, asbestos, nickel, sulfur, and petroleum. The deposits of nickel are substantial. Mining, however, is little developed. International oil companies have conducted intensive exploration, but in the early 1970s no wells were in operation.

Guatemala, Area and Population

	area		population	
	sq mi	sq km	1964 census	1972 estimate
Departments (*departamentos*)				
Alta Verapaz	3,354	8,686	261,000	317,000
Baja Verapaz	1,206	3,124	96,000	121,000
Chimaltenango	764	1,979	163,000	195,000
Chiquimula	917	2,376	150,000	178,000
El Progreso	742	1,922	65,000	80,000
Escuintla	1,693	4,384	270,000	426,000
Guatemala	821	2,126	810,000	1,160,000
Huehuetenango	2,857	7,400	288,000	359,000
Izabal	3,490	9,038	117,000	182,000
Jalapa	797	2,063	99,000	118,000
Jutiapa	1,243	3,219	195,000	239,000
Petén	13,843	35,854	26,000	36,000
Quezaltenango	753	1,951	271,000	342,000
Quiché	3,235	8,378	250,000	311,000
Retalhuleu	717	1,856	117,000	164,000
Sacatepéquez	180	465	81,000	97,000
San Marcos	1,464	3,791	337,000	423,000
Santa Rosa	1,141	2,955	157,000	196,000
Sololá	410	1,061	108,000	127,000
Suchitepéquez	969	2,510	185,000	238,000
Totonicapán	410	1,061	143,000	176,000
Zacapa	1,039	2,690	97,000	118,000
Total Guatemala	42,042*	108,889	4,287,000†	5,604,000†

*Converted area figures do not add to total given because of rounding.
†Figures do not add to total given because of rounding.
Source: Official government figures.



Population density of Guatemala.

Forest resources. Fifty percent of Guatemala is forested. There are over 12,350,000 acres of tropical forest, and coniferous forests abound on the highland plains. They are little exploited because of their inaccessibility.

Hydraulic resources. Conservative estimates indicate that the country has a large electric-power potential in its rivers and mountains; the National Institute of Electrification, a governmental department, is responsible for their development.

Export crops

Sources of national income. *Agriculture, forestry, and fishing.* The national income is drawn principally from agriculture and cattle raising, which occupy two-thirds of the economically active population. Crops raised for domestic consumption include corn (maize), beans, rice, wheat, sorghum, and potatoes. The leading commercial crops are bananas, cardamom, citronella, coffee, cotton, and sugarcane, which are grown principally for export. The 1969 agricultural census counted 1,376,000 cattle, 200,000 horses, 778,000 pigs, and 800,000 sheep. In 1969 the output of the agricultural and livestock sector was estimated at about a quarter of the gross national product. The forestry industry produces lumber and chicle (a gum used in the manufacture of chewing gum) for export. Fish, crustaceans, and mollusks are also exported.

Manufacturing. Industrial activity has increased in recent years, although in 1969 it still accounted for less than 5 percent of the gross national product. The principal industries are cement, sugar, flour, alcoholic beverages and soft drinks, canned food, textiles, tobacco, tires, medicines, chemical products, lumber, and vegetable oils.

Trade and finance. The United States is Guatemala's chief trading partner, for both imports and exports. Significant trade also exists with the other countries of Central America and with Japan, West Germany, the United Kingdom, and Venezuela. The leading imports include machinery, petroleum, paper, metal products, electrical material, medicines, textiles, and books. Major exports include coffee, cotton, sugar, fruits and vegetables, textiles, and chemicals.

The Bank of Guatemala is the note-issuing authority and controls the country's banking system. It also handles the international accounts. Guatemala's monetary unit, the quetzal (1 quetzal = $1 U.S., and 2.4 quetzales = £1 sterling, on December 1, 1970), is maintained at par with the U.S. dollar.

The country's recent industrial development has been assisted by loans from private foreign concerns.

Guatemala has made several agreements with the Cen-

The Central American Common Market

tral American states of El Salvador, Honduras, Nicaragua, and Costa Rica. Together, in 1957 they created the Central American Common Market (Mercomún) for the purpose of fostering free trade among the member countries and establishing a common tariff on products imported from outside the area. The secretariat of the organization is in Guatemala city. It has had considerable success in eliminating or reducing trade barriers among the members. Trade among them, though still not large, increased considerably by the early 1970s.

Management of the economy. The government taxes income, property, bequests and gifts, imports, and sales. Its major revenues are derived from customs duties, sales taxes, and excises on liquor and tobacco. Local taxes are also levied by municipalities.

Through the Consejo Nacional de Planificación Económica (the National Council for Economic Planning) and other specialized agencies, the government has addressed itself to Guatemala's socioeconomic problems: its predominantly rural population; the unequal distribution of property, particularly land; its high rate of illiteracy and other deficiencies in public education; inadequate communications; and a general lack of economic opportunity for the poor. Development plans under way call for substantial investments in roads, public works, electric plants, housing, education, and land reform, financial and administrative reforms, and encouragement of economic development in areas other than Guatemala city, which is now the only industrial centre.

Transportation. There are more than 5,500 miles of paved and surfaced two-lane highways connecting all parts of the country. There are also more than 2,000 miles of dirt roads that can be used during all seasons of the year. The antiquated narrow-gauge railway system has about 600 miles of track, extending from the Pacific ports to Puerto Barrios on the Atlantic and from the Mexican frontier to El Salvador by way of the capital city. It was taken over by the state in 1968.

The ports of Puerto Barrios and Santo Tomás de Castilla on the Atlantic have excellent facilities for loading and unloading. The Pacific ports of Champerico and San José are of lesser significance. Since they are on the open sea, and vessels cannot approach the piers, special barges must be used for loading and unloading.

There is only one international airport, which is in Guatemala city. It is modern and functional and has adequate landing strips for jet service. Various international airlines service it regularly. The government-owned company Aviateca provides the principal domestic service. There are other airports for domestic flights and for private aircraft.

ADMINISTRATION AND SOCIAL CONDITIONS

Constitutional provisions

Government. *Administration.* The constitution defines the country as a sovereign democratic republic. The government has three branches, legislative, executive and judicial. Legislative power is delegated to a Congress of 55 representatives elected every four years by popular, direct suffrage. Executive power is vested in the president and vice president, who are also elected every four years by popular vote. There is a Supreme Court of Justice, which has jurisdiction over all the tribunals of the country. The constitution contains broad guarantees of human, individual, and social rights.

The national territory is divided into 22 departments, each headed by a governor appointed by the president. Governors are usually high-ranking military officers. The departments in turn are divided into autonomous *municipios* (municipalities), which have freedom to dispose of their own revenues. The *municipios* are governed by councils presided over by mayors, elected directly by popular ballot.

Politics. All men and women over the age of 18 are considered citizens and are obliged to register and to participate in all elections. Voting is compulsory for all who are literate and optional for illiterates. (An official 1967 estimate showed that 60 percent of the population over seven years of age was illiterate.)

Broad guarantees are given for the organization and functioning of democratic parties, except the Communist Party and any other that is dedicated to the overthrow of the democratic process. Authorized political parties must show a minimum of 50,000 affiliates, of which at least 20 percent must be literate. Only those parties may nominate candidates for president, vice president, and Congress. Candidates for mayor and other municipal officers do not have to be nominated by political parties. Some of the constitutional guarantees are occasionally suspended.

Instability of governments

Since the liberal reform of 1871, which modified and modernized the political structure of Guatemala, three fundamental constitutions have been promulgated. Each of these was modified more than once to legalize revolutionary coups d'etat (*golpes de estado*). In that period only three presidents have finished their terms (Manuel Lisandro Barillas, in 1892; Juan José Arévalo, in 1951; and Julio César Méndez Montenegro, in 1970). The leaders of coups usually suspend the constitution and govern by provisional measures granting them executive and legislative powers until a new constitution can be drawn up. The constitution now in force was drafted in 1965 by a constitutional assembly convoked by the government that overthrew Gen. Miguel Ydígoras Fuentes (1958–63).

Justice. The Supreme Court of Justice consists of a president and seven magistrates elected by Congress for a period of four years. The constitution provides for free access to justice for all, based on individual rights and guarantees. There are tribunals of various kinds in all the cities and towns. Municipal mayors act as judges in special cases. The death penalty is applicable only in extraordinary cases and may not be applied to women or minors, to those over 70 years of age, or to political prisoners.

The armed forces. There is no official information on the size of the armed forces. The military services include land, air, and sea forces. Most of the troops are Indians. Although constitutionally outside of politics, the army nevertheless represents a powerful element in political struggles.

Social conditions. *Education.* Education is in theory free, secular, and compulsory through the first six years. There are two-year kindergartens, six-year primary schools, and normal high schools (five to six years). The secondary schools train teachers, agricultural experts, industrial technicians, and candidates for universities. There are also private schools. The system is not adequate to cover all the educational needs of the country, and many children go without schooling. Others are unable to finish their primary education, and only a minority reach the secondary level.

Higher education

There are four universities. The state university (Universidad de San Carlos de Guatemala, founded in 1676) is supported from the national budget. None of the recognized professions can be practiced without a diploma from the national university or one of the private universities. Graduates of foreign universities must take board examinations before they are permitted to practice their professions in Guatemala.

Health and social welfare. The Ministry of Public Health and Social Welfare maintains 36 hospitals, which provide free care. There are several private hospitals in the capital and principal cities. Medical and health services are not adequate for the population as a whole, particularly among the poor and in rural areas. In the early 1970s infant mortality rates were still very high, and the numbers of hospital beds, doctors, and dentists were very low. Most physicians live in Guatemala city and the larger provincial capitals. Along with poor sanitation, malnutrition is an important factor contributing to Guatemala's high mortality.

Since 1946 the Guatemalan Social Security Institute has provided medical insurance for public and private employees. The benefits cover accident and common illness, as well as maternity care. The institute also maintains several hospitals.

Housing. The types and quality of housing vary considerably. The rural population lives in huts made from mud, brick, wood, or cane. In the cities and towns, build-

ing follows traditional methods using modern materials and techniques. In the capital city, where the higher income families live, houses have all the modern comforts. Central heating and air conditioning are not considered necessary.

Housing construction has increased considerably since 1945, particularly in the capital and other important cities, where the population has grown rapidly. The government housing agency is committed to the large-scale construction of homes for low-income families, financed with foreign technical aid and state funds. In 1964, of the nation's 840,000 houses, about 275,000 were in towns and the remainder in rural areas.

Living standards

Social classes. Guatemala's upper income group consists of businessmen, plantation owners, merchants, and bankers. There is a middle stratum of industrialists, professionals, and university graduates. The lowest income group consists of rural workers, labourers, artisans, and small merchants. There are no racial barriers among the different social classes.

Incomes vary greatly. The basic salary paid by the government to teachers was 100 quetzales per month in 1971. It was estimated that 40 percent of the population received salaries of less than 50 quetzales per month, that 31 percent were paid from 50 to 100 quetzales, and that only 8 percent earned over 400 quetzales per month. The government-established minimum wage for urban labourers was 1.50 to 1.60 quetzales per day in 1971. Legal working hours cannot exceed eight per day or 48 per week.

Cultural life. Guatemala belongs culturally to the Hispanic–American world, but the presence of the Indian—as already mentioned, at least four-fifths of the population is either Indian or of mixed blood—has given it a distinct accent of its own. Despite the high illiteracy rate, educational opportunities have greatly expanded since World War II in the capital, in cities of the interior, and even in agricultural areas.

Several state institutions promote the arts. The capital city has the Dirección General de Bellas Artes, the National Conservatory of Music, the national orchestra (Orquesta Sinfónica Nacional), the Ballet Guatemala, various museums, and the School of Arts. There is, however, little artistic activity in the provinces. Guatemala has excellent painters and sculptors, many of whom have exhibited and studied abroad. Indian art is a marvel of colour and taste, particularly the handwoven textiles. The Indian pottery, clay, and wood carvings are also of high artistic quality. Traditional dances, music, religious rites, and games have survived in the Indian regions.

Traditional arts

The Universidad de San Carlos de Guatemala is the major intellectual centre. Other scientific and cultural institutions include the Institute of Nutrition of Central America and Panama, the Institute of Anthropology and History, the National Library, the General Archives, and the Society of Geography and History of Guatemala.

The constitution guarantees freedom of press, but this and other guarantees are sometimes suspended. Most of the newspapers are in the capital city. There are 70 radio stations in operation and more than 200,000 receivers. Transistor radios have become very popular, even among rural labourers. There are three television channels.

Prospects for the future. Guatemala faces problems similar to those of many other underdeveloped countries. On the one hand, there is the need to develop the economy, build schools and housing, and improve living conditions. On the other hand, there is the threat to civil peace from the left-wing guerrilla movement that developed in the 1950s in the mountains of the eastern part of the country; the guerrillas have since then appeared in the capital city and other urban centres. Violent deaths, assaults, and kidnappings of high officials have been frequent.

In this continuing crisis, the future of the country is difficult to predict. The government, elected by popular vote, appears to be strong. Nevertheless, while the social evils that foster subversion persist (unjust distribution of the national wealth, illiteracy, inadequate financial resources, lack of confidence on the part of the people in the government), there seems little possibility of a peaceful solution in the near future to the undeclared civil war.

BIBLIOGRAPHY

Geography: Further data is available in the *Diccionario geográfico de Guatemala*, 2 vol. (1961–62, suppl. 1968), published by the INSTITUTO GEOGRAFICO NACIONAL; ALFREDO GUERRA BORGES, *Geografía económica de Guatemala* (1969); G. DENGO, *Estructura geológica, historia tectónica, y morfología de America Central* (1968); and F.W. MCBRYDE, *Cultural and Historical Geography of Southwest Guatemala* (1947). General geographic aspects are covered in M. MONTEFORTE TOLEDO, *Guatemala: Monografía sociologica* (1959); and in N.L. WHETTEN, *Guatemala: The Land and the People* (1961).

Economy and policy: For information on economic development, see the periodical publications of the Banco de Guatemala and of the Dirección General de Estadística; and in the publications of the Secretaría Permenente de Integración Económica Centroamérica (SIECA); the Instituto Centroamericano de Investigación y Tecnología Industrial, Guatemala (ICAITI), and the Consejo Nacional de Planificación Económica. Critical studies on politics and economics may be found in N. AMARO *et al.*, *El reto del desarrollo en Guatemala* (1970); VILLAGRAN-COLON MIJANGOS, *Bases para el desarrollo económico y social de Guatemala* (1966); and K.H. SILVERT, *A Study in Government: Guatemala* (1954).

The social situation: Books on ethnology, anthropology, and demography are abundant, but a basic study may be found in the collection of 29 titles published by the SEMINARIO DE INTEGRACION SOCIAL GUATEMALTECA, such as, RICHARD N. ADAMS *et al.*, *Cultura indígena de Guatemala* (1956); and *Integración social en Guatemala*, 2 vol. (1956–59), collections of articles by several authors. Other social studies are SOL TAX, *Penny Capitalism: A Guatemalan Indian Economy* (1963); R.N. ADAMS, *Encuesta sobre la cultura de los ladinos en Guatemala*, 2nd ed. (1964); and F. TERMER, *Etnología y etnografía de Guatemala* (1957). See also the periodical publications of the Instituto Indigenista Nacional, Guatemala; and *Cuadernos* of the Seminario de Integración Social.

(J.D.C.R.)

Guerrilla Warfare

Guerrilla warfare is a type of warfare characterized by irregular forces fighting small-scale, limited actions, generally in conjunction with a larger political–military strategy, against orthodox military forces. Guerrillas are usually nondescript in dress, unconventional in weapons and equipment, lack formal supply lines and employ highly unorthodox tactics. In addition to extremely mobile, aggressive operations, these tactics embrace all aspects of psychological warfare, including the use of sabotage and terrorism. Although this type of warfare occurs throughout history, the word guerrilla (the diminutive of Spanish *guerra*, "war") stems from the Duke of Wellington's Iberian campaigns (1809–13), during which Spanish–Portuguese irregulars, or *guerrilleros* (also referred to at the time as partisans and insurgents), helped drive the French from the peninsula. In World War II the word partisan became synonymous with guerrilla; later the word insurgent came into vogue.

Guerrilla warfare is by tradition a weapon of protest employed to rectify real or imagined wrongs levied on a people either by a foreign invader or by the ruling government. As such it may be employed independently or used to complement orthodox military operations, in which case it can be employed either inside enemy territory or in areas that have been seized and occupied by an enemy.

The importance of guerrilla warfare has varied considerably through history. After World War II it came to play a significant role in what has been termed revolutionary or insurgency warfare or what Communists call people's wars and wars of national liberation.

Since World War II, guerrilla warfare has been employed by non-Communist insurgencies in such countries as Indonesia, Cyprus, and Algeria, where it was successful, and by Communist insurgencies in Malaya and the Philippines, where it ultimately failed. In a complementary role, in which the guerrilla force first fights independently and later evolves into an orthodox insurgent army, it has been successfully employed by Communist insurgencies in China, Indochina, and Cuba.

HISTORY

Ancient and medieval chronicles offer countless examples of guerrilla actions, usually of an independent type undertaken by peasant bands and normally resulting in little more than temporary embarrassment to the incumbent ruler or temporary harassment to the invader. These chronicles also describe numerous campaigns undertaken by marauding tribes that practiced an offensive style of warfare often marked with definite guerrilla overtones. The genesis of modern guerrilla warfare, however, is found in the American Revolution, during which the colonists, many of them veterans of Indian fighting, formed loosely knit bands of riflemen that practiced highly unorthodox tactics against the formally trained British redcoats. Despite George Washington's later, more standard approach to warfare, vestiges of the guerrilla tendency remained. In 1780–81 one of the former Indian fighters, Francis ("the Swamp Fox") Marion, organized a ragtag group of guerrillas that complemented orthodox warfare in South Carolina by continual, devastating raids in the rear of British Gen. Charles Cornwallis' lines.

Guerrilla campaigns of the Napoleonic wars

A far more important role, however, was played by Spanish–Portuguese guerrillas in Wellington's campaigns in Portugal and Spain. Throughout this long war, effectively commanded guerrilla bands made life miserable for the French armies by completely disrupting their lines of communication—"by blocking the roads, or intercepting couriers and convoys" and even "waging regular war" (Charles Oman, *A History of the Peninsular War*, 6 vol., 1902). Numbering no more in the field than 20,000 and

> despite their weakness in the open field, their intestine quarrels, their frequent oppression of the countryside, and their ferocity, they rendered good service . . . by pinning down . . . twice their own numbers of good French troops.

In 1812 Napoleon was to suffer heavily from guerrilla strikes during the long retreat from Moscow. Bands of Russian peasants, working with mounted Cossacks, harassed the French until they had been driven out.

Guerrilla warfare, in both its independent and its complementary roles, has been employed extensively since the time of Napoleon. A striking example of the protest role of guerrilla warfare is the Taiping Rebellion (1850–64) in China. Begun by impoverished peasants and by jobless coolie porters, opium smugglers, and pirates, the unsuccessful rebellion against the Manchu dynasty cost an estimated 20,000,000 lives and, in the opinion of some sinologists, constituted one of the great social upheavals of modern times.

Lesser but, nonetheless, significant independent guerrilla actions were fought against the British in India and Africa, the French and Spanish in Morocco, the Turks and Austro-Hungarians in the Balkans, and the Americans in the opening of the West. In 1899–1900 the Boxer Rebellion in China constituted a protest action against a foreign invader (in the form of the Western powers whose influence, ironically, had grown as a result of the Taiping Rebellion), as did both the Philippine insurrection (1899–1902), in which the guerrillas ultimately lost to American regulars, and the South African War (1899–1902), in which the Afrikaners quickly abandoned orthodox tactics in favour of highly mobile, irregular operations undertaken by mounted groups called commandos, which the British regulars defeated only with the greatest difficulty.

Equally impressive is the concurrent record of complementary guerrilla operations. The most successful guerrilla leader in the American Civil War was a Confederate cavalry officer, John Mosby. Leading a small band of mounted volunteers, Mosby so disrupted Union operations by his constant, dashing raids in northern Virginia that Union forces were finally forced to devastate the region in order to deprive the guerrilla force of its base. Mexican guerrillas led by Emiliano Zapata and Pancho Villa played a leading role in the Mexican Revolution.

The static nature of World War I prevented guerrilla warfare on the western front, but subsidiary theatres offered scope for guerrilla activity. In the Middle East a British officer, T.E. Lawrence, led a revolt of Arab tribesmen in a prolonged guerrilla action that claimed the lives of some 35,000 Turkish soldiers and resulted in another 35,000 captured or wounded; the guerrillas finished the war in control of about 100,000 square miles (259,000 square kilometres)—a significant contribution to the British victory in Palestine. In German East Africa (modern Tanzania), a German officer, Lieutenant Colonel Lettow-Vorbeck, led a small force of German regulars supplemented by a few hundred tribesmen in a holding operation against much larger British forces, which could have been used on the western front. Although bereft of supply and physically nearly exhausted, the group had still not surrendered when the war ended.

Guerrilla activity in World War I

Guerrilla fighting of a different nature broke out in southern Ireland, which was in rebellion against British rule. Beginning in Dublin on Easter Monday, 1916, the original insurrection did not prove popular and was instantly put down by the British army garrison. But then the military authorities made a major political and psychological error by court-martialling and condemning to death the 15 principals. The easygoing Irish public was horrified, and the result was a guerrilla war characterized by the most brutal terroristic killings and ambushes, which lasted until 1921.

A different type of guerrilla action was fought in Russia, where in 1918 Lenin's Bolsheviks had taken control of the revolution and were fighting White Russian counterrevolutionary forces supported by various of the great powers, including the United States. This support was probably unwise, since the intervention of foreign powers brought a great wave of patriotism to the Russian masses, many peasants joining partisan movements.

During the interwar period the Communist Party in China and its army, after many vicissitudes and under the leadership of Mao Tse-tung, fought the invading Japanese and, after World War II, wrested control of the country from Chiang Kai-shek's Nationalist government. In Spain the Popular Front left-wing government had to rely primarily on guerrilla forces to hold the Nationalist armies while the government built up conventional forces. The war assumed a distinctively ideological character, the Nationalists being supplied with arms and air power primarily by the Fascist governments of Germany and Italy, the Loyalists being supplied by Russia.

Communist guerrilla activity in World War II

Guerrilla warfare in World War II was also marked with strong ideological overtones. Since Communist parties had been operating, usually clandestinely, in most of the invaded countries, their members were ideally suited for guerrilla warfare. In the West, primarily in France and Italy, the Communists either formed their own bands or joined other bands, such as the French and Belgian maquis, covert organizations engaged in espionage, sabotage, and terrorist activities. Communist cadres in the Balkans and the Far East formed guerrilla bands that usually operated independently, sometimes in competition with guerrillas representing the legal governments, as Tito did in Yugoslavia. Although some of these groups spent as much time eliminating indigenous opposition while consolidating their own hold on the country as in fighting the enemy, they, nonetheless, contributed sufficiently to the war effort to win impressive shipments of arms and equipment from the Western powers—the result of an Allied decision in which the political goal was subordinated to the demands of military strategy. The decision resulted in Communist bands in Yugoslavia, Greece, Burma, Thailand, Indochina, the Philippines, and Indonesia receiving arms from the Allies.

Another important guerrilla action of World War II was fought in the Ukraine. The peasantry, stung by German atrocities, formed into numerous partisan bands. Once these semi-independent bands were organized by the Soviet high command (which never entirely trusted them), they caused widespread and, at times, vital damage to German communications. In the autumn of 1943, in addition to large police and security forces, the German command in the U.S.S.R. was expending 10 percent of its strength—25 field divisions—in fighting the partisans. Although estimates vary, these guerrillas probably killed more than 250,000 German soldiers, while blowing up

thousands of trains and trucks and inflicting an inestimable psychological pressure.

In some countries the Communist guerrilla forces formed in World War II facilitated the quick establishment of Communist regimes. In Yugoslavia the take-over of government was simple and direct, while in other cases, such as Czechoslovakia and China, it was complicated and delayed. In Vietnam it was only partially accomplished; in Malaya and the Philippines it was foiled. Non-Communist insurgencies used guerrilla warfare with considerable success in Cyprus, Kenya, and Algeria. By the early 1970s widespread guerrilla activity was continuing in Indochina (including South Vietnam, Laos, Cambodia, and Thailand). In Latin America the failure of several attempts to emulate the guerrilla campaign that resulted in the overthrow of the Batista regime in Cuba in 1959 led to a concentration on revolutionary activity within the cities, or urban guerrilla warfare, as it came to be termed. The extent to which the intermittent bombings and kidnappings in such countries as Brazil, Argentina, and Guatemala indicate the existence of the discipline, organization, and overall strategy that are the hallmark of guerrilla warfare remains uncertain. But the increasing urbanization of the population does suggest that the cities present the best opportunities for guerrillas to engage in disruption and to engender the lack of confidence in the ability of the government to control the situation that is essential to the success of a guerrilla campaign.

METHODS OF GUERRILLA WARFARE

Strategy and tactics. The broad strategy underlying successful guerrilla warfare is that of protracted harassment accomplished by extremely subtle, flexible tactics designed to wear down the enemy while gaining time either to develop sufficient military strength to defeat him in orthodox battle or to subject him to internal and external political and military pressures sufficient to cause him to seek peace. This strategy embodies political, social, economic, and psychological factors to which the military element is often subordinated. It is essentially a strategy for the morally strong and materially weak.

Writings of Sun-tzu

Many of the essential rules of guerrilla tactics are to be found in *The Art of War* by the Chinese general Sun-tzu (*c.* 350 BC). Sun-tzu instructed his generals in words familiar to successful, latter-day guerrilla leaders: "And therefore I say: Know the enemy, know yourself; your victory will never be endangered. Know the ground, know the weather; your victory will then be total." A successful general "avoids strength and strikes weakness"; the use of tactics based on deception and surprise is the hallmark of a victorious commander (Sun-tzu, *The Art of War*, trans. by S.B. Griffith, 1963).

Sun-tzu's indirect approach was largely ignored in the written commentary on wars of later centuries. Such an approach does appear now and again—Xenophon wrote in the 4th century BC of the importance of psychological factor in warfare, while in the 18th century the French commander Marshal Saxe suggested that it is possible to win a war without fighting battles. Saxe was writing in a time of limited, formal wars, which soon gave way to the total warfare introduced in the Napoleonic era and which was subsequently treated by the Prussian officer-scholar Carl von Clausewitz.

Clausewitz argued that a weaker adversary does not have to destroy the enemy's army in order to gain victory, but rather that he must destroy the other's will to wage war. He argued that partisan warfare could further the wearing process, provided the theatre of operations was large enough, the terrain sufficiently rugged, and the partisans themselves temperamentally suited to this type of fighting. A contemporary, the Cossack Gen. Denis Vasiliyevich Davidov, who led a partisan force during Napoleon's retreat from Moscow, later wrote that this type of warfare "is concerned with the entire area which separates the enemy from his operational base." Its objectives are

> to cut the communication lines, destroy all units and wagons wanting to join up with him, inflict surprise blows on the enemy left without food and cartridges and at the same time block his retreat. This is the real meaning of partisan war (Otto Heilbrunn, *Warfare in the Enemy's Rear*, 1963).

Tactics of T.E. Lawrence

Nearly a century later T.E. Lawrence offered the world a dramatic demonstration of Davidov's definition. Sent to lead dissident Arab tribes in revolt against the Turks, Lawrence followed a Clausewitzian precept by isolating the Arab political aim, which "was unmistakably geographical, to occupy all Arabic-speaking lands in Asia."

Lawrence wrote,

> The Turkish army was an accident, not a target. Our true strategic aim was to seek its weakest link, and bear only on that till time made the mass of it fall. The Arab army must impose the longest possible passive defense on the Turks (this being the most materially expensive form of war) by extending its own front to the maximum. Tactically, it must develop a highly mobile, highly equipped type of force, of the smallest size, and use it successively at distributed points of the Turkish line.

By making the Arabs "an influence, a thing invulnerable, intangible, without front or back, drifting about like a gas," Lawrence would gain "five times the mobility of the Turks [thus] the Arabs could be on terms with them with one-fifth their number."

About the time that Lawrence was incorporating these thoughts into an article for the *Encyclopædia Britannica*, Mao Tse-tung was developing a doctrine of peasant warfare in China. Mao was a young, devoted student of revolution as preached by Marx, practiced by Lenin, and qualified by Mao's own considerable experience. Since 1927 he and a band of comrades had been on the run from the Nationalist Generalissimo Chiang Kai-shek. In the Fukien–Kiangsi borderlands, Mao had helped turn peasants and bandit bands into the first crude Chinese Communist army, one that spent the next eight years fighting for its life against Chiang's forces. Mao's experience had led him to defy his Russian teachers by concluding that the Communist revolution in China could come only from the country peasants, not from the urban proletariat. His theory was tested when pressure from Chiang's armies forced him and his followers to undertake a 6,000-mile (9,656-kilometre) march to the north, to Yen-an, a mountain hideout in Shensi Province.

The rebel leader then began to codify a doctrine of revolutionary warfare. Mao saw two enemies: the Japanese invader and the regular Kuomintang armies headed by Chiang Kai-shek. He looked on a country

> half colonial and half feudal; it is a country that is politically, militarily, and economically backward . . . a vast country with great resources and tremendous population, a country in which the terrain is complicated and the facilities for communication are poor. All these factors favor a protracted war; they all favor the application of mobile [that is, orthodox] warfare and guerrilla operations (Mao Tse-tung, *On Guerrilla Warfare*, trans. by S.B. Griffith, 1961).

Mao's view of the role of guerrilla warfare

Mao came to the conclusion that

> The concept that guerrilla warfare is an end in itself and that guerrilla activities can be divorced from those of the regular forces is incorrect . . . in sum, while we must promote guerrilla warfare as a necessary strategical auxiliary to orthodox operations, we must neither assign it the primary position in our war strategy nor substitute it for mobile and positional warfare as conducted by orthodox forces.

He borrowed freely from Sun-tzu's thesis of the indirect approach:

> Guerrilla strategy must be based primarily on altertness, mobility, and attack. It must be adjusted to the enemy situation, the terrain, and the existing lines of communication, the relative strengths, the weather, and the situation of the people.

It should be used

> to exterminate small forces of the enemy; to harass and weaken large forces; to attack enemy lines of communication; to establish bases capable of supporting independent operations in the enemy's rear; to force the enemy to disperse his strength; and to coordinate all these activities with those of the regular armies on distant battle fronts.

To accomplish these goals, Mao demanded tactics based on surprise and deception:

> In guerrilla warfare, select the tactic of seeming to come from the east and attacking from the west; avoid the solid, at-

tack the hollow; attack, withdraw; deliver a lightning blow, seek a lightning decision.

As opposed to orthodox warfare, which is frequently static, Mao wanted

> constant activity and movement. There is in guerrilla warfare no such thing as a decisive battle; there is nothing comparable to the fixed, passive defense that characterizes orthodox war. In guerrilla warfare, the transformation of a moving situation into a positional defensive situation never arises. The general features of reconnaissance, partial deployment, general deployment, and development of the attack that are usual in mobile warfare are not common in guerrilla war.

Instead of fixed defense, Mao calls for

> alert shifting . . . when the enemy feels the danger of guerrillas, he will generally send troops out to attack them. The guerrillas must consider the situation and decide at what time and at what place they wish to fight. If they find that they cannot fight, they must immediately shift.

Although the guerrilla will defend his own operational bases, these must be abandoned when necessary.

> We must observe the principle, "To gain territory is no cause for joy, and to lose territory is no cause for sorrow."

Such tactics demand

> careful planning . . . those who fight without method do not understand the nature of guerrilla action. A plan is necessary regardless of the size of the unit involved; a prudent plan is as necessary in the case of the squad as in the case of the regiment.

As Lawrence put it, "Guerrilla war is far more intellectual than a bayonet charge."

Good planning depends on superior intelligence, of course, and this can be gained only from the people, who, in turn, must withhold such from the enemy:

> Many people think it impossible for guerrillas to exist for long in the enemy's rear. Such a belief reveals lack of comprehension of the relationship that should exist between the people and the troops. The former may be likened to water and the latter to the fish who inhabit it. How may it be said that these two cannot exist together? It is only undisciplined troops who make the people their enemies and who, like the fish out of its native element, cannot live.

Mao's basic strength came from the people—from the water that produced, then supported, the fish. From the Yen-an haven, his agents went forth to select suitable base areas for organization and consolidation, a process in which volunteers were trained and indoctrinated as agitators and propagandists, who in turn went forth to the countryside to enlist peasant support.

In Mao's scheme of things, this phase merges into one of limited direct action, mainly sabotage and terrorism designed to eliminate members of the opposition, and to gain arms and supplies for the embryo guerrilla force. This expansion phase may last for years, but if it succeeds it merges into a decisive phase: the destruction of the enemy largely with orthodox military forces.

Mao's teachings, though perhaps only partially utilized, nonetheless underlie most of the revolutionary wars fought since World War II. In fact, his doctrine has become a blueprint for the "national wars of liberation" that China and Cuba have promised to foment and support in Asia, Africa, and Central and South America.

Motivation. Fundamental to the revolutionary process is a cause, which unfortunately is not difficult to find in the underdeveloped countries of the world. The cause may assume several guises: to the world it may be presented as liberating a country from the colonial yoke; to the peasant being converted to Communism it may be freedom from serfdom, from oppressive taxation, or from payment of oppressive rents to absentee landlords.

Whether real or artificial, whether inspired by Communism or by virulent nationalism, the political goal is fundamental in motivating people to action. Mao leaves no doubt as to its importance:

> Without a political goal, guerrilla warfare must fail, as it must if its political objectives do not coincide with the aspirations of the people and their sympathy, cooperation, and assistance cannot be gained.

Popular support. The guerrillas' affiliation with the people is constantly stressed in revolutionary writings. Guerrillas spring from the people, who in turn support their spawn, not only by furnishing their sons to the cause but also, when called upon, by furnishing money, food, shelter, refuge, transport, medical aid, intelligence—support that they must attempt to deny the enemy. Although Lawrence called for no more than

> a friendly population, not actively friendly, but sympathetic to the point of not betraying rebel movements to the enemy,

he also wrote that his rebels

> had won a province when the civilians in it had been taught to die for the ideal of freedom: the presence or absence of the enemy was a secondary matter.

Gen. Georgios Grivas, the non-Communist professional soldier who led the Cypriot rebellion, wrote that a guerrilla war stands no chance of success unless it has "the complete and unreserved support of the majority of the country's inhabitants."

Organization. Protracted revolutionary warfare as defined by Mao demands a complicated organization on both the political and the military levels. Mao recommends a clandestine system of parallel hierarchy beginning with the cadre or cellular party structure at the hamlet–village level and proceeding to the top via district, province, and regional command structures.

The tactical organization of guerrilla units varies according to operational demands. Mao called for a guerrilla squad of nine to 11; Grivas employed sabotage groups of four or five. In the Vietnam fighting, the Viet Minh and, later, the Viet Cong ranged from small squads up to battalion and even regimental strengths.

Arms. The guerrilla by necessity employs a wide variety of weapons, some self-manufactured, some captured, and some supplied from outside sources. In the earlier stages of the war, the weapons are usually primitive. Americans in Vietnam have frequently encountered homemade rifles, hand grenades, and Claymore mines; trails booby-trapped with *punji* stakes soaked in urine; and shallow pits lined with nail boards. Nearly every guerrilla war has produced ingenious improvisation, both from necessity and to avoid a cumbersome logistic "tail." Nothing can be simpler to construct and use than a Molotov cocktail or a *plastique* bomb, yet under certain conditions nothing can be more effective.

Rural- and urban-based guerrilla campaigns

Terrain. It is axiomatic to Mao and his followers that revolution begins in the country. Once sufficient base and guerrilla areas are established, it is possible to extend operations to include cities and lines of communication susceptible to attack. This rural strategy is influenced by such factors as the political goal, geography, the insurgent strength, and the government's strength.

Such was the combination of these in Russia that the 1917 revolution was decided in the cities and only later successfully defended in the country by orthodox Communist armies employing guerrilla forces in a complementary role. The Irish Rebellion was also fought largely in the cities, and General Grivas opened the semisuccessful Cypriot rebellion with a few combat groups especially trained in terrorist–sabotage tactics. As his strength grew, he resorted to guerrilla warfare across the entire island. On the basis of this and other examples, Grivas later argued that, contrary to Mao's teachings, guerrilla warfare need not be rural based and, further, that it is sometimes possible for guerrilla warfare alone to accomplish the political objective.

Terror. One of the most hideous characteristics of guerrilla warfare is the use of terror: assassination, a hand grenade thrown into a crowd, an indiscriminate bombing—actions familiar to any insurgency.

Terror is used for several reasons: to focus world attention on the rebel cause, to eliminate opposition leaders, to paralyze normal government activity, to intimidate the general populace, and to keep one's own guerrillas from defecting. It is difficult to assess the psychological impact of terrorist tactics on the general population. It would seem that even those sympathetic to the guerrilla cause may be alienated by the indiscriminate use of terrorism. The defending forces, moreover, may reply in kind, so that the population is subject to terror from both sides.

The question of political sympathy or loyalty may then become irrelevant, the local populace being ready to cooperate with whichever side is in control at a particular moment.

Sanctuary. Guerrilla forces cannot fight all the time. They must control safe areas to which they can retire, voluntarily or involuntarily, for rest, recuperation, and repair of arms, clothing, and equipment and where recruits can be indoctrinated, trained, and equipped. Such areas traditionally are located in remote, rugged terrain, usually mountains, forests, and jungles; but guerrilla areas may be developed in which whole villages and hamlets serve a sanctuary role. The sea can also provide sanctuary, as in the Peninsular War, when the British navy succored Wellington's cooperating Portuguese–Spanish guerrillas.

Sanctuary may also be provided by sympathetic neighbouring countries: during the Greek Civil War (1946–49) the Communist guerrillas frequently retreated into Yugoslavia, which offered not only physical sanctuary but also arms and supplies; and it was only after the Yugoslavs closed the border that the guerrillas were finally defeated, though not for that reason alone. Similarly, Ho Chi Minh's guerrillas, in the later stages of the war against France, relied on China for refuge, training, and supply of arms and equipment.

The people offer a final form of sanctuary. At one time during the Cypriot revolt, Grivas was surrounded by a British force for nearly two months but, though spotted, was repeatedly able to escape capture or death. An Algerian rebel leader was able to install himself within 200 yards of the army commandant's headquarters in Algiers, avoiding capture for several months. In South Vietnam in 1964 American officials discovered that several thousand supposedly government-controlled, "fortified" hamlets were in fact controlled by Viet Cong guerrillas "who often used them for supply and rest havens" (Staff of the Senate Republican Policy Committee, *The War in Vietnam*, 1967).

OTHER ASPECTS OF GUERRILLA WARFARE

Guerrilla forces. *Leaders.* The unusual requirements of guerrilla warfare call for outstanding leadership at all levels if a guerrilla force is to survive and prosper. The vicissitudes of these wars demand a leader not only endowed with extraordinary intelligence and courage but also buttressed by an almost fanatic belief in himself and his cause. Lenin, Trotsky, Lawrence, Mao, Tito, the Filipino Luis Taruc, the Kenyan Jomo Kenyatta, Ho Chi Minh, Vo Nguyen Giap, the Algerian Ahmed Ben-Bella, Castro, Ernesto "Che" Guevara—these and dozens of their lieutenants at lower levels have all been unusual, unorthodox personalities, generally with civilian backgrounds. But all were able to attract followers, to organize them, and to instill a disciplined zeal matched only in the most elite military organizations.

Recruits. Selectivity is the key to effective guerrilla recruitment. The guerrilla recruit must be resourceful and enduring and must be committed totally to the cause if he is to withstand the hardships and dangers that guerrilla fighting involves. The strength of a guerrilla movement is thus directly related to the degree to which the objectives of the campaign can evoke a response from youths of military age. Historically, the cause has invariably been a desire for national independence and autonomy, sometimes combined with an ideological commitment to Communism or the overthrow of corrupt and repressive regimes or both. Recruitment to such causes has rarely presented much difficulty to guerrilla leaders, although, in those cases in which the conflict has been large-scale and extended over a long period of time, it has been necessary to abandon selectivity and resort to conscription. Such has been the case in Indochina, where Communist recruiting normally begins at the village level, generally by inducing peasant candidates to join one or more front groups and participate, if only indirectly, in the war effort. After exposure to political indoctrination, the candidate joins the village guerrilla cell, after which he can be promoted to the regional and regular forces.

Use of conscription

Discipline. Since it is essential for the guerrilla to win and retain popular support, he is also taught to practice circumspect behaviour when among the people. Communist leaders in China, Cuba, and Vietnam have drawn up lengthy codes of individual behaviour: the Chinese guerrilla, for example, was required to pay a peasant for food, to respect his property, and not to offend propriety by undressing in front of a peasant woman.

It is questionable to what depth party-imposed discipline descends in the average Communist guerrilla. Unquestionably, the hard-core guerrilla practices an almost ascetic association with the people, while approaching his military tasks with dogged determination. But, judging from interrogations of Viet Cong guerrilla prisoners and from the number of Viet Cong defectors, the guerrilla's discipline varies as ideological beliefs ebb and flow in the physical and moral tides of war.

Counterguerrilla warfare. In waging this type of warfare, which calls for location, isolation, and elimination of the entire guerrilla apparatus (political as well as military), orthodox military commanders have employed and continue to employ a wide range of weapons and tactics that, judged by results, are more appropriate to a conventional-warfare situation. Wholesale bombings and mass artillery interdictions of suspected sanctuary areas, division- and corps-strength "sweeping" operations in which only a few guerrillas are captured or killed but whole villages are destroyed, the establishment of defended but isolated chains of military outposts, mass arrests and interrogations—each has failed to achieve triumphs.

Throughout history, nations have impelled insurgencies because of political errors, and military commanders have failed to quell insurgencies because of political ignorance. In Communist revolutionary warfare, especially, the political factor is paramount. If the Communist guerrilla is not supported by the people—if he is not a fish in the sea of humanity—he cannot effectively operate for any length of time. Consequently, the government must win the people's support, both to deprive the guerrilla of this support and to obtain information on which to base tactics of destruction. It is not enough to break up guerrilla bands and kill individual guerrillas. A government can claim victory only when the subversive organization behind each level of an insurgency has been destroyed and when viable government has been achieved.

Mixed strategy of counterguerrilla warfare

Such counterinsurgency campaigns as those conducted in the Philippines and in Malaya prove that the Communist guerrilla ultimately can be defeated (but not necessarily eliminated). The means of defeat, however, lie only in a patient and judicious application of a host of civil and military measures, involving social and economic reform, effective policing, and military security. Although in the course of a counterinsurgency campaign one or another of these factors may assume a temporary supremacy, in the end each must remain integral to an overall political consideration.

Indeed, political realism is the first essential in conducting a counterinsurgency: the recognition of weakness as well as strength, of failure as well as success. An insurgency indicates a breakdown of government in that a minority is able to defy law and order while coercing others to offer either active or passive support to the cause. The opening phase of an insurgency, in which subversion is supported by selective terrorism, must be met by specific governmental measures, both covert and overt. Because this phase is generally covert, the government, usually the police, must practice considerable subtlety in a difficult environment in which internal weakness, ineptness, or corruption is at work. If the insurgency is not contained at this level it usually flares into the second, "armed struggle" phase, which brings guerrilla warfare into the open.

In the opening phase of an insurgency, a government is usually on the defensive. In Malaya this was a period in which the government was able to prevent the enemy from taking over and to keep the insurgency from escalating. The general strategy was security, both by maintaining police functions and, militarily, by splitting up (but not attempting to destroy) the larger guerrilla units. These

holding operations gave the government time to marshal its forces (not just police and military) in order to fight the second, or offensive, phase, in which the enemy's power was broken, and, finally, the third, or victory, phase, which destroyed the last remnants of guerrilla forces while establishing a stable, independent government.

One veteran of the Malayan campaign ascribes the government's success to certain basic principles, which he holds essential for the conduct of any counterinsurgency: (1) the government must have a clear political aim, ideally "to establish and maintain a free, independent and united country which is politically and economically stable and viable"; (2) in ferreting out, neutralizing, and destroying guerrillas, the government, no matter how tempted, must function in accordance with law and with a carefully developed counterinsurgency plan, which grants priority to defeating the political subversion, not the guerrillas per se; and, (3) finally, in fighting the offensive phase the government must first develop base areas before commencing aggressive tactics (Robert Thompson, *Defeating Communist Insurgency*, 1966).

These principles need not unduly restrict the counterguerrilla efforts. Most authorities agree that emergency regulations, often harsh, must be legally invoked. In Malaya these included compulsory census, an enforceable identity-card system, suspension of habeas corpus (but with carefully publicized safeguards), permission to search private property without a warrant, the death sentence for persons caught with unauthorized weapons, harsh sentences for those aiding the Communists, flexible power of curfew; later extraordinary measures included the right to shoot on sight in prohibited areas, the right to resettle whole villages, and the right to control food distribution with harsh penalties, including death, for those found guilty of aiding the enemy.

Such regulations, or course, are not attractive. If indiscriminately applied, as, unfortunately, is sometimes the case, they will lead to an increasing alienation of the civil population from the government. When properly applied, however, they will greatly aid the police and military forces in their essential mission of providing security to the civil populace, which, in turn, will then feel more free to provide essential information on which to base further counterguerrilla operations.

Holding and clearing operations

The exact nature of such operations must vary in accordance with the enemy's strength and the area concerned. The first priority of government is to re-establish law and order, which, in a rural area, means revitalizing the rural police function. The military effort, the strength of which is dictated by necessity, concentrates on clearing operations, designed to break up and disperse large guerrilla formations, then to keep them deprived of the initiative by small-unit tactics, mainly patrols and ambushes based on valid intelligence. The clearing operation is followed by the holding operation, which is designed to "restore government authority . . . and to establish a firm security framework" (Thompson, *op. cit.*). The holding operation is the period of "winning the hearts and minds" of the people, first by providing security, which will be maintained by strategic hamlets defended by organized hamlet militias working in conjunction with government forces where necessary, and second by providing social reforms (land reform, schools, hospitals, community projects) that will identify the government with the people's best interests. Once won over, the people will deprive the guerrillas of vital support, besides furnishing information necessary for police and military forces to penetrate and destroy the local insurgent organization.

The clearing and holding operations provide the key to successful counterguerrilla warfare. When they have been carefully applied, as in Malaya, they have proved successful; when applied quantitatively, as in Vietnam, they have in large part failed. Even under the most favourable circumstances, most governments lack the necessary civil, police, and military resources to carry out clear and hold operations simultaneously in all areas. For this reason, the military effort may have to extend to secondary operations in lower priority areas. These are designed to keep the guerrilla off balance until the civil effort can be enlarged. Such operations may include large-scale sweep and clear actions, in which large numbers of aircraft and helicopters are used; they may include long-term, deep-penetration operations, in which units are parachuted into guerrilla sanctuaries and supplied by airdrop while establishing and maintaining permanent ambushes, sometimes with the aid of friendly tribes.

Such may be the strength of the insurgency and the weakness of the legal government that outside aid is called for, as happened in Vietnam. Unless limited to supply, technical training, and professional advice, outside aid may well prove a two-edged sword. If the donor government underestimates the dimensions of the conflict, as happened with the United States in the case of Vietnam, it is persuaded to a military intervention that, by escalating to the extent needed for victory, tends to take over the war from the host government, thus widening the gulf between the people and their government and providing the guerrillas with propaganda for the charge of imperialist aggression.

Legal status. For understandable reasons, the orthodox military commander has always placed the guerrilla in an extralegal status. After the British were stung several times by Francis Marion in the Revolutionary War, they complained that he fought neither "like a gentleman" nor like "a Christian." Napoleon's marshals on the Spanish peninsula were driven to violent reprisals against Spanish–Portuguese guerrillas, with the result that for every guerrilla shot a French prisoner paid with his life, a "barbarous system" finally concluded by mutual agreement. The problem arose again in the American Civil War; when Gen. Eleazar A. Paine, the Union commander in western Kentucky, was unduly harassed by guerrillas, he published this proclamation:

> I shall shoot every guerrilla taken in my district, and if your Southern brethren retaliate by shooting a Federal soldier, I will walk out five of your rich bankers, and cotton men, and make you kneel down and shoot them. I will do it, so help me God" (Richard Bennett, *The Black and Tans*, 1959).

Rules applicable to guerrillas

The Brussels international conference of 1874 provided that, in order to be recognized as lawful belligerents, guerrillas must answer to a specific commander, wear a distinctive badge, carry arms openly, and conform in operations to the laws and customs of war. The Hague conferences on the rules of land warfare in 1899 and 1907 adopted this definition with a few modifications, and it is also contained in the Geneva Conventions (1949) on the laws of war.

The Hague ruling has not been complied with, mainly because conformance would nullify the advantages of guerrilla warfare but also because sabotage and terrorist tactics often breed brutal reprisals. The guerrilla has invariably been held fair game for torture or for execution without trial. As in the Peninsular War of the early 19th century, the guerrilla, unable to expect just treatment, has continued to render unjust treatment, a cycle of horror reaching its zenith in the Spanish Civil War, in the partisan actions of World War II, and in Vietnam.

BIBLIOGRAPHY. No definitive history of guerrilla warfare exists. Nearly all ancient chroniclers offer examples of guerrilla warfare, if only in passing, as do most classical and medieval historians. In addition to the writings of Theodor Mommsen, Charles W. Oman, J.B. Bury, A.R. Burn, and John Beeler, see REGINALD HARGREAVES, *Beyond the Rubicon: A History of Early Rome* (1966); FERDINAND SCHEVILL, *History of the Balkan Peninsula, from the Earliest Times to the Present Day* (1922, reprinted 1966); J.E. MORRIS, *The Welsh Wars of Edward I* (1901, reprinted 1969); R.C. SMAIL, *Crusading Warfare, 1097–1193* (1956); and EDOUARD PERROY, *La Guerre de Cent Ans* (1945; Eng. trans., *The Hundred Years War*, 1951). Also helpful are general military histories such as JOHN FORTESCUE, *A History of the British Army*, 14 vol. (1899–1930; 6 vol. of maps and plans, 1906–30); J.F.C. FULLER, *The Decisive Battles of the Western World*, 3 vol. (1954–56); R.A. PRESTON and SYDNEY F. WISE, *Men in Arms*, 2nd rev. ed. (1970); CORRELLI BARNETT, *Britain and Her Army, 1509–1970* (1970); RUSSELL WEIGLEY, *History of the United States Army* (1967); and R.D. HEINL, *Soldiers of the Sea* (1962). Lesser known but invaluable works include C.E. CALLWELL, *Small Wars*, 3rd ed. (1906); and CHARLES GWYNN, *Imperial Policing*, 2nd ed.

(1936). For the American Revolution, see especially CHRISTOPHER WARD, *The War of the Revolution*, 2 vol. (1952); and JOHN ALDEN, *The South in the Revolution, 1763–1789* (1957). Studies of more recent guerrilla wars number in the hundreds, and only a few can be mentioned. General studies include M.R.D. FOOT, *SOE in France* (1966); OTTO HEILBRUNN, *Warfare in the Enemy's Rear* (1963); DAVID GALULA, *Counterinsurgency Warfare* (1964); CHARLES THAYER, *Guerrilla* (1963); and GEORGE TANHAM, *Communist Revolutionary Warfare: The Vietminh in Indochina*, rev. ed. (1967). The following studies of specific actions are recommended: LEON WOLFF, *Little Brown Brother: How the United States Purchased and Pacified the Philippine Islands at the Century's Turn* (1961); RONALD ATKIN, *Revolution! Mexico, 1910–1920* (1969); T.E. LAWRENCE, *Seven Pillars of Wisdom* (1935, reprinted 1966); EDGAR HOLT, *Protest in Arms: The Irish Troubles, 1916–1923* (1960); LEONARD MOSLEY, *Duel for Kilimanjaro* (1963); DAVID WOOLMAN, *Rebels in the Rif* (1968); FITZROY MacLEAN, *Disputed Barricade: The Life and Times of Josip Broz-Tito, Marshall of Jugoslavia* (1957); JOHN ARMSTRONG (ed.), *Soviet Partisans in World War II* (1964); MAO TSE-TUNG, *On Guerrilla Warfare* (1961); N.D. VALERIANO and C.T.R. BOHANNAN, *Counter-Guerrilla Operations* (1962); EDGAR O'BALLANCE, *The Greek Civil War, 1944–1949* (1966); MENAHEM BEGIN, *The Revolt: Story of the Irgun* (1951); RICHARD CLUTTERBUCK, *The Long, Long War* (1966); ROBERT THOMPSON, *Defeating Communist Insurgency* (1966) and *No Exit from Vietnam* (1970); PETER PARET, *French Revolutionary Warfare from Indochina to Algeria* (1964); GEORGE GRIVAS, *On Guerrilla Warfare* (1964; orig. pub. in Greek, 1961); BERNARD B. FALL, *The Two Viet-Nams*, 2nd rev. ed. (1967); VO-NGUYEN-GIAP, *People's War, People's Army* (1962); DOUGLAS PIKE, *Viet Cong* (1966); and CHE GUEVARA, *Guerrilla Warfare* (1961).

(R.B.A.)

Guevara, "Che"

A prominent figure in the Cuban Revolution of the 1950s and '60s and an important official in Fidel Castro's first Communist government there, Ernesto "Che" Guevara exerted an important influence on the theories and tactics of guerrilla warfare.

Lee Lockwood—Black Star



Guevara.

Ernesto Guevara de la Serna was born in Rosario, Argentina, on June 14, 1928. The eldest of five children in a middle class family of Spanish-Irish descent and leftist leanings, Ernesto was known for his dynamic and radical views even as a boy. Though suffering from the crippling bouts of asthma that were to handicap him throughout his life, he excelled as an athlete and a scholar, completing his medical studies in record time in March 1953. He spent many of his holidays travelling in various countries of Latin America, and his first-hand observations of their economic and political problems—specifically of the great poverty of its masses—convinced Guevara that the only solution lay in violent revolution. His travels also taught him to look upon Latin America not as a collection of separate nations but as a cultural and economic entity, the liberation of which would require an intercontinental strategy.

In 1953 Guevara went to Guatemala, where Colonel Jacobo Arbenz Guzmán headed a progressive régime that, through various reforms, especially land reform, was attempting to bring about a social revolution. Around this time Guevara acquired his famous nickname, "Che," from a verbal mannerism of Argentines who punctuate their speech with the interjection *che*.

The overthrow of the Arbenz régime by a United States Central Intelligence Agency-backed coup in 1954 persuaded Che that the United States would always oppose progressive leftist governments in Latin America and in the other developing countries of the world. This conviction, which he later expressed in many of his speeches and writings, became the cornerstone of his plans to bring about Socialism through a world revolution. Che left Guatemala for Mexico, where he met the Cuban brothers, Fidel and Raúl Castro, political exiles who were preparing to return to Cuba with an expeditionary force in an attempt to overthrow the dictatorship of Fulgencio Batista. Che joined Castro's force and began to train for guerrilla warfare. On November 25, 1956, Che and 82 other young men left Mexico on the yacht "Granma" and landed in the Cuban province of Oriente a few days later. Immediately detected by Batista's army, they were almost wiped out in the first encounter; the few survivors, including the wounded Che, were able to reach the mountain chain known as the Sierra Maestra, where they set up the nucleus of a guerrilla army. The rebels slowly gained in strength, seizing weapons from the enemy, winning support and new recruits from the local peasantry and the intellectuals and workers in the cities. Che exhibited great courage and skill in combat and soon became one of Castro's ablest and most trusted aides. In July 1957 he was made a major and appointed leader of the rebel army's second column. Che recorded the two years spent in overthrowing Batista's government in a detailed account entitled *Pasajes de la guerra revolucionaria* (English trans., *Reminiscences of the Cuban Revolutionary War*, 1968), first published in 1963.

Association with Castro

After Castro's victorious troops entered Havana on January 2, 1959, and established a progressive government along Marxist lines, Che became a Cuban citizen, divorced his Peruvian wife, Hilda Gadea, by whom he had one daughter, and married a member of Fidel's army, Aleida March, by whom he had four children. He became as prominent in the new government as he had been in the revolutionary army, representing Cuba on many commercial missions and delegations to African, Asian, and Socialist countries. He also became well-known in the West for his outspoken opposition to all forms of imperialism and neocolonialism and for his fiery attacks on U.S. foreign policy in Africa, Asia, and especially Latin America. An active participant in the economic and social reforms brought about by Castro's government, he occupied such important posts in his adopted country as chief of the Industrial Department of the National Institute of Agrarian Reform, president of the National Bank of Cuba, and minister of industry. During this period, he defined Cuba's policies and his own views in many speeches, articles, letters, and essays, the most important of which are "El socialismo y el hombre en Cuba" (1965; "Man and Socialism in Cuba," 1967), which is an examination of Cuba's new brand of Communism, and a highly influential manual on guerrilla strategy and tactics (Eng. trans., *Guerrilla Warfare*, 1961).

Disappearance from public life

After April 1965 Che dropped out of public life and then vanished altogether. His disappearance was variously attributed to the relative failure of the industrialization scheme he had advocated while minister of industry, to pressures exerted on Castro by Soviet officials disapproving of Che's pro-Chinese Communist outlook, and to serious differences between Che and the Cuban leadership regarding Cuba's economic development and ideological line. In October of that year, Castro revealed a letter written to him by Che some months earlier in which Che reaffirmed his enduring solidarity with the

Cuban Revolution but explained that "other nations are calling for the help of my modest efforts" and that, having "always identified with the world outcome of our Revolution," he had decided to go and fight as a guerrilla in other liberation struggles being waged in different parts of the world.

Che's movements and whereabouts remained a secret for the next two years, although he sent a letter that was read aloud at the first conference of the Tricontinental Solidarity Organization, held in Havana in April 1967. In this letter, known as the "Message to the Tricontinental," Che analyzed the world situation and called for all-out war against imperialism and the forces of reaction. During this period, Che may have gone to North Vietnam and to several countries in Latin America; it is also probable that he secretly returned to Cuba in 1966, but the facts have never been established positively. It was later learned that he had spent some time in the Congo with other Cuban guerrilla fighters, helping to organize the Patrice Lumumba Battalion, which fought in the civil war there.

In the autumn of 1966, Che went to Bolivia, incognito, to create and lead a guerrilla group in the region of Santa Cruz. He spent the next 11 months training the Bolivian guerrillas and fighting with them, meticulously recording their progress and setbacks in a diary that was found and published after his death. On October 8, 1967, the group was encircled by a special detachment of the Bolivian Army and almost annihilated. Che was captured after being wounded and shot soon afterward.

Capture and execution

While pictures of the dead guerrilla chief were being circulated and the circumstances of his death were being debated, Che's legend began to spread. Demonstrations in protest against his death occurred in many parts of the world, and articles, tributes, and poems were written about his adventurous life and tragic death. Even those liberal elements that felt little sympathy with Che's militantly Communist ideals during his lifetime expressed regret for his death and admiration for his integrity and spirit of self-sacrifice. He is singled out from other revolutionaries by many young people in the West because he rejected a comfortable middle class background to fight for those who were deprived of political power and economic stability. An intellectual and a thinker, Che believed in putting his theories into action. Called "the most complete human being of our age" by the French philosopher Jean-Paul Sartre, Che's supporters believe he may yet prove to be the most important thinker and activist in Latin America since Simón Bolívar, leader of the South American independence movement.

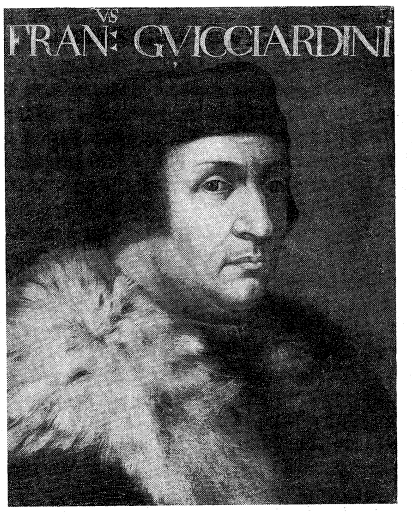
BIBLIOGRAPHY. ANDREW SINCLAIR, *Che Guevara* (1970), a biography of Che and a critical study of his works; RICARDO ROJO, *Mi amigo el Ché* (Eng. trans., *My Friend Ché*, 1968), an account of Che's life by a childhood friend; M. ALEXANDRE (ed.), *Viva Che* (1968), an anthology of tributes, articles, and poems about Che by many authors, including Fidel Castro; *On Trial: Fidel Castro, Régis Debray* (1968), contains Debray's self-defence at his trial in Bolivia for his role in Guevara's guerrilla group; J. GERASSI (ed.), *Venceremos: The Speeches and Writings of Ernesto Che Guevara* (1970).

(A.A.Si.)

Guicciardini, Francesco

The *Storia d'Italia* ("History of Italy") of the Italian historian and statesman Francesco Guicciardini has remained one of the monuments of Italian historiography. Written by a statesman closely associated with many of the events he described, and by a historian who in his critical use of evidence followed and surpassed his humanist predecessors, the work is the most important contemporary history of Italy during the Italian wars.

Francesco Guicciardini was born at Florence on March 6, 1483. He came of an aristocratic family that played a prominent role under Lorenzo de' Medici ("The Magnificent"). From 1498 to 1505 Guicciardini studied civil law at Florence, Ferrara, and Padua and subsequently set up legal practice at Florence. In 1508 he married Maria, daughter of Alamanno Salviati. In the same year he began to write his family memoirs and his *Storie fiorentine* ("History of Florence") from 1378 to 1509. The latter constitutes one of the major sources for the history of the republican regime after 1494; it reveals Guicciardini's gifts for historical analysis and narrative. Elected in 1511 as Florentine ambassador to King Ferdinand of Aragon, he was at the Spanish court when in 1512 the Florentines restored the Medici, exiled in 1494, under the pressure of Spanish troops. On his return to Florence in 1514, he resumed his legal practice; in 1514 he was member of the Otto di Balìa, who were in charge of internal security, and in 1515 of the Signoria, the highest magistracy. In 1513 Cardinal Giovanni de' Medici became Pope Leo X; in 1516 he appointed Guicciardini governor of Modena, and in 1517 also of Reggio. Until 1534 Guicciardini served the papacy almost continuously.



Guicciardini, oil painting by an unknown artist. In the Galleria degli Uffizi, Florence.
By courtesy of the Galleria degli Uffizi, Florence

Appointed governor of papal states

As governor of an exposed and recently acquired part of the papal states, in which he had to face internal disorders as well as external dangers, Guicciardini showed outstanding administrative gifts. His severe and sometimes ruthless measures were effective in restoring order but also caused him unpopularity. The outbreak of the war in northern Italy between King Francis I of Spain and the Holy Roman emperor Charles V, with whom Leo had concluded an alliance, turned Reggio into a military outpost of the papal states, and in July 1521 Guicciardini was appointed commissioner general of the papal army. During this time, Guicciardini became also a prolific political writer, composing numerous memoranda and treatises, mostly in the form of discourses on political problems of the day, often in connection with his official duties. A number of them deal with the government of Florence, on which he also wrote, between 1521 and 1525, the *Dialogo del reggimento di Firenze* ("Dialogue on the government of Florence"). In this he advocates an aristocratic regime on the Venetian model as the ideal constitution for his city. In his capacity as commissioner general, he prevented, by his courage and determination, Parma from falling into French hands in December 1521. But the death of Leo X in the same month jeopardized his career temporarily; after the election of Adrian VI, he was at first deprived of the governorships of Modena and Reggio but recovered them at the end of 1522. In 1523, after Adrian's death, he had to defend both cities against their original ruler, the Duke of Ferrara. Reggio capitulated, but Modena was held by Guicciardini against superior odds. After the election of Cardinal Giulio de' Medici as Clement VII, he earned his reward by being appointed, in 1524, president of the Romagna, the northernmost papal province. In the critical situation after the battle of Pavia, when the army of the emperor Charles V was preparing to advance south, Guicciardini conveyed to the Pope much advice, and in January 1526 he was summoned to Rome. There he

played a prominent role in the papal counsels, advocating an alliance with France against Charles V. The resulting League of Cognac, concluded in May 1526, was to no small extent his work, and in June he was appointed papal lieutenant general with the army of the league. The failure of the league to prevent the imperial army under the duke of Bourbon from advancing on Florence and Rome involved him once more in the fate of his native city.

The danger in which Florence found itself as a result of Clement's policy had increased the opposition to the Medici regime. When, on the arrival of the Duke of Urbino with his army near Florence, the Medici left the city to welcome him (April 26, 1527), a revolt broke out. Guicciardini, who had arrived shortly before to help protect the city, succeeded in preventing the Duke from assaulting the palace of the Signoria by negotiating a free pardon in return for surrender. A few days later, Bourbon's army captured Rome, and this was followed by the expulsion of the Medici from Florence and the restoration of republican government in that city.

The collapse of Pope Clement's authority in Rome rendered Guicciardini's position as his lieutenant general untenable, while his long association with the Medici made him suspect in republican Florence. The victory of the intransigent republican faction and the fall of the gonfalonier Niccolò Capponi, who had been trying to come to terms with the Pope (April 1529), followed by the advance of the imperial army on the city, endangered Guicciardini's position and, in September 1529, he left Florentine territory for the papal court. Thereafter he fully supported Clement's bid for a Medicean restoration in Florence, while seeking to obtain favourable conditions for the Florentines. In March 1530 he was condemned as a rebel at Florence. Between 1528 and 1530 Guicciardini worked on his second history of Florence and compiled the most concise and varied expression of his views on society and politics in his collection of maxims and observations, the *Ricordi*. His political thought is frequently akin to, and sometimes more radical than, that of his friend Niccolò Machiavelli, with whom he shares, despite his long service with the papacy, a criticism of the contemporary church. He disagreed, however, in his *Considerazioni intorno ai "Discorsi" del Machiavelli* ("Considerations on the 'Discourses' of Machiavelli," *c.* 1530), with Machiavelli's interpretation of Roman history as evidence for a political science. After the city's surrender, he returned as papal representative and took a leading part in the persecution of the republicans. In 1531 Clement appointed him governor of Bologna, but he lost this post after the accession of Paul III in 1534. Back in Florence, he acted as legal adviser to Duke Alessandro de' Medici and began work on a history of Italy during his lieutenantship that, redrafted during the following years, became the nucleus of his far more ambitious *Storia d'Italia* (*History of Italy*) from 1494 to 1534. He began the work probably in 1536; the final revision was not completed when he died.

His history of Italy

After the murder of Alessandro in 1537, he helped secure the succession for Cosimo, probably hoping to limit the ducal powers which he considered excessive. Disappointed in his hopes and personal ambitions, although still holding high office under the new ruler, he devoted the last years of his life, in his villa at S. Margherita a Montici, to the composition of his *Storia d'Italia*, the crowning achievement of his life. He died on May 22, 1540.

BIBLIOGRAPHY. *La Storia d'Italia* was edited by ALESSANDRO GHERARDI, 4 vol. (1919); and by CONSTANTINO PANIGADA in *Scrittori d'Italia*, 5 vol. (1929); most of Guicciardini's other works were published in 10 vol. between 1857 and 1867 by G. CANESTRINI in *Opere inedite*. New editions of the *Storie fiorentine*, the *Dialogo del reggimento di Firenze* and other political treatises, the *Ricordi*, the family memoirs and diaries, by ROBERTO PALMAROCCHI, are included in *Scrittori d'Italia* (1931–36). The fundamental critical edition of the *Ricordi* is by RAFFAELE SPONGANO (1951); the draft of the second history of Florence was edited for the first time by ROBERTO RIDOLFI under the title *Cose fiorentine* (1945). There are English translations of the *Storia d'Italia* by SIDNEY ALEXANDER (1968, abridged); of the *Storie fiorentine* (1970) and of the *Ricordi* (*Maxims and Reflections of a Renaissance Statesman*, 1965, introd. by NICOLAI RUBINSTEIN), both by MARIO DOMANDI; of the *Ricordi*, the *Considerazioni intorno ai "Discorsi" del Machiavelli*, and the diaries by MARGARET GRAYSON (*Selected Writings*, 1965, introd. by CECIL GRAYSON). The new edition of Guicciardini's letters (*Carteggi*), in the *Fonti per la storia d'Italia*, is almost completed (1938–70): vol. 1–4 (1499–1521) ed. by ROBERTO PALMAROCCHI; vol. 5–16 (1522–40) by P.G. RICCI. The definitive biography is ROBERTO RIDOLFI, *Vita di Francesco Guicciardini* (1960; Eng. trans., 1967); see also ANDRE OTETEA, *François Guichardin, sa vie publique et sa pensée politique* (1926). On Guicciardini as historian and political thinker, see FELIX GILBERT, *Machiavelli and Guicciardini: Politics and History in Sixteenth-Century Florence* (1965).

(N.Ru.)

Guinea

The Republic of Guinea (République du Guinée) is an independent nation of West Africa. It is bordered by the Atlantic Ocean to the west; by Portuguese Guinea, Senegal, and Mali to the north and east; by the Ivory Coast to the southeast; and by Liberia and Sierra Leone to the south. Its area of 94,925 square miles (245,857 square kilometres) supports a largely rural population of more than 4,000,000. The national capital of Conakry is the country's main port and its only city.

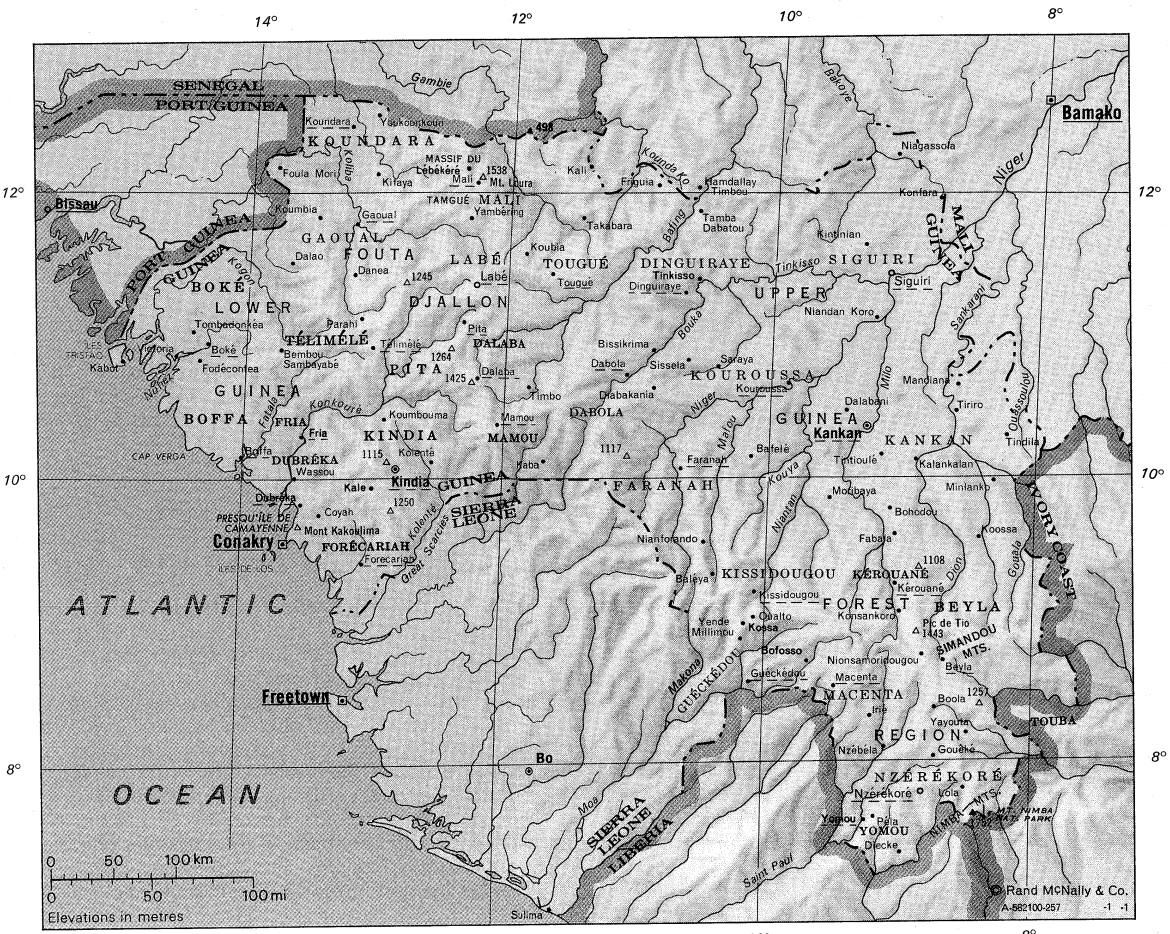
In 1958 Guinea became the second sub-Saharan nation after Ghana and the third French African possession after Tunisia and Morocco to achieve independence. Beset with difficulties of adjustment and reorganization, the nation successfully avoided being overwhelmed by internal discord or external pressures under the leadership of its president Ahmed Sékou Touré and his Parti Démocratique de Guinée. Since the 1960s Guinea has continued to occupy a special position among African states in its unqualified rejection of colonial control or economic domination by more developed nations. A militant pan-Africanist stance on the continent, "positive neutralism" in the Cold War, and a unique articulation of African Socialism and cultural revolution in internal affairs contribute to Guinea's image as one of the most radical experiments in social and political development in West Africa. (For history, see WEST AFRICA, HISTORY OF. For related physical features, see GAMBIA RIVER, NIGER RIVER, and SENEGAL RIVER.)

The landscape. *Relief.* There are four geographic regions; Lower Guinea, the Fouta Djallon, Upper Guinea, and the Forest Region. Lower Guinea includes the coast and coastal plain. The irregular coast is marked by drowned river valleys that form inlets and tidal estuaries and by offshore islands. There are few lagoons or sandbars, but the estuaries are muddy and lined with mangrove swamps.

Immediately inland the gently rolling coastal plain rises to the east, being broken by rocky spurs of the Fouta Djallon highlands in the north at Cape Verga and in the south at the Kaloum Peninsula. Between 30 and 50 miles wide, the plain is wider in the south than the north. Its base rocks of granite and gneiss (coarse-grained rock containing bands of minerals) are covered with laterite (red soil with a high content of iron oxides and aluminum hydroxide) and sandstone gravel.

The Fouta Djallon highlands rise sharply from the coastal plain in a series of abrupt faults. More than 5,000 square miles of the highlands' total extent of 30,000 square miles lie above 3,000 feet. The enormous sandstone block is comprised of level plateaus broken by deeply incised valleys and dotted with dykes or peaks of igneous rock, such as the Kakoulima Massif, which attains 3,273 feet (998 metres) northeast of Conakry. The highest point in the highlands, Mt. Loura, rises to 4,970 feet (1,515 metres) near the town of Mali in the north.

Upper Guinea is comprised of the Niger Plains, which slope northeastward toward the Sahara. The flat relief is broken by rounded granite hills and outliers of the Fouta Djallon. Composed of granite, gneiss, schist (crystalline rock), and quartzite, the region has an average elevation of about 1,000 feet.



GUINEA

MAP INDEX

Political subdivisions

Beyla..... 8·55n 8·25w
Boffa.....10·20n 14·00w
Boké.....11·00n 14·20w
Conakry..... 9·31n 13·43w
Dabola.....10·36n 11·07w
Dalaba.....10·45n 12·18w
Dinguiraye.....11·30n 10·55w
Dubréka.....10·00n 13·40w
Faranah.....10·00n 10·50w
Forécariah..... 9·30n 13·15e
Fria.....10·30n 13·40w
Gaoual.....11·45n 13·12w
Guéckédou..... 8·40n 10·15w
Kankan.....10·10n 9·15w
Kérouané..... 9·10n 8·50w
Kindia.....10·00n 12·45w
Kissidougou..... 9·15n 9·55w
Koundara.....12·25n 13·10w
Kouroussa.....10·40n 9·55w
Labé.....11·23n 12·07w
Macenta..... 8·30n 9·25w
Mali.....12·05n 12·05w
Mamou.....10·30n 12·00w
Nzérékoré..... 7·50n 8·45w
Pita.....11·00n 12·45w
Siguiri.....11·30n 9·15w
Télimélé.....11·00n 13·30w
Tougué.....11·28n 11·36w
Yomou..... 7·35n 9·10w

The name of a political subdivision if not shown on the map is the same as that of its capital city.

Cities and towns

Bafelé.....10·09n 10·08w
Baléya..... 9·15n 10·29w
Bembou Sambayabé.....10·55n 13·44w
Beyla..... 8·41n 8·37w
Bissikrima.....10·51n 10·56w
Boffa.....10·10n 14·02w
Bofosso..... 8·40n 9·42w
Bohodou..... 9·46n 9·04w
Boké.....10·56n 14·18w
Boola..... 8·22n 8·43w
Conakry..... 9·31n 13·43w
Coyah..... 9·43n 13·23w
Dabola.....10·45n 11·07w
Dalaba.....10·42n 12·15w
Dalabani.....10·28n 9·27w
Dalao.....11·29n 13·40w
Danea.....11·27n 13·12w
Diabakania.....10·38n 10·58w
Diecke..... 7·21n 8·58w
Dinguiraye.....11·18n 10·43w
Dubréka..... 9·48n 13·31w
Fabala..... 9·44n 9·05w
Faranah.....10·02n 10·44w
Fodécontea.....10·50n 14·22w
Forécariah..... 9·26n 13·06w
Foula Mori.....12·10n 13·51w
Fria.....10·27n 13·32w
Gaoual.....11·45n 13·12w
Gouéké..... 8·02n 8·43w
Guéckédou..... 8·33n 10·09w
Irié..... 8·17n 9·11w
Kaba.....10·09n 11·40w
Kabot.....10·48n 14·57w
Kalankalan.....10·07n 8·54w
Kale..... 9·55n 13·06w
Kankan.....10·23n 9·18w
Kérouané..... 9·16n 9·01w
Kifaya.....12·10n 13·04w
Kindia.....10·04n 12·51w
Kintinian.....11·36n 9·23w
Kissidougou..... 9·11n 10·06w
Kolenté.....10·06n 12·37w
Konfara.....11·55n 8·50w
Konsankoro..... 9·02n 9·00w
Koossa..... 9·32n 8·32w
Kossa..... 8·57n 10·05w
Koubia.....11·35n 11·54w
Koumbia.....11·48n 13·30w
Koumbouma.....10·24n 12·56w
Koundara.....12·29n 13·18w
Kouroussa.....10·39n 9·53w
Labé.....11·19n 12·17w
Lébékéré.....12·07n 12·24w
Lola..... 7·48n 8·32w
Macenta..... 8·33n 9·28w
Mali.....12·05n 12·18w
Mamou.....10·23n 12·05w
Mandiana.....10·38n 8·41w
Minianko..... 9·58n 8·22w
Moribaya..... 9·53n 9·33w
Niagassola.....12·19n 9·07w
Niandan Koro....11·05n 9·15w
Nianforando..... 9·32n 10·31w
Nionsamoridougou..... 8·43n 8·50w
Nzébéla..... 8·05n 9·06w
Nzérékoré..... 7·45n 8·49w
Oualto..... 9·01n 10·06w
Parahi.....11·09n 13·07w
Péla..... 7·37n 9·07w
Pita.....11·05n 12·24w
Saraya.....10·46n 10·24w
Siguiri.....11·25n 9·10w
Sissela.....10·49n 10·37w
Takabara.....11·50n 11·30w
Tamba Dabatou..11·48n 10·40w
Télimélé.....10·54n 13·02w
Timbo.....10·38n 11·50w
Tindila.....10·16n 8·15w
Tinkisso.....11·15n 10·37w
Tintioulé.....10·13n 9·12w
Tiriro.....10·27n 8·39w
Tombadonkéa....11·00n 14·23w
Tougué.....11·27n 11·41w
Victoria.....10·50n 14·33w
Wassou.....10·02n 13·39w
Yambéring.....11·49n 12·21w
Yayouta..... 8·11n 8·30w
Yende Millimou.. 8·53n 10·11w
Yomou..... 7·34n 9·16w

Physical features and points of interest

Atlantic Ocean... 9·00n 14·00w
Bafing, *river*.....12·15n 10·20w
Bouka, *river*.....11·00n 10·50w
Camayenne, Presqu'île de (Kaloum, Presqu'ile de) *peninsula*..... 9·33n 13·40w
Dion, *river*.....10·12n 8·39w
Fatala, *river*.....10·13n 14·00w
Forest Region, *physical region*.. 8·30n 8·35w
Fouta Djallon, *physical region*..11·15n 12·20w
Gouala, *river*..... 9·57n 8·10w
Îles Tristao, *islands*.....10·53n 14·58w
Kakoulima, Mont, *mountain*..... 9·46n 13·27w
Kaloum, Presqu'île de, see Camayenne, Presqu'île de
Kogon, *river*.....11·09n 14·42w
Kolenté, *river*.... 9·15n 12·57w
Koliba, *river*.....12·15n 13·55w
Konkouré, *river*.. 9·58n 13·42w
Koulountou, *river*.....12·40n 13·30w
Kouya, *river*.....10·09n 9·45w
Los, Îles de, *islands*..... 9·30n 13·48w
Loura, Mont, *mountain*.....12·06n 12·17w
Lower Guinea, *physical region*..10·27n 13·33w
Mafou, *river*.....10·32n 10·08w
Makona, *river*.... 8·16n 10·42w
Milo, *river*.....11·04n 9·14w
Niantan, *river*....10·30n 10·26w
Niger, *river*.....11·35n 8·45w
Nimba Mountains..... 7·30n 8·30w
Nunez, *river*.....10·36n 14·40w
Ouassoulou, *river*.....10·30n 8·07w
Sankarani, *river*..11·20n 8·20w
Simandou Mountains..... 9·00n 8·30w
Tamgué, Massif du, *mountains*..12·00n 12·18w
Tinkisso, *river*....11·21n 9·10w
Tio, Pic de, *mountain*..... 8·52n 8·54w
Upper Guinea, *physical region*..10·40n 9·50w
Verga, Cap, cape.....10·12n 14·27w

The Forest Region, or Guinea Highlands, is an isolated area of hills in the country's southeastern corner. The region attains 5,747 feet (1,752 metres) at Mt. Nimba, which rises at the borders of Guinea, Liberia, and the Ivory Coast. The area is composed of the same rocks as those of Upper Guinea.

Three major West African rivers

Drainage and soils. The Fouta Djallon is the source of West Africa's three major rivers. The Niger River and several tributaries, including the Tinkisso, Milo, and Sankarani, rise in the highlands and flow in a general northeast direction across Upper Guinea to Mali. The Bafing and Bakoye rivers, headwaters of the Sénégal River, flow northward into Senegal before uniting to form the main river. The Gambia River flows northwestward before crossing Senegal and Gambia.

The Fouta Djallon also gives rise to numerous smaller rivers, such as the Fatala, Konkouré, and Kolenté, which flow westward across the coastal plain to enter the Atlantic. The Forest Region generally drains to the southwest through Sierra Leone and Liberia. The St. Paul River enters the Atlantic at Monrovia, Liberia, and the Moa River has its mouth at Sulima, Sierra Leone.

The most common soils are formed of hydrated aluminum silicates and other materials that often concretize into hard iron-rich conglomerates called laterites. In the northeast, sandy brown soils predominate, while along the coast black, heavy clay soils accumulate in the backwaters. There are alluvial soils along the major rivers. Soil conservation is extremely important because most soils are thin, and the heavy rainfall causes much erosion.

Climate. The average temperature of Guinea is more than 68° F (20° C). There are two alternating seasons—a dry season (November through March) and a wet season (April through October). The low-pressure zone caused by the summer solstice and its rapidly rising humid air brings the June rains. As the low-pressure area shifts southward in November, the hot, dry wind known as the harmattan blows southwest off the Sahara.

On the coast a rhythm of six months of dry weather is followed by six months of rain. The average rainfall at Conakry is 170 inches a year, and annual temperatures average about 81° F (27° C). In the Fouta Djallon, January diurnal temperatures range between 86° F and 95° F (30° C and 35° C), while evening temperatures dip to 50° F (10° C). Rainfall varies between 91 and 63 inches annually, and the average annual temperature is about 77° F (25° C).

In Upper Guinea rainfall drops to about 59 inches a year. During the dry season temperatures of more than 104° F (40° C) are common in the northeast. The drying harmattan winds blow hot in the day, but the night can be uncomfortably cool. In the Forest Region at Macenta, there may be 106 inches of rain annually. Only the month of January is relatively dry, with less than one inch of precipitation. At low altitudes, temperatures resemble those of the coastal areas.

Vegetation and animal life. The coast is fringed with mangrove trees, and the coastal plain supports stands of oil palms. The Fouta Djallon is mostly open, with trees growing along the wider stream valleys. In Upper Guinea, the savanna grassland is comprised of several species of tall grasses that reach heights of five to ten feet during the rainy season. Deciduous trees grow in scattered clumps, but few have commercial value; baobab and shea trees furnish fruit and oil. The Forest Region contains several extensive patches of rainforest, with teak, mahogany, and ebony trees; agriculture, however, has diminished the forests and resulted in a shift largely toward open savanna.

Vanishing animal life

Guinea is not a big-game hunter's paradise. Baboons, a variety of monkeys, and hyenas are common, while an occasional wild boar, several types of antelope, and a rare leopard may be sighted. Two or three small elephant herds exist in the savanna woodlands; several chimpanzee families are still found in the Fouta Djallon valleys; and lions have been noted in the northeast. The hippopotamus and manatee inhabit the rivers of both Lower and Upper Guinea. Poisonous snakes include mambas, vipers, and cobras, along with pythons and a variety of harmless snakes. Crocodiles and several varieties of fish are found in most rivers.

The landscape under human settlement. In the Fouta Djallon, the Fulani (Fulbe or Peoul) live in small hillside hamlets of 75 to 95 persons each, and the lower classes occupy the valleys. In the heart of the highlands the countryside is thickly settled with hamlets every few miles, while in the east the land is less settled.

In Lower Guinea, villages are grouped together at the bases of hills, in the open plain, or in a valley floor. Village solidarity is more marked in this area than in the highlands, and each village may contain between 100 and 200 people.

The majority of the Malinke (Mandingo) people of Upper Guinea live in moderately large villages of about 1,000 inhabitants located near permanent water sources, the adjacent soils of which are used for cultivation. The villages are tightly grouped; there are empty brush areas in which farming is unprofitable.

In the Forest Region the effects of human occupation are less apparent. Among the Kissi people in the west, rice is grown on most hillsides and in every low-lying and swampy area. Villages are small and rarely contain more than 150 people; they are often tucked away inside a grove of kola, mango, and coffee trees. Further east among the Loma and Kpelle people, fire-cleared land is used to plant vegetables and rice. Larger villages are usually located on remote hillside terraces that are often surrounded by secondary forest growth.

Guinea's urban centres

The only true urban centre is Conakry, with more than 197,300 inhabitants. The old city, located on Tumbo Island, retains the segregated aspect of a colonial town, while the Kaloum Peninsula community, which has grown up since the 1950s, has a few buildings of the colonial period. From the tip of the peninsula, a 20-mile long industrial zone has a growing salaried population that is truly urbanized.

The second largest town of Kankan (population 40,000), in Upper Guinea, is more a cluster of Malinke villages around an administrative and trading core area than a Westernized town. Labé (30,000), located well into the Fouta Djallon, serves as an old market town and an administrative and educational centre. Nzérékoré (23,000), in the Forest Region, serves the same functions as Labé. Other important towns are the trading centres of Kindia and Mamou and the industrial settlements of Fria and Boké (11,000).

People and population. *Population groups.* The four major geographical regions largely correspond to the areas inhabited by the major linguistic groups. In Lower Guinea the major language of Susu has gradually replaced many of the other indigenous languages and is a lingua franca for most of the coastal population. In the Fouta Djallon the major language is Fulani (Poular), while in Upper Guinea the Malinke language is the most widespread. The Forest Region contains the linguistic areas, from east to west, of Kpelle, Loma, and Kissi.

Eight official languages besides French are taught in the schools. They are Basari, Fulani, Kissi, Koniagi, Kpelle, Loma, Malinke, and Susu. All official texts are written in French and at least one of the indigenous languages, and all governmental personnel are required to have bilingual reading and writing ability. It is hoped ultimately to replace French with the Susu or Malinke languages.

All of Africa has been plagued by a proclivity on the part of Europeans to overemphasize distinctions between ethnic groups. Most definitions of what constitutes an ethnic group in Guinea are incomplete and subject to criticism because they often emphasize one criterion at the expense of others. The Fulani have come to dominate the Fouta Djallon culturally; the Malinke have widely influenced Upper Guinea and the northern Forest Region; and the Susu are dominant in Lower Guinea. In the Forest Region, however, the Kissi, Loma, and Kpelle each retain their own common historical and cultural identities.

Except for the diplomatic community and some expatriate teachers and technical advisors, there are few foreigners. The alien community also includes about

1,000 Lebanese and Syrian traders and a few Frenchmen engaged in plantation agriculture and technical occupations.

Religious affiliations

At least 62 percent of the population is Muslim, and about 2 percent of the population is nominally Christian, mostly Roman Catholic. Since 1967, all priests have been required to be either Guinean nationals or Africans. There are about 2,200 Protestants.

Demography. Guinea's population was estimated at 4,010,000 in 1971. The growth rate is about 2.2 percent per year. The birth rate is a high 47 per thousand, with ten births for every female; the infant mortality rate, however, is such that only five or six children in ten reach 18 years of age. Mortality is also high (25 per thousand), with an average life expectancy in 1955 of 28 years for females and 26 years for males.

Immigration is minimal, but emigration—especially from the Fouta Djallon and Upper Guinea—is high. About 15 percent of the male population from these areas has migrated in search of work, leaving an imbalance of females. Emigration is directed toward the neighbouring countries of Sierra Leone, Senegal, Mali, the Ivory Coast, and Liberia.

The heaviest population concentration is in the Fouta Djallon; the area around Pita and Labé sustains more than 130 people per square mile. The rich rice areas around Guékédou and some of the land around Nzérékoré in the Forest Region have a density of 65 to 130 people per square mile. Most of the rest of the country has fewer than 65 people per square mile, and over half the land area has fewer than 26 persons per square mile. Conakry and the Kaloum Peninsula suffer from rapid population growth caused by a never-ending exodus from the rural areas to the city.

The population is projected to be 5,016,000 by 1980. Except for the Fouta Djallon, this population explosion poses no serious immediate threat to development because there is no pressure on the land and no landholding class.

The national economy. *Natural resources.* There are more than 1,000,000,000 tons of bauxite reserves at Boké, Dabola, and Fria. Guinea has among the largest iron-ore deposits in the world, totalling at least 200,000,000 tons at the Kaloum Peninsula and the Nimba and Simandou mountains. Gold occurs along the Niger and its tributaries, and diamonds occur in the gravels of the Makona River tributaries. The southeastern rain forests offer valuable tropical hardwoods, and the ocean and rivers contain food fishes. Hydroelectric potential is considerable because of the high rainfall and deep gorges of the Fouta Djallon.

Bauxite and diamonds

Sources of national income. Guinea is an agricultural nation. The high plateaus of the Fouta Djallon are little more than part-time pastures, with hillsides given over to the growing of peanuts (groundnuts) and fonio (a sorghum-like grain). Along the streams and rivers, rice, bananas, tomatoes, strawberries, and citrus fruits are grown commercially. Most families have truck gardens, and tsetse resistant Ndama cattle, sheep, goats, horses, donkeys, chickens, and an occasional Muscovy duck are raised.

In Lower Guinea, oil and coconut palms, rice, bananas, salt, and fish are important elements of trade. Except for poultry, there are relatively few domestic animals. In Upper Guinea, grains and manioc (cassava) are important food crops; vegetables, tobacco, and *karite*, or shea butter, are traded locally; and domestic animals are common.

In the southeast rice is the chief food crop along with manioc, peanuts, and maize (corn). Gardens of tomatoes, peppers, and tobacco are scattered in the shade of fruit trees, and coffee trees and oil palms provide important cash crops. Goats and fowl are the most common domestic animals.

By the early 1970s, experiments with large-scale cooperative agricultural production had failed. Relatively low government farm prices and the high cost and scarcity of consumer goods caused many producers to return to subsistence agriculture or to resort to smuggling. The production of coffee—the major cash crop—varied little between 1958 and 1968. Rice output has grown, but the staple continues to be imported. Other cash crops, such as palm kernels, peanuts, pineapples, and citrus fruit, only began to be important in the 1970s.

Commercial fishing is little developed, although two boats furnished by the Soviet Union have raised production considerably. Traditional ocean and inland fishermen continue to be important, and smoked and fresh fish are sold in local markets.

Forestry is hampered by the lack of transportation. The government-owned sawmill and plywood plant at Nzérékoré and wood panelling plant at Seredou cannot function to capacity because neither area has sufficient forests to continue exploitation.

Mining and quarrying

Guinea depends heavily upon the bauxite plant at Fria, owned by an international consortium of private companies from the United States, France, the United Kingdom, Switzerland, and West Germany. Guinea's production of alumina furnishes more than 60 percent of all exports, and Guinea is the sixth largest producer of bauxite. The Boké project, which is scheduled to begin in 1972, will produce 5,000,000 tons of bauxite annually and have a total yearly capacity of 8,000,000 tons. Combined with Fria's production of about 1,600,000 tons annually, the Boké project will eventually raise Guinea's total yearly bauxite production to more than 9,000,000 tons. Two other projects of the 1970s were a Soviet-financed mining operation at Dabola and a Swiss-aided scheme for the exploitation of bauxite in the Tongue Mountains of the Fouta Djallon.

The iron-ore deposits of the Nimba and Simandou mountains are scheduled to produce for export in the 1970s, replacing the nearly exhausted Kaloum Peninsula deposit. Limestone deposits in the northeast were to be developed by a Spanish company that was constructing a cement works in Lebekere. Diamond and gold production are minimal.

Food-processing plants do not run at full capacity because agricultural production is insufficient. Establishments include a fruit-juice factory at Kankan, a tea factory in Macenta, a palm-oil works at Kossa, and a peanut-oil works at Dabola. Government factories produce leather, shoes, matches, cigarettes, processed meat, textiles, and furniture. Semiprivate operations include the brewery in Conakry, a soup cannery, and several small-scale fruit-juice and plastics concerns.

In 1970 the hydroelectric station near Pita began to supply electricity to Pita, Labé, Dabola, and Mamou. The hydroelectric plant at Kale, backed by two diesel-powered thermal centres, supplies power to Conakry. There is a small hydroelectric station in Tinkisso; Fria has its own thermally-produced electricity supply; and many regional towns have their own diesel-operated generators. In the late 1960s total electric production amounted to 212,000,000 kilowatt-hours annually.

Financial services

Guinea has a nonconvertible currency; it is not tied to any foreign currency, and its value is dictated by the Guinean government. In consequence, domestic financial affairs are really a matter of national bookkeeping. The national bank, the Banque Centrale de la République de Guinée, issues currency, the Banque Guinéenne du Commerce Extérieur handles all foreign exchange, and the Crédit National pour le Commerce, l'Industrie et l'Habitat serves as a savings-and-loan bank. Agricultural credit is obtained through the Banque Nationale de Dévéloppement Agricole.

All foreign trade is theoretically a state monopoly. Specialized government agencies, such as Alimag, are in charge of importation of various goods: Guinexport is the sole export agency. Guinea trades in about equal measure with the east European Socialist bloc countries and China; the franc-zone nations, such as France and Senegal; and other Western countries, including the United States and West Germany.

Management of the economy. The private sector is largely a paper presence or a clandestine operation. Approximately 500 private businesses continue to operate, but they must obtain merchandise from government

sources. Most of the larger foreign companies are local capital investments that are no longer profit-making enterprises.

The wholesale trade is largely controlled by the Ministry of Commerce, while the retail trade is theoretically in the control of the state. Retail stores are operated in the larger towns, but, within rural areas, cooperatives are expected to distribute imported and some local produce.

Government revenue is chiefly derived from mining concessions, import and export duties, excise taxes, a petroleum-products tax, and taxes on commercial transactions and production. There are also various other surtaxes, stamp duties, and registration fees. Business and other licenses and personal-property, building, dwelling, and vehicle taxes are handled by the regional administration. Taxes on salaries and wages contribute little revenue because few people are salaried and because most of the wage earners work within the government.

The national trade union

All workers in Guinea are required to join the National Confederation of Guinean Workers (CNTG), which is affiliated with the Union of the Workers of Black Africa (UTAN). Both organizations are actually an arm of the party-government apparatus. Wages are fixed by government decree and a 48-hour workweek is officially in force for industrial workers. The Chambre Économique de Guinée (Chamber of Commerce) is a vestige of the former French- and Lebanese-owned "free-enterprise" establishments. Banana marketing boards and other cooperative agricultural marketing systems have given way to state enterprises.

Guinean policy seems to be based on the belief that increased mining will stimulate the entire economy. Associated projects are viewed as improvements of transportation and communications that will, in turn, help agricultural production. There is a serious shortage of trained personnel, and finances suffer from misappropriation, tax evasion, and smuggling. Many of the processing industries are held back by inadequate supplies of raw materials. Internal production is not sufficiently high, in agriculture particularly, and there is a permanently deficient trade balance. It is doubtful whether the disproportionate development of bauxite to the neglect of capital investment in agriculture will redress economic problems.

Transportation. Guinea's transportation system is largely based upon the road and railway from Conakry to Kankan and a lesser road to Nzérékoré. This forked axis is intersected at Mamou by a road north to Senegal. East of Kouroussa the main road branches northeastward through Siguiri to Bamako, Mali. The main road continues northeast of the railhead at Kankan to Sikasso, Mali. The regional centres, like pods strung out on a vine, lie along thin lines of communication that, in turn, radiate feeder routes.

There are an estimated 4,725 miles (7,604 kilometres) of roads, of which 1,650 miles were unimproved in 1970. The 411-mile railroad from Conakry to Kankan is a single-track metric line. It needs track and roadbed repairs throughout most of its length from Kindia. There is also an 89-mile line linking Conakry to the Fria bauxite mines.

The port facilities of Conakry are extensive. There is a channel 26 to 66 feet deep that needs constant dredging, and more than 7,400 feet of dock space with modern loading equipment. Coastal shipping, however, is limited.

The international airport at Conakry serves jets of all sizes. Air Guinée operates weekly domestic flights to the packed-earth landing strips at Kankan, Boké, Labé, Kissidougou, Nzérékoré, and Siguiri and maintains occasional service to Bamako, Mali; Dakar, Senegal; Freetown, Sierra Leone; and Monrovia, Liberia.

Administration and social conditions. *Governmental structure.* Guinea is a one-party state in which the Parti Démocratique de Guinée (PDG) and government are one. Since 1968, major government functions have been in the hands of the president and the six ministers of foreign affairs, economics, finance, social affairs, internal affairs, and trade. There are also four deputy ministers, appointed by the president, who represent each of the four supra-regions in the council of ministers. The president and ministers also compose the Bureau Politique National (BPN), which, as party executive, controls administration. The BPN is elected by the party's National Congress every four years. The president is charged with general supervision of the party and the presidency, with secretaries of state in charge of ideology, the army, scientific research, and administrative coordination.

The 75-member National Assembly ratifies the party's decisions. Its members are elected by universal suffrage for a five-year term from a single slate of candidates. The president is elected for a seven-year term. Those who hold key posts in government or national enterprise also hold party positions and may be leaders in the national trade union.

Local government

Local administration is theoretically straightforward. The village commune is administered by a council that consists of the party's village executive council, the president of which also serves as mayor. The village council and mayor are, in practice, appointed by the commandant of the next administrative level, the *arrondissement*. The commandant is appointed by the president. A party executive council of 12 members constitutes the *arrondissement* council, with the commandant as 13th member, party executive, and chairman. The *arrondissements* are grouped into 29 regions, each of which has an assembly.

Each of the four supra-regions of Upper Guinea, Forest Guinea, Middle Guinea, and Maritime Guinea is headed by a governor appointed by the president. The Regional Assembly is composed of the party's executive group at the regional level; party leaders from the *communes* and *arrondissements* act in an advisory capacity. Administrative affairs are coordinated by the governor and the federal secretary, who heads the assembly. At the top of this structure are four minister delegates, one for each of the four supra-regions appointed by the president.

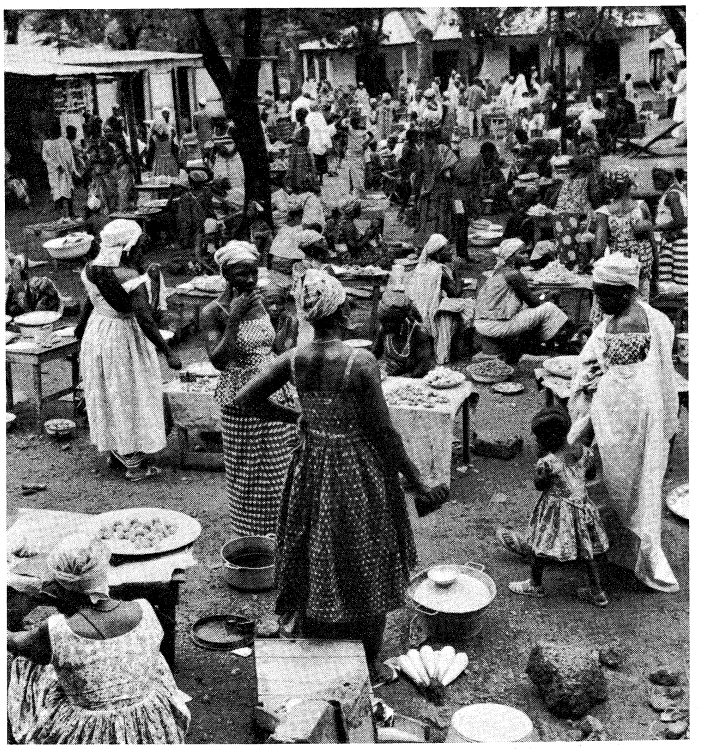
Citizen participation in government is a function of participation in the party. At the village or neighbourhood level, the general assembly of the Base Committee meets every Friday. At the *arrondissement* level the Sectional Congress of Base Committee delegates meets every two years, while the Federal Congress, composed of members elected by the Sectional Congress, meets every three years and the National Congress of delegates from the federal congresses meets every four years. PDG membership is universal for all over 17 years of age who have no criminal record since independence and who are not "exploiters." The party decides most civil-court matters; the criminal courts are controlled by appointment from the presidency and are closely linked to the PDG.

The army consists of 5,000 men dispersed throughout the country who are employed in rural economic development projects. The 200-man air force is one of the best equipped in black Africa; the navy consists of 200 men equipped with patrol boats. As well as these regular forces, the PDG has a People's Militia of about 30,000; militia instruction is compulsory in all schools. Functions of the militia include party mobilization, internal police services, and external defense. All forces are united under the Defense Council, which is headed by the president.

Administrative services. Education is free and compulsory for nine years. The government sees it as one of the chief means of restoring authentic African values. Teaching of the French classics and arts curriculum has been reduced in favour of an emphasis on technical training. In the early 1970s there were about 1,600 primary schools, and about 35 secondary schools. The country's three colleges—Institut Polytechnique de Conakry, École Nationale des Arts et Métiers, and École Supérieure d'Administration—are all located in Conakry. In the mid-1960s, 18 percent of the primary school-age children attended classes.

Health and welfare services

Since independence, Guinea has devoted much effort to health services. In 1970 there were five major hospitals, 29 regional hospitals, and almost 300 *arrondissement* dispensaries. Vaccinations against smallpox and polio were not universally available in the early 1970s, but malaria,



Market day in Dabola, Guinea.
Paul Conklin—Pix from Publix

trypanosomiasis (sleeping sickness), and onchocerciasis (river blindness) had been curbed. Medical care and personnel, however, are inadequate. The Institute of Traditional Medicine was created in 1967 to aid in making indigenous medical remedies more available. Such social-welfare services as infant clinics and child-care services are largely a function of the extended family.

A housing problem exists only in the most urbanized areas because local construction materials are otherwise available. In the industrialized areas near Conakry, attempts have been made to construct low-cost prefabricated housing in workers' cities near the plant sites.

Social conditions. The salaried labour force numbers little more than 110,000. In general, salaries are low, and little more than subsistence is the lot of more than half of the salaried workers. Rents have been lowered by government decree; purchasing cooperatives have been established; and salaries have been progressively adjusted in favour of the least paid. Inflation, however, has produced a continual need to resort to extra-salary means in order to eke out a livelihood.

Guinea has largely succeeded in creating a classless society. Although there are inequalities in wealth and although party and government officials abuse their positions for self-gain, the general Guinean pattern has been one of control of corruption and excessive profit taking. The rural farmer is perhaps no better off in the 1970s than he was in the 1950s, but neither is he much worse off.

Cultural life and institutions. A profound cultural revolution was launched in 1967, aimed at involving all of the people in national construction by developing a sense of national responsibility and an awareness of the part that communal action must play to achieve complete decolonization of the mind, the economy, and national politics.

Members of the professional Ballet National Guinéen are chosen from ballet, theatre, and traditional and modern music groups from each village, section, and federation. They form two national troupes that are creating a national Guinean form of the performing arts. Because the performers are often from *griot* families (traditional families of artistic specialists) and because they are often paid by the federation or village groups they represent, their traditional roles are still maintained.

A distinctively Guinean literature is difficult to define. Such authors as Camara Laye, Djibril Tamsir Niane, and Mamadou Traore Rayautra have made contributions to a fusion of traditional forms and patterns into a universally understandable literary genre.

Handicrafts have almost become extinct because they are unable to compete with manufactured consumer goods. Those remaining include leatherwork, masks, statuettes, art objects, and jewelry for the tourists.

The mass media

The PDG is the only source of information in Guinea. The only daily newspaper published is *Horoya*, the party daily written in French, and Radio-Conakry, the "voice of the revolution," is the main source of foreign and national news. Daily programs reach a large percentage of the population with broadcasts in French, English, Portuguese, Krio (a language derived from English and spoken in coastal Guinea and Sierra Leone), Arabic, and the official Guinean languages.

Prospects for the future. Guinea is underpopulated and has a rich natural-resource base, and its government has largely succeeded in creating a strong sense of national identity. Whatever problems its anticolonial and revolutionary stance may have created, the country is in control of its own political and economic future. The vocabulary of the PDG has, perhaps, been more Socialist than libertarian, but the pragmatic course of political and social progress seems to have created a truly Guinean system based upon an African model.

The main problem is that the somewhat simplistic reliance on inevitable progress and the overdependence upon mineral resources has yet to be tempered with a consistent and realistic appraisal of world economics. Neither bauxite nor iron ore are Guinean monopolies, and, without the cooperation of international capital, they have little practical value. Agricultural production must be improved in order to maintain internal harmony and to achieve self-sustaining economic growth.

BIBLIOGRAPHY. There are two major sources of information on Guinea. JEAN SURET-CANALES, *La République de Guinée* (1970), is the best single source by a trained social scientist and long time resident of Guinea. The *Special Warfare Area Handbook for Guinea* (1961), from the American University for the Department of the Army is dated and less reliable. GUY DE LUSIGNAN, *French Speaking Africa Since Independence* (1969), though somewhat simplistic and superficial, should also be consulted for political and economic interpretations. More current and accurate information may be found in the yearly *Africa Contemporary Record*. Statistical information may be gleaned from the *United Nations Statistical Yearbook* and the *AID Economic Data Book: Africa* (1970). The yearly encyclopaedia *Africa*, available from the African Publishing Corporation, is also useful. LADIPO ADAMOLEKUN, "Politics and Administration in West Africa: The Guinean Model," in the *Journal of Administration Overseas*, 8:235–242 (1969), is a lucid examination of political and administrative realities. Works by President SEKAU TOURE describing social and political change within Guinea are *Toward Full Re-Africanization: Policy and Principles of the Guinea Democratic Party* (1959) and *La Révolution guinéenne et le progrès social* (1963).

(T.E.O'T.)

Guinea, Gulf of

The Gulf of Guinea, narrowly defined, is part of the eastern tropical Atlantic off the West African coast, from Cape Lopez, near the Equator, to Cape Palmas at 7° west. Its coastline forms part of the western edge of the African tectonic (rock structure) plate and is the mirror image of the South American coastline of Brazil and the Guianas. The remarkable coincidence between the geology and the geomorphology of these two coastlines constitutes one of the clearest confirmations of the theory of continental drift.

The continental shelf of the Gulf of Guinea is almost uniformly narrow, widening to as much as 100 miles only from Sierra Leone to the Bijagós Archipelago, Portuguese Guinea, and in the Bight of Biafra. The shelf deposits are predominantly of Quaternary sands, with linear Holocene (Recent) fossil coral reefs near the shelf edge. Recent river muds overlie these deposits off the mouths of the rivers—Casamance, Volta, and Niger—which empty

into the Gulf. The Niger has built a great delta of Holocene muds—and it is only here that the fit between the African and South American tectonic plates is seriously disturbed—while the Congo has built a unique submerged delta on the continental slope.

The only active volcanic region is the island arc aligned with Mt. Cameroon (13,353 feet, or 4,070 metres) on the coast of the Cameroon Republic; the islands (Fernando Po, Príncipe, São Tomé, and Annobón) extend 450 miles offshore to the southwest.

The entire northern coast of the Gulf is washed by the eastward flow of the Guinea Current, which extends 250–300 miles offshore from Senegal to the Bight of Biafra. Its tropical water is separated from the cool Equatorward flow of the Benguela and Canary currents by sharp frontal regions off the Congo and Senegal rivers respectively. The Benguela Current, as it swings westward, forms the South Equatorial Current to the south of, and running counter to, the Guinea Current.

The warm tropical water, of relatively low salinity because of river effluents and high rainfall along the coast, is separated from deeper, more saline, colder water by a shallow thermocline (a layer of water, between upper and lower levels, in which temperature drops more than 1° C per metre of depth) lying usually less than 30 metres deep. Coastal upwelling, and hence rich production of plant and animal life, occurs seasonally and locally off Senegal and the Congo, and off the central Gulf coasts of Ghana and the Ivory Coast.

The variety of the marine flora and fauna of the Gulf of Guinea is limited, when compared with the western tropical Atlantic and, especially, with the Indo-Pacific biogeographic realm. This relative biological poverty results from (1) a lack of coral reef ecosystems (except on Annobón Island) because of low salinity and the high turbidity of Guinea Current water, and (2) the climatic regression to cool conditions during the Miocene Epoch during which far fewer faunistic refuges for tropical species of animals and plants were available in the Atlantic than in the Indo-Pacific region.

Because of the difficulty of the coast, most of which is low-lying and without natural harbours, and which is largely separated from the dry land of the interior by a belt of muddy mangrove-infested creeks and lagoons, the African coastal peoples have usually not taken easily to seafaring on the Gulf. Tribes located in Senegal and Ghana, where the coast is easier, and coastal fisheries are relatively productive, form an exception. Demersal (*i.e.*, sea bottom) fisheries of the continental shelf are restricted to a coastal zone only a few miles wide and can support only modest fleets of small trawlers, though in some areas shrimp resources may yet prove to be a valuable export resource.

Sardine fisheries off Senegal and Ghana are incapable of supporting exploitation by large factory trawlers, such as the Soviet vessels that have fished there in recent years; the offshore tuna fisheries, on the other hand, now support a large and modern fleet of Spanish, French, and American purse-seining vessels and yield about 30,000 tons annually, little of which is consumed in West Africa.

Exploration for offshore oil proceeds rapidly, and exploitation on a large scale has begun off the Niger Delta. Extraction of heavy metals (titanium from rutile sands) from beach sand deposits in Sierra Leone has begun, and continental shelf deposits in this region may also prove rich in metal ores and diamonds.

Phoenician, Roman, and—from the 15th century—European trading vessels sailed the Gulf of Guinea, which was also the scene of naval efforts to suppress the slave trade in the 19th century. In the future, development of its resources will depend primarily on the availability of capital from outside Africa. Perhaps the major question now posed is to what extent the African states bordering the Gulf will share in the economic benefits from its exploitation.

BIBLIOGRAPHY. For sources of additional information on the oceanographic region of which the Gulf of Guinea forms a part, see bibliography of the article ATLANTIC OCEAN; for information on fisheries, see the bibliography of the article FISHING, COMMERCIAL.

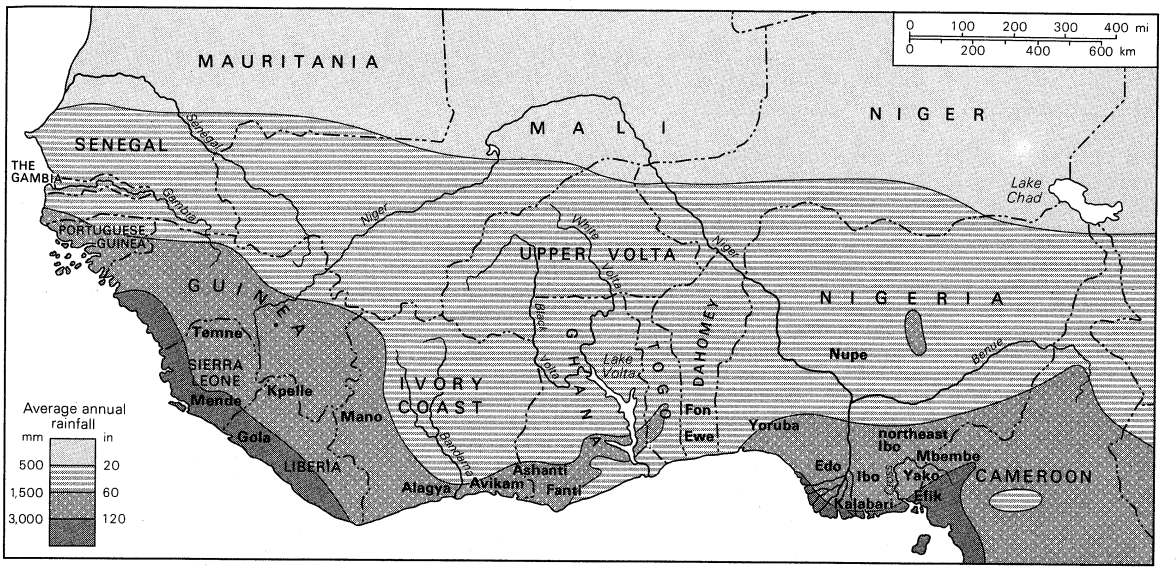
(A.R.L.)

Guinea Coast, Cultures of the

Guinea is a term used for the coastlands and adjacent forests of West Africa between the Republic of Guinea and Equatorial Guinea. There have been conflicting interpretations of the derivation of the name Guinea, but it would seem to be a version of the Berber word *aguinaw*, or *gnawa*, meaning "black man," or "Negro."

THE ENVIRONMENT AND THE PEOPLE

In West Africa, in the general absence of major mountainous areas, natural regions are determined primarily by climate and vegetation, and the Guinea coast societies are those broadly associated with the equatorial forest zone. This forest has been erased by agriculture in some areas, but it is the natural vegetation of most districts within 100 miles of the coasts—that area within which average monthly temperatures always exceed 70° F (21° C) and the annual rainfall is well distributed, with a dry season that does not exceed three months (from December to the end of February) and with a total precipitation that is more than 45 inches (110 centimetres). In the east a heavy forest formerly extended from the borders of the Cameroon highlands to the west of the Niger River. In the west a forest stretched from



Distribution of Guinea coast peoples.

Sierra Leone to the west of Ghana. Between these two belts of forest there is, for very complex reasons, a drier region, and here for centuries the forest has been thin, and the land, when cleared for farming, has changed to grassland if left fallow. Societies here have been culturally very similar to the true forest societies but have shown some significant differences.

Cultural effects of the forest

The forest greatly influenced the cultural development of the Guinea coasts, for it influenced the movements of peoples and the development of agriculture and commerce. Men occupied the forest areas relatively late, because farming there had to await the development of suitable tools and new crops. Iron axes are needed to clear equatorial forests efficiently, and, although iron was introduced south of the Sahara between 500 BC and AD 200, the quantities available were too small to be significant until much later. Moreover, the crops that were indigenous or introduced early to West Africa were unsuited to the dense shade of small forest clearings. The yam, for long the most important staple crop in much of the forest region, succeeds only where large areas are cleared. Only with the introduction of shade-tolerant crops, the plantain and cocoyam (taro or eddo) brought from Asia in the 9th century AD, could forest farming become an economic alternative to hunting and gathering. Moreover, until recently forest farming did not include animal husbandry, for the forest has harboured species of tsetse fly that are particularly virulent to cattle and horses. This fact, however, had its advantages for the forest peoples, because the tsetse fly and the trees protected them from marauding cavalry. Gradually, too, the forest gave its inhabitants commercial advantages: kola nuts and later palm oil were so highly desired by distant peoples that traders by caravan and ship were drawn to the Guinea coasts.

Language distribution

The cultural significance of the original forest environment seems shown even today in the linguistic map of West Africa. In detail there is a host of different languages in the forest area, some spoken by millions, some by a few thousand people. What is striking, however, is that the boundaries between the major language families roughly coincide with the old boundaries of the forest. In the extreme west, in the area of Sierra Leone, a history of multiple small invasions has brought a confused linguistic pattern to the forest areas, but there is a clear division between these forest language groups and the great groups of Mande languages in the bordering savanna belt. The forests of Liberia and the western Ivory Coast are inhabited mainly by speakers of Kru languages and bordered again by Mande speakers in the grasslands. From the Bandama River in the Ivory Coast eastward beyond the Niger River there is the clearest evidence of all of a cultural boundary between the old forest zone and the grasslands to the north, for almost all the different languages of the forest region, including those of the forest gap, are classified as belonging to the Kwa family and are bordered in the grasslands to the north by Gur and Hausa languages. Only in the extreme east does this clear division disappear. The linguistic division between the Guinea coast and the hinterland, in any event, provides good grounds for distinguishing Guinea coast societies.

This article cannot deal individually with all the groups in the area but only the more important: (1) certain peoples of the western forest, including the Mende and Temne of Sierra Leone, the Kpelle of Liberia, and the Akan of Ghana; (2) certain peoples living mainly in the lightly forested gap, including the Fanti of Ghana, the Fon and Ewe of Dahomey, the Yoruba of western Nigeria, and the Nupe; and (3) certain peoples of the eastern forest, including the Edo of Benin, the Ibo, the Cross River peoples, and, on the coast, the Efik and Delta peoples.

CULTURAL PATTERNS

Guinea coast societies vary enormously. Today, differences between groups emerge most clearly in differences in density of population and types of settlement. Even within a small area such as southern Nigeria great variations exist: in certain Ibo rural areas there are 700 or more people per square mile, whereas in the equally fertile forest hinterland of the Cross River densities are well under 100 per square mile (40 per square kilometre). Moreover, Ibo live characteristically in scattered hamlets, whereas to the west the Yoruba have built the largest indigenous towns in Africa.

Interplay of environmental and historical influences

Such cultural differences are usually caused by the complex interplay of environmental and historical factors, even when the direct influence of the physical environment seems most obvious. This can be seen by looking at patterns of crop production. Sometimes crop choice seems dictated solely by physical factors—for instance, when farmers in dense forest have to grow cocoyams rather than the yams that they would prefer. But sometimes tradition seems also to be involved. In the Ivory Coast, for instance, the staple crop to the west of the Bandama River is rice, whereas to the east it is the yam. Although this difference seems, in a general way, to correspond to a difference in rainfall—the yam, requiring even watering, growing in an area of better distributed rainfall—the fact that the Bandama River also marks a linguistic boundary suggests that there is a strong element of historically determined cultural choice in this crop boundary.

That other types of cultural differences also have explanations less exclusively environmental than at first appears can be seen from a study of house types. Most Ibo, for example, live in typical forest-area houses—mud-walled, rectangular structures with gable-ended roofs thatched with mats made from single palm fronds. The northeastern Ibo, however, having moved into a grassland region where palm trees are scarce, today build a house with low circular walls and a conical roof thatched with grass. Environmental factors forced these people to adopt new materials, but the design details of their present houses owe as much to cultural borrowings from their new neighbours as to environmental necessity.

The more complex the cultural pattern is, the more complex is the interplay of environmental and historical factors behind it. The development of the large Yoruba towns owed much to the necessity for defense in a relatively open territory; the Ibo, by contrast, in their sheltering forests, needed less protection from attack and could risk living in scattered houses. Here at first sight is a simple relationship of environmental and military factors. But not all forest dwellers live like the Ibo. The farming villages next in size to the Yoruba settlements were built by the Cross River Yako, even though they live in a heavily forested area and have experienced little serious warfare. Thus, a full explanation of differences in settlement pattern would require the detailed examination of many factors influencing tradition.

Historical background of trade and politics. Given the circumstances outlined above, it can be appreciated that factors accounting for the political variations of Guinea coast societies are extremely complex. At the end of the 19th century, there existed both tiny, independent political units of four or five villages and large kingdoms of medieval foundation holding sway over hundreds of thousands of people. Older writers sometimes tried to account for such differences in simple environmental terms. Environmental factors undoubtedly were extremely important, but explanations must also take account of the economic patterns and political struggles of past centuries, and especially of the contacts between the Guinea coasts and the outside world, both northward and across the seas.

Shift from Arab to European trade

Before the end of the 15th century, most of the region's external contacts were made through the savanna kingdoms to the north, whose merchants wanted slaves, gold, and kola nuts. From the end of the 15th century, however, the interests of the Guinea coast peoples were partly reoriented toward trade with European merchants, who sought successively gold, slaves, and palm oil. European trade was significant partly because within the Guinea coast areas it was entirely controlled by local people. Europeans were prevented by climate and disease and

also by the express action of African authorities from penetrating inland. Consequently, all the inland traffic in gold, slaves, and palm oil was organized by African traders; Europeans merely provided inducements to sell.

The European trade was significant also because of the nature of the goods involved. Imports from Europe consisted mainly of such consumer goods as clothes but also included such capital goods as iron, guns, and gunpowder, which gradually introduced a crucial new factor into Guinea coast warfare, making it militarily vital for groups to acquire the new weapons. Moreover, even imported consumer goods had a political significance, for often it was the political authorities able to tax European merchants or their own traders who were able to acquire most goods. This put new wealth, new economic and political resources, and thus new power into the hands of some authorities, with important political repercussions.

Effects of the slave and palm-oil trade

The development of Guinea coast societies was also strongly influenced by the nature of what they exported. Slave trading did not everywhere lead to raiding and depopulation. Ibo slaves, for example, seem to have been sold mainly by authorities as a punishment for crime or by senior kinsmen who had got into debt; there is no evidence of depopulated Ibo districts. Nevertheless, the slave trade was commonly associated with warfare directed toward the acquisition of captives for sale. Sometimes, as in an area of northern Ghana, a low population even today is considered attributable to such raids. More generally significant for political development was the stimulation that raiding gave to military and political organization; groups that raided for slaves roused the antagonism of their neighbours, who made retaliatory attacks, and all this increased the tendency to military centralization, as, for example, in Dahomey, where a strong government was closely involved with slave raiding in every dry season. Moreover, wherever war captives were important as slaves, the trading position of central authorities who controlled captives was better than that of private traders.

In the 19th century, palm oil gradually became the most important Guinea coast export because of its rising use as an industrial lubricant and because of the overseas restraints on the slave trade imposed by European humanitarians. Nevertheless, slave dealing and raiding remained important internally throughout the century. Even when slave exports declined, to the anger of the local traders, the increased palm-oil traffic in itself stimulated internal slaving; the reason was that the transportation of the bulky oil required the use of slaves, either to paddle the canoe transports or to headload the oil to the ports. Whenever a trader had many men engaged in oil transport, moreover, he needed other slaves to produce food for them; thus, slaves were important economically and politically to the Mende, Ashanti, Fanti, Dahomean, Yoruba, Niger Delta, and Efik peoples.

The growth of the palm-oil trade brought other economic and political changes. African traders exporting palm oil needed more capital than slave dealers did, because slaves transported themselves and worked while awaiting shipment, whereas oil was expensive to transport and constituted idle capital while at the ports. Palm-oil exporters therefore needed capital and became increasingly reliant on European firms who advanced them goods on credit and took increasing interest in their affairs. This was politically significant when the African exporter was also a political leader, and it was one factor leading to colonialism. Paradoxically, in the rural areas many men who could never have been slave traders could easily gather and sell palm nuts. In Dahomey most palm oil came from government-fostered large plantations, but elsewhere the oil was drawn mainly from trees growing naturally in the bush; in the eastern forest area especially, participation in this production and trade was very general.

Markets and traders

By the end of the 19th century, a network of local markets had been developed over much of the hinterland of the Guinea coasts. One indication of the importance of markets is that in many societies the days of the week are named after local places holding markets on those days—village life revolved around "our" market and those of "our" neighbours. The great centralized kingdoms were naturally associated with great trade routes, but there is evidence that traders travelled quite safely, even in the absence of strong governments able to enforce peace, because the desire for regular supplies of trade goods protected the accredited trader even from ambitious headhunters.

Kingdoms and chiefdoms of Guinea. Although trade could flow across political boundaries, the political development of West African societies was much influenced by the growth of trade—through the warfare and struggle for trade routes that accompanied it. In the area of Sierra Leone and Liberia, where trade was initially rather limited, most political units remained very weak, making possible the later settlement there of freed slaves. Yet later, however, Mende tribes, located in rich oil-palm areas, fought for the control of trade routes and developed into more centralized groups under warrior chiefs. At the other end of the western forest, the Ashanti confederation developed quite differently. The area early had commercial importance as a source of gold and of the best kola nuts (one stimulant allowed to Muslims), which were wanted by northern traders. From the 17th century the Ashanti exploited their gold resources, easily made a government monopoly, in order to gain local control over the import of firearms. Extending their influence over their immediate neighbours by diplomacy and over those more distant by warfare, they eventually subjugated peoples as far southeast as Accra and as far north as the savanna.

In the gap in the forest, where there was easier movement, there developed particularly powerful kingdoms. Dahomey, essentially a militaristic state of 18th-century origins, had centralized control of raiding and trading. The Yoruba kingdoms, whose medieval growth was linked to northern trade, became in the 18th and 19th centuries increasingly concerned with southern trade routes. This, plus a complex internal power struggle, helped to plunge the Yoruba kingdoms into war with one another. Northward, the Nupe kingdom owed much of its original importance to its key position on trade routes to the north.

In the eastern forest, northwest of the Niger Delta, the Kingdom of Benin was also of medieval origin and was so well established in the 16th century that it became the first sub-Saharan state to exchange ambassadors with a European power (Portugal). From the 16th century Benin rather unsuccessfully competed with its Yoruba neighbours for the control of trade routes to the coast. Eastward across the Niger there grew up very different political systems. The Niger Delta is one of the greatest mangrove swamps in the world, an area without land for farming and unsuited to the development of large states. Here there developed small, independent trading towns wherever a deep-water anchorage suitable for European ships existed adjacent to rivers giving good access to the interior. These towns were interested in exerting commercial control rather than political power over the slave-rich and oil-rich Ibo hinterland.

It was once assumed that Guinea coast kingdoms must have borrowed their political institutions from northern Islāmic states, but this idea is no longer accepted. In contradistinction to the situation in Islāmic states, the Guinea coast king was usually the keystone of the political system because of his ritual relation to royal ancestors believed to exercise power over the kingdom. Moreover, the king, sometimes transformed into a semi-divine personage through his installation rites, had contact with the gods so close that all his actions had to be circumscribed lest by breaking taboos he brought disaster to his people.

Political structure of the kingdoms

In Oyo, one of the best described Yoruba kingdoms, the king, at the culmination of his installation rituals, ate the heart of his predecessor and was transformed into a personification of his ancestors. Thereafter, on his only public appearances, at rituals held three times a year,

he appeared veiled, his face hidden by a beaded fringe. Those who formally represented him in judicial, religious, military, and administrative capacities were slave eunuchs, chosen because, having neither kin nor affines, they presumably had no interests to serve but their master's. Though secluded, the king was involved in important political manoeuvrings, playing one group of hereditary chiefs off against a second and trying to avoid the great danger that would ensue if both groups were to unite against him.

This description gives the barest outline of the political structure of Oyo, which, at its most powerful, in the 18th century, controlled an area from the borders of Benin to present-day Ghana. Its political structure had many parallels in other Guinea coast kingdoms. Such political structures can be described as if they were well-integrated systems of checks and balances persisting unaltered for generations, but modern research suggests that these structures were altered whenever the balance of political power altered. Kings and hereditary chiefs might compete, for example, over the rules for choosing the king's successor, for it was recognized that the greater the freedom of choice possessed by the hereditary kingmakers, the weaker the king would be. Similarly, there was competition between king and hereditary chiefs over the king's power to appoint subordinate officials. In the Oyo and the Ashanti kingdoms there is evidence that this was a cause of struggle whenever new territories were conquered, for there was competition to claim the administrative offices essential to the new territories. In Ashanti, successfully, and in Oyo, unsuccessfully, the king tried to gain in these territories the right of appointment that he lacked in the metropolitan districts.

Variety of chiefdoms

In many ways the small chiefdoms of the Guinea coast had political characteristics similar to those of the kingdoms. Even the very small groups of the Cross River area, for example, surrounded their priest-chiefs with ritual similar to that of the Oyo. On the other hand, some small chiefdoms had completely different political systems. Mende chiefs, for instance, were essentially secular leaders. The chiefs or kings in the Niger Delta towns were hardly more than ritual figureheads, for power lay openly with wealthy traders. In many Ibo groups there were few formal political roles at all, and decisions were often taken through public discussions at village meetings.

Kin groups and other associations. In most tribal societies a man's household is the main unit of production, and consequently a husband and wife or a man and his junior kin are bound by strong economic bonds. Understandably, however, the more wealthy the community is, the greater is the value of inheritance; the heavier the population is, the more important is the position of those who control land; and the more farmers are involved in the market, the more commercial considerations will influence ties between kinsfolk.

Husband–wife relations

In the Guinea coast areas the marketing of crops is intimately involved in the rights and obligations of husband and wife and tends to add monetary considerations to domestic arrangements. Among the Cross River Mbembe, for example, the husband has the right to sell yams that his wife has helped to grow, but she is traditionally regarded as justified in seeking divorce if her husband spends the acquired money irresponsibly—especially if he uses it to make marriage gifts for another wife. Among the Nupe and Yoruba the men work the farms and the women market the produce. A husband must freely give his wife the staple crops for sale or home consumption; other crops, however, she must buy from him. She, in turn, is obliged to cook without charge her husband's staple foods but not his delicacies, which he must buy if he wants them. Even in the modern urban environment the tradition of financial independence between husband and wife continues. One of the problems associated with slum clearance in the Yoruba city of Lagos has been that wives, moved out to suburbs, can no longer earn by trading in the central markets, and this has led to matrimonial stress.

Age-sets

In most tribal societies age is an important basis for group formation, and in some there are what are called "age-sets"—compulsory groupings of individuals of comparable age who advance through life together. Age-sets had greatest political power in East Africa but they were significant on the Guinea coast in either of two circumstances: (1) wherever there were highly competitive kin groups, there tended to be age-sets as a kind of compensating factor to link non-kin together; (2) wherever, in highly competitive societies, kinsmen were wont to quarrel over inheritance, there were age-sets to provide the individual with loyal partisan friends. Mbembe villages, for example, are deeply divided by kin-group competition, but age-sets unite everyone of the same age throughout the village. Girls early form strong ties with one another, because each age-set of girls collectively undergoes public rites and ceremonies before its members are allowed to marry. This produces such collective solidarity that recently teen-age sets have acted like trade unions, stipulating the minimum personal gifts necessary before any member will accept a suitor. Relationships with age-mates remain important throughout a woman's life, especially in matrimonial quarrels. Boys also form very close ties with their age-mates and from their mid-teens onward can be called to work collectively on tasks that are necessary for the village as a whole. A man in need of financial help may meet with selfish refusal from his kinfolk, but his age-mates will generally always help him; indeed, in all the crises that an individual faces, especially in sickness and death, his age-mates are there to support him. Age-mate links are of enormous value when kin-group elders meet to deal with intergroup arguments, because the heads of the disputing parties often have the closest personal links with one another.

Secret societies

One of the most characteristic of Guinea coast institutions, especially in areas in which central government was weakly developed, was the so-called secret society. Such societies had a significance similar to that of age-sets since they cut across kin-group lines and united men (and sometimes women) of different groups. The influence of secret societies was even greater, because membership often cut across village and sometimes even tribal boundaries. Moreover, the fact that members were often of different grades and that membership in higher grades was open to those who could pay the fees meant that these societies provided the means whereby successful individuals could exercise wide, if covert, influence. In two major areas—in Sierra Leone and Liberia, and in the area east of the Niger Delta—these associations exercised real political power and were crucial to the traditional systems of law and order. Among the tribes in the former area, the men's Poro Society and women's Sande Society were primarily responsible for punishing such serious offenses as incest and homicide; moreover, if the death sentence were imposed, it was imposed and executed by a masked Poro member. There were local Poro councils composed of members of the highest grade, and a chief's authority often rested on his Poro rank. Among the Gola, Kpelle, and Mano of Liberia and the Mende of Sierra Leone, Poro forged links between autonomous chiefdoms; in 1898 the Poro of the Mende even organized a general uprising to try to oppose British expansion in Sierra Leone.

Similarly, though on a smaller scale, among the Yako and Mbembe in the Cross River area social control was largely vested in important secret societies. The most interesting development took place in 19th-century Calabar, where the Ekpe Society became the main instrument of the trading oligarchy. In the absence of a strong government Ekpe members ensured that commercial debts were honoured and united members and freemen in countering any signs of rebellion among the many slaves owned by the rich traders of Calabar. Even among the Ibo there were rich men's societies, whose leaders exerted considerable influence on village life. Thus, in general, secret societies were institutions for translating slight advantages of wealth into political influence. Moreover, for Africa, they reflected an unusual measure of social stratification.

Social stratification. Stratification among freemen seems to have been greatest among the most wealthy and centralized states: the Nupé had a clear division between commoners and aristocrats, and the tendency was incipient in Ashanti, Dahomey, and Benin. Among the Yoruba there was a kind of stratification through occupational craft inheritance, but deliberate efforts were made to prevent the emergence of a royal aristocracy. Slavery was an element in stratification mostly in Ashanti, Dahomey, Yoruba, and Calabar, where slaves were used in the commercial production of crops.

Freed slaves from North America

In the 19th century new forms of stratification emerged in Sierra Leone and Liberia when freed slaves educated in North America and more receptive to missionary education settled there to become shopkeepers and white-collar workers—an elite vis-à-vis the natives. Some Sierra Leonians moved to other British West African possessions on the coast, where they joined with tiny indigenous elites drawn from wealthy, educated coastal families to form with them a new bureaucratic class.

Belief systems. The similarity between the various belief systems on the Guinea coast demonstrates the great measure of contact between the various tribes and peoples. Most systems contain these features: belief in a withdrawn high god; belief in lesser gods of lesser power but useful because easily manipulated; reverence for the dead, usually ancestors, who exercise influence over the groups to which they belonged in life; belief in witches and sorcerers, whose existence explains undeserved misfortune; and, finally, the existence of diviners who can determine the cause of a particular misfortune. The pantheon of gods tends to be more hierarchical and complex in kingdoms such as Dahomey that have more social stratification and differentiation. There are countless cults, some of the same name, occurring over wide areas, some being adopted ad hoc to guard against some temporary misfortune. Distant cults are often deemed more powerful than local ones, so that priests may be brought long distances to establish local shrines to protect against misfortune.

Metaphysical explanations of individual success

Most significantly, many of these belief systems make special provision for explaining the success of an outstanding individual. The Kalabari of the Niger Delta believe in a special supernatural being whose activities aid in individual success. Among the Cross River societies there is a belief in a kind of sorcerer who achieves individual riches and influence by sacrificing, in a supernatural or spiritual sense, the lives of his junior kinsmen; an individual identified as such a sorcerer is not abhorred. Very widespread from the Niger Delta to Ghana is the concept of prenatal fatalism; individuals choose before birth success or failure in life, and this accounts for individual variation in wealth, as well as in fertility and health.

PROBLEMS OF MODERNIZATION

When the new states of the Guinea coast are considered in the context of the modern world, it is important to be aware that the several peoples in the area are culturally more distinct from one another than, for example, are the peoples of different European nations. This cultural diversity poses a real threat to national cohesion. Traditional social cohesion depended on the ritual authority of leaders, on the active political interests of kin groups, and on the influence of age-sets and secret societies—not on the administrative authorities governing within colonially determined boundaries. Although missionaries and colonial powers did tend to undermine traditional authorities and although national party politics today are no longer based on the old groups, old loyalties remain remarkably intact beneath the surface of modern politics. The Nigerian civil war of the 1960s showed that. Indeed, as long as most of the people remain primarily subsistence farmers, much of the old tribal order will remain.

On the other hand, in such major centres as Freetown and Ibadan, tribalism is simply a veneer of traditionalism that masks radically new associations, as rural migrants seek to cope with the problems of city life by banding together. In all the new states, moreover, education has developed new elites. Except in Liberia, where the Afro-American minority hangs on to its elite status, it is the indigenous educated people who are in control in all walks of life. The emergence of this elite is one factor that makes for the growth of a sense of national unity across the old tribal boundaries. It is a development, however, that brings social problems in its wake. There are now vast differences in income and power between many rural farmers and the politically—and economically—successful minority. One of the greatest political problems is that the man who achieves success through politics alone, without other claims to elite status, clings to power at all costs, because, should he fall, he loses economically so much.

Sometimes those who write about tribal societies suggest that most of their modern ills spring from a new selfishness and individualism engendered by modernization. A study of the Guinea coast societies suggests, on the contrary, that, although modernization may give the selfish individual more room to manoeuvre, it has not created him. These societies have undergone much social change over past centuries, and some individuals always have seized the advantage of the moment. Change today is perhaps more rapid, and the differences in wealth between individuals is perhaps greater, but it is doubtful if modern social problems arise from any new development of self-interest.

BIBLIOGRAPHY. The literature on West Africa is very considerable, and much of it is in French because much of the area was formerly under the control of France. A most valuable source of French material is the *Bulletin I.F.A.N.* (Institute français d'Afrique noire). Two other journals that are important general sources of information are *Africa* and the *Journal of African History* (both issued quarterly). For an understanding of the area some standard geographical text, such as WALTER FITZGERALD, *Africa*, 10th ed. rev. (1967), is useful. On the prehistory of West Africa, see the review article on African prehistory in the *Biennial Review of Anthropology* (1965). On the general history of West Africa, a most valuable book is the collection of essays by different authors edited by J.F.A. AJAYI and MICHAEL CROWDER, *History of West Africa*, 2 vol. (1972). JOHN R. GOODY, *Technology, Tradition, and the State in Africa* (1971), is interesting in its attempts to account for the differences in state organization between the forest and the savanna regions; a very important book on kingdoms in both areas is that edited by C. DARYLL FORDE and P.M. KABERRY, *West African Kingdoms in the Nineteenth Century* (1967). There are many ethnographic studies of Guinea Coast peoples. Particularly relevant to this article are the following: on the Mbembe, ROSEMARY L. HARRIS, *The Political Organization of the Mbembe* (1965); on the Ibo, M.M. GREEN, *Ibo Village Affairs* (1947); on the Nupé of the Niger–Benue confluence, S.F. NADEL, *A Black Byzantium* (1942); on the Yoruba, P.C. LLOYD, *Yoruba Land Law* (1962); on the Akan states, K.A. BUSIA, *The Position of the Chief in the Modern Political System of Ashanti* (1951, reprinted 1968); on Dahomey, MELVILLE J. HERSKOVITS, *Dahomey: An Ancient West African Kingdom*, 2 vol. (1938, reprinted 1967); on the Mende, KENNETH L. LITTLE, *The Mende of Sierra Leone* (1951). Most ethnographic accounts give some information on religious beliefs, but particularly recommended are the relevant articles on the Mende and the Ashanti in C. DARYLL FORDE (ed.), *African Worlds* (1954); and an unusual book dealing mainly with peoples in Ghana is MEYER FORTES, *Oedipus and Job in West African Religion* (1959). There are a number of useful books that examine the modern political problems of Guinea Coast states against the social background of their peoples. Particularly interesting are DAVID E. APTER, *Ghana in Transition*, rev. ed. (1963); and J.G. LIEBENOW, *Liberia: The Evolution of Privilege* (1969). For information on the urban situation, the following are recommended: MICHAEL P. BANTON, *West African City: A Study of Tribal Life in Freetown* (1957); HILDA KUPER (ed.), *Urbanization and Migration in West Africa* (1965); and the introduction and relevant chapters in P.C. LLOYD (ed.), *The New Elites of Tropical Africa* (1966).

(R.L.Ha.)

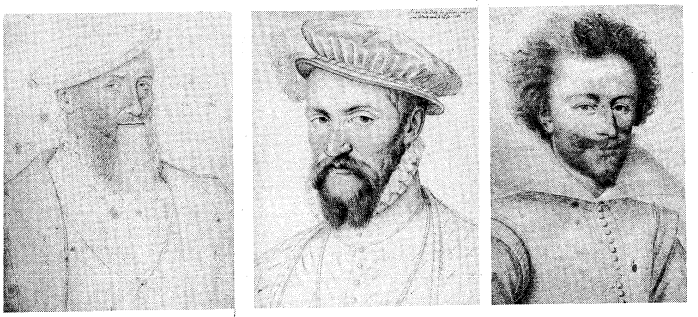
Guise, House of

The House of Guise played a dominant role in France during the 16th-century Wars of Religion, when it stood for Roman Catholic intransigence against the Calvinist Huguenots. It was a branch of the House of Lorraine, a ducal dynasty of royal descent ruling territory external to

France; thus its members ranked in France as "foreign princes" and so claimed precedence over the purely French nobility.

The founder of the Guises' fortunes was Claude de Lorraine (1496–1550; see Table), who inherited a number of fiefs in France, including the countship of Guise (in the modern *département* of Aisne), and served so well in France's wars that King Francis I made him duc de Guise and a peer of the realm in 1528. His brother Jean (1498–1550), cardinal de Lorraine, further promoted the family's prestige in France by political and diplomatic services.

By courtesy of (centre) the Bibliotheque Nationale, Paris; (left and right) Giraudon



(Left) Claude de Lorraine, 1st duc de Guise, *c.* 1547. In the Musée Condé, Chantilly, France. (Centre) François de Lorraine, 2nd duc de Guise, *c.* 1550. In the Bibliothèque Nationale, Paris. (Right) Henri de Lorraine, 3rd duc de Guise, *c.* 1585. In the Musée Condé, Chantilly, France. Portraits by the School of Clouet.

François de Lorraine (1519–63), Claude's eldest son, already high in King Henry II's favour, became duc d'Aumale in 1547 before he succeeded his father as the 2nd duc de Guise three years later. He then ceded Aumale to his brother, Claude. His military successes—the defense of Metz against the Germans in 1552 and the capture of Calais from the English in 1558—greatly enhanced his reputation. His niece Mary Stuart, heiress to the Scottish crown, was married to Henry II's eldest son in 1558; and when the latter ascended the throne as Francis II, in 1559, the Duc de Guise and his brother Charles (1524–74), known first as the cardinal de Guise, then as the cardinal de Lorraine, became the most influential men in France. They completely overshadowed their erstwhile rivals, the constable of France, Anne de Montmorency and the princes of the House of Bourbon, a cadet branch of a former ruling house, the Capetians.

The Conspiracy of Amboise

In 1560 an attempt by the Huguenots, with Bourbon connivance, at a coup d'etat against the Guises—the so-called Conspiracy of Amboise—was savagely put down.

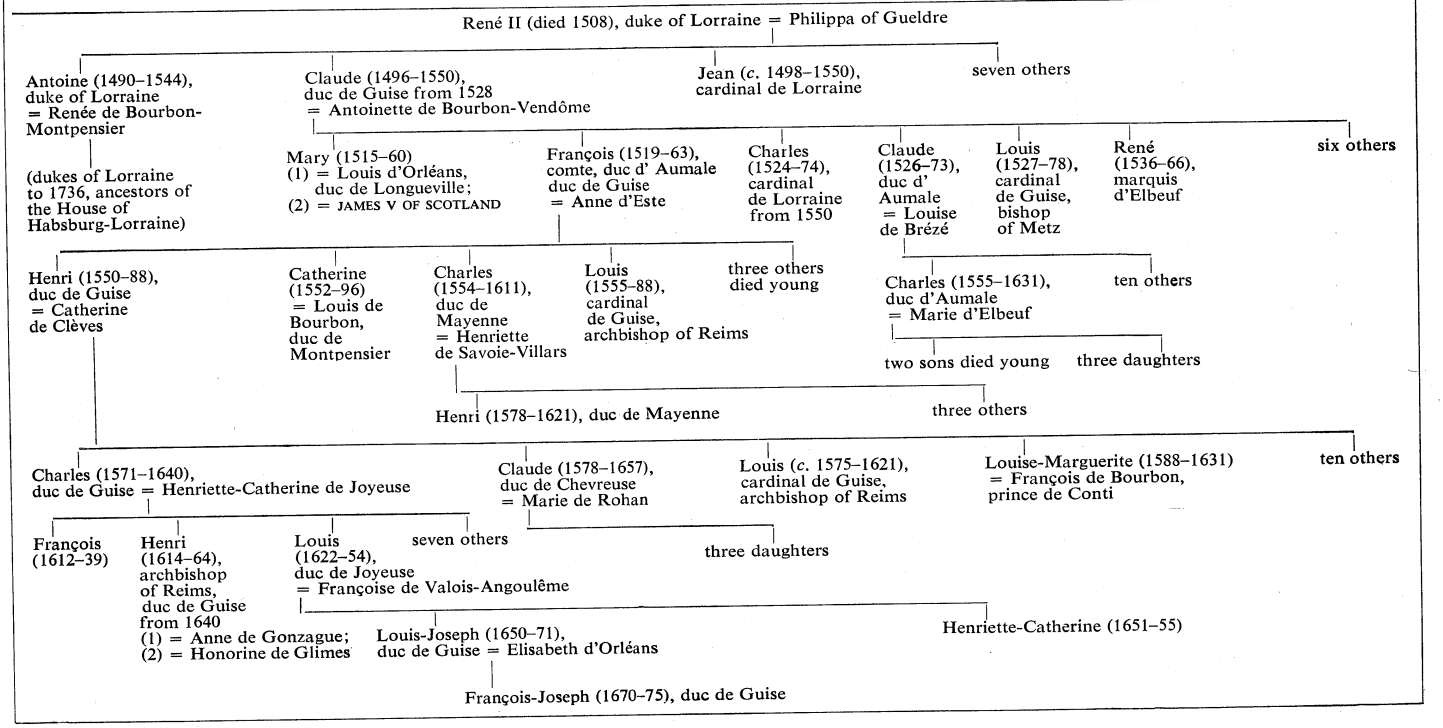
Francis II's sudden death in the same year led to a realignment of forces. Catherine de Médicis, regent for the new king, her son Charles IX, decided to try a policy of religious toleration and brought both Montmorency and the Bourbons back into the government. Guise, however, in 1561 reconciled himself with Montmorency in order to form a Catholic front against the Huguenots and their Bourbon sponsors. His attack on a Huguenot congregation at Vassy in 1562 led to France's Wars of Religion. The Battle of Dreux, where his intervention ensured the Huguenot defeat, returned virtual control of the government to him; but in February 1563, besieging Orléans, he was mortally wounded by a Huguenot assassin.

Henri de Lorraine (1550–88) succeeded his father as 3rd duc de Guise. Though he did not play a great role in the second of the Wars of Religion (1567–68), his conspicuous bravery in the third (1568–70) made him the idol of the young nobility at court; and after that war he began to ingratiate himself also with the people of Paris. Meanwhile, he was biding his time for vengeance on the Huguenot leader Gaspard de Coligny, whom he regarded as the instigator of his father's assassination.

In the summer of 1572, when Coligny's growing influence over Charles IX was drawing France toward open war against Spain, Catherine de Médicis seems to have appealed to the Guises to eliminate him. At that moment many Huguenots had come to Paris for the marriage of the Bourbon king of Navarre (later Henry IV of France); and when the attempt on Coligny's life miscarried, Catherine decided that it must not only be repeated with success but also accompanied by a general slaughter of his infuriated supporters. In the Massacre of Saint Bartholomew's Day (August 23–24), Guise supervised the killing of Coligny; but he sheltered 100 other Huguenots in his house before leaving Paris for a campaign in the provinces (fourth of the Wars of Religion, 1572–73).

Charles IX's brother became king of France, as Henry

House of Guise: Lines of Guise, Mayenne, and Aumale



III, in 1574. He married a princess of Lorraine, one of Guise's cousins; and in October 1575, during the fifth of the Wars of Religion (1574–76), Guise saved Paris for him by defeating the Huguenots' German mercenaries at Dormans. The peace terms of 1576, however, appeared to Guise as a betrayal of Catholic interests; and so, to promote the latter, he formed a Holy League of Catholic nobles—an almost overt challenge to the King's authority. The King countered by declaring himself head of the league and by denouncing the peace; but at the end of the sixth of the Wars of Religion (1577) he dissolved the league and again granted a measure of toleration to the Huguenots. The Guises in resentment strengthened their links with Spain, the one reliably Catholic power.

Formation of the Holy League

When Henry III's last brother died in 1584, Henry of Navarre became heir presumptive to the French throne. To exclude this Huguenot from the royal succession, the Guises promptly revived the league, with firm support from Spain. Henry III submitted to their dictation (1585), and religious war broke out once more. Guise won two notable victories over the Huguenots in 1587; but Henry III was now more anxious to be rid of Guise's tyranny than to consolidate it by helping to destroy its opponents, Henry of Navarre and the Huguenots. On the Day of the Barricades (May 12, 1588), when the Parisians rose in favour of Guise against Henry, Guise might easily have deposed the King: instead he pacified the insurgents and Henry was able to escape. The Edict of Union (July 1588) was a further surrender by the King to the league; and in the following month Guise was nominated lieutenant general of the realm. The States-General, in session at Blois from October, were dominated by the league until, on December 23, 1588, Guise was suddenly called to the King's apartments: on entering the antechamber he was stabbed to death by the bodyguards. His brother Louis, cardinal de Guise, was murdered the following day.

The 3rd Duc's sister Catherine, duchesse de Montpensier, was active in exciting the Parisians against Henry III and Henry IV; her brother Charles, duc de Mayenne, continued the military struggle against Henry IV until 1595; and their cousin Charles, duc d'Aumale, was energetic in the same cause until 1594. The family remained prominent in the 17th century but ceased to play any dominant role.

BIBLIOGRAPHY. The only reasonably complete accounts of the House of Guise are RENE DE BOUILLE, *Histoire des ducs de Guise*, 4 vol. (1849–50); and HENRI FORNERON, *Les Ducs de Guise et leur époque*, 2nd ed., 2 vol. (1893). AUGUSTE BAILLY, *Henri le Balafré, duc de Guise* (1953), is a popular biography of the 3rd Duc. For a useful up-to-date survey of the background, see GEORGES LIVET, *Les Guerres de religion (1559–1598)*, 2nd ed. (1966). Relevant sources are indicated in HENRI HAUSER, *Les Sources de l'histoire de France au XVIe siècle (1494–1610)*, 4 vol. (1906–15).

(R.J.K.)

Gujarāt

One of the 21 states of the Indian Union, Gujarāt lies on the west coast of India. It is bounded by the Arabian Sea to the west, by Pakistan to the northwest, the Indian states of Rājasthān to the north, Madhya Pradesh to the east, and Mahārāshtra to the south. Its coastline is 992 miles long, and no part of the state is more than 100 miles from the sea. Its area is 72,236 square miles (187,091 square kilometres), and its population at the 1971 census was about 26,700,000.

The capital of the state is Gandhinagar on the northern outskirts of Ahmadābād (Ahmedabad), the former capital, largest city in the state, and one of the greatest cotton-textile centres in India. It was here that Mahatma Gandhi built his Sābarmati *āśrama* ("retreat," or "hermitage") as a headquarters for his campaigns.

Gujarāt derived its name from the Gurjara (supposedly a subtribe of the Huns), who ruled the area during the 8th and 9th centuries AD. The state in its present form came into existence in 1960 when the former Bombay state was divided between Mahārāshtra and Gujarāt on a language basis.

Though most of the people are engaged in agriculture, there is a cohesive and comparatively prosperous merchant community that thrives on the trade and commerce generated in Gujarāt because of its long coastline. For general historical background, see INDIAN SUBCONTINENT, HISTORY OF THE.

History. Human settlements have been traced back to the Stone Age period in the valleys of the Sābarmati and Mahi rivers in the eastern part of the state. The historic period is linked with the spread of the Harappan (Indus Valley) civilization, which flourished in the 3rd and 2nd millennia BC. Centres of this civilization have been found at Lothal, Rangpur, Amrī, Lakhabaval, and Rozdi (mostly in the Saurāshtra [Kāthiāwār] Peninsula).

The Mauryan period

The known history of Gujarāt began with the extension over the area of the rule of the Mauryan dynasty, as is evidenced by the edicts of the emperor Aśoka (*c.* 250 BC), carved on the Girnār Rocks in the Saurāshtra Peninsula. After the fall of the Mauryan Empire, Gujarāt came under the rule of the Śakas (Scythians), or western Kṣatrapas (AD 130–390). The greatest of these, Mahākṣatrapa Rudradāman, established his sway over Mālwa, Saurāṣṭra (Saurāshtra), Kutch, and Rājasthān.

During the 4th and 5th centuries, Gujarāt formed a part of the Gupta Empire, until the Guptas were succeeded by the Maitraka dynasty of the kingdom of Valabhī, which ruled over Gujarāt and Mālwa for three centuries. The capital Valabhīpura (near the eastern coast of the Saurāshtra Peninsula) was a great centre of Buddhist, Vedic, and Jain learning. The Maitraka dynasty was succeeded by the Gurjara-Pratihāras (the imperial Gurjaras of Kannauj), who ruled during the 8th and 9th centuries; they, in turn, were followed shortly afterward by the Solaṅki dynasty. The boundaries of Gujarāt reached their furthest limits in the time of the dynasty, when remarkable progress was made in the economic and cultural fields. Siddharāja Jayasiṃha and Kumārapāla were the best known Solaṅki kings; the famous writer Hemacandra flourished during this period (12th century). Karṇadeva Vāghelā, of the Vāghelā dynasty that followed, was defeated in about 1298 by 'Alā'-ud-Dīn Khaljī, sultan of Delhi; Gujarāt then came under Muslim rule. It was Aḥmad I, the first independent sultan of Gujarāt, who founded Ahmadābād (1411). The end of the 16th century saw Gujarāt under Mughal rule; this lasted until the mid-18th century, when the Marāthās overran the state.

The founding of Ahmadābād

Gujarāt came under the administration of the British East India Company in 1818. After the Indian rebellion of 1857, the area became a province of the British crown and was divided into Gujarāt province, with an area of 10,000 square miles, and numerous native states. With Indian independence in 1947, all of Gujarāt except the states of Kutch and Saurāshtra was included in Bombay state; the province was enlarged in 1956 to include the two states. In 1960 Bombay state was split into present-day Gujarāt and Mahārāshtra states.

In April 1965 fighting broke out between India and Pakistan in the Rann of Kutch, an area that had long been in dispute between them. A ceasefire came into force on July 1, and the dispute was submitted to arbitration by an international tribunal. The tribunal's award published in 1968 gave nine-tenths of the territory to India and one-tenth to Pakistan.

The landscape. *Physical geography.* The present-day state of Gujarāt is one of great contrasts; it stretches from the wet, fertile, rice-growing plains of the west coast, north of Bombay city, to the almost rainless salt deserts of Kutch. It is easiest to treat the state on the basis of its natural divisions. In the northwest, Kutch comprises a single district so arid as to be almost desert; it is bounded on the south by the Gulf of Kutch and on the north and east is separated from Pakistan and the mainland of India by the Rann of Kutch, best described as a vast salt marsh covering about 8,000 square miles. The Rann floods during the rainy season, slight though the rains may be, and Kutch is converted into an island; in the dry season it is a sandy, salty plain, plagued by duststorms. To the south of Kutch is the large peninsula of Saurāshtra (Kāthiāwār), lying between the Gulf of Kutch and the Gulf of Cambay. It also is arid and rises

The Rann of Kutch

from the coasts to a low, rolling area of hill land in the centre, covered with scrub or sparse woodland. The chief towns are found in the more fertile spots and were formerly the capitals of small states. Soils are mostly poor, having been derived from a variety of old crystalline rocks, but among the state's valuable products are the fine building stones of Porbandar. Rivers, except for seasonal streams, are absent. On the southern shores of the peninsula is the former Portuguese territory of Diu. Northeastern Gujarāt is mainly a country of small plains and low hills, through which runs the main line of a railway from Bombay to Delhi. Rainfall is low; January temperatures may drop almost to freezing point, while a temperature of 118° F (48° C) has been recorded in the hot season. Crops grown are millet and some cotton.

Southward in central Gujarāt the rainfall increases; temperature ranges are less; and soils are more fertile, being derived partly from the basalts of the Deccan region. The focus of this area is the city of Baroda, formerly the capital of a rich and powerful state. South of what is now the Baroda district, the important river, the Narmada, empties into the Gulf of Cambay, and it is the silt deposited by this river and the Tāpti that is responsible for the shallowness of the Gulf of Cambay and the decline of its former ports.

Southern Gujarāt, the districts of Broach and Surat, are famed for their rich soils and fine crops of cotton. The great Tāpti River, flowing in a deep trench from the east, cuts through Surat; and in the eastern parts of south Gujarāt the country is mountainous. This is, indeed, the northern extension of the Western Ghāts, which attract a heavy rainfall from the rain-bearing summer monsoon winds. Farther south, the mountains are forested. The small district of the Dangs is in this area. Along the coastal plains conditions begin to approach an equable climate, with rainfall nearing 80 inches.

Vegetation. Rainfall and altitude variations are such that the vegetation changes from north to south and from the east highlands to the west coast. With less than 25 inches of rainfall, the species in the scrub forests are the babul acacia, the caper, the Indian jujube, and the toothbrush tree (*Salvadora persica*). Where rainfall is up to 40 inches the following deciduous species are found: teak, catechu or cutch, bakligum, axlewood, and Bengal kino or butea gum. These are found in the tablelands of Saurāshtra and the east mainland area. Most of the deciduous forests are concentrated in areas with more than 40 inches of rain in the south and east. They produce valuable timbers: wooly tomentosa, Vengai padauk (resembling mahogany), Malabar simal, and the heartleaf adina. The west coast of Saurāshtra is known for its algae, and the east coast produces the paper-making plant *Cyperus papyrus* (a sedge known as the paper reed or paper rush).

Animal life. The forests of Gīr in Saurāshtra contain the last Indian lions, the only remaining numbers of the Asiatic species; in the Little Rann of Kutch the only surviving Indian wild asses are found.

The peacock is common throughout the state, and there is a bird sanctuary at Lake Nalsarovar, near Ahmadābād, which attracts about 140 species of birds migrating from the Siberian plains and elsewhere in winter. These include the sarus crane, the Brahmini duck, the bustard (a game bird), the pelican, the cormorant, the ibis (a wading bird related to the heron), the stork, the heron, and the egret.

There is excellent offshore and inland fishing in Gujarāt. Catches include pomfret (a food fish), salmon, hilsa (a type of shad), jewfish (scianid fish), prawn, Bombay duck, and tuna.

Population. Ethnically, there are two dominant racial strains in Gujarāt: a brachycephalic (short-headed) leptorhinian (straight-nosed) element, represented by the Parsi and the Nāgar Braham, Bhāṭiā, Bhādela, Rabāri, and Miana castes; and a dolichocephalic (long-headed) mesorrhinian (broad-nosed) element, represented by the aboriginal Bhīl and the Bhaṅgī, Kolī, Dubla, Naikda, and Mācchi-Khārwā tribes. The rest of the population are in intermediate groups; Gujarāt as a whole is a mesocephalous region (*i.e.*, one in which heads are approximately midway between brachycephalic and dolichocephalic). Members of the former untouchable castes (about 7 percent) and of the aboriginal tribes (about 13 percent) formed nearly one-fifth of the state's population at the time of the 1961 census. Of India's tribal population, about 10 percent lived at that time in Gujarāt, with the highest accumulation in the entirely tribal district of Dangs. Ahmadābād district had the highest proportion of scheduled (untouchable) castes.

The Gujarati language

The Gujarati language belongs to the Indo-Aryan family of languages and is derived from Sanskrit through Prākrit, ancient Indic languages other than Sanskrit, and Apabhraṃśa, a language spoken in northern and western India from the 10th to the 14th century. Gujarāt's contact by sea with foreign countries has also led to the introduction of Persian, Arabic, Turkish, Portuguese, and English words. The prodigious writings in Gujarati of Mahatma Gandhi (*q.v.*; architect of India's freedom and apostle of the doctrine of nonviolence), which are noted for their vigour and simplicity, have had a strong influence on modern Gujarati prose.

Hinduism was the religion of almost 90 percent of the population at the time of the 1961 census, whereas 8 percent adhered to Islām and 2 percent to Jainism. The policy of the state has always been marked by the religious tolerance of its rulers.

The density of population in the state was 369 per square mile in 1971, as compared to 434 per square mile for India as a whole. Of the total population, about 28 percent was urban, as compared to 20 percent for India as a whole. The most highly urbanized part of the state was the Ahmadābād-Baroda industrial belt. More than half the urban population lived in 24 of the 217 towns in the state. The major industrial towns are Ahmadābād (population, 1971 census, 1,600,000), Baroda (470,000), and Surat (470,000). Major towns that were once capitals of princely states are Rājkot (300,000), Bhavnagar (Bhaunagar) (230,000), and Jāmnagar (215,000).

Administration. A governor, appointed by the president of India, is the head of the government of Gujarāt. A council of ministers, led by the chief minister, aids and advises the governor in the exercise of his functions. The Legislative Assembly, an elected unicameral legislature, consists of 168 members.

Justice. The High Court, with 14 judges, including the chief justice, is the highest judicial authority in the state. There are also city courts, the courts of district and sessions judges, and courts of civil judges in each district.

Local government. The state is divided into 19 districts, the revenue and general administration of each being looked after by the district collector, who also functions as the district magistrate for the maintenance of law and order. The 19 districts are: Ahmadābād; Baroda; Banās Kāntha; Broach; Mehsāna; Kaira; Pānch Mahāls; the Dangs; Surat; Sābar Kāntha; Jāmnagar; Bhavnagar; Amreli; Junāgadh; Kutch; Rājkot; Surendranagar; Bulsār; and Gandhinagar.

The *pañcāyat*s

With a view to associating the people living in the remotest villages with the administration, rule by *pañcāyat* (local governing committee) was introduced in 1963. The scheme comprised (1) a *grām* (village) *pañcāyat* for each village or group of villages or a *nāgar* (town) *pañcāyat* for towns with a population of between 10,000 and 20,000; (2) a *taluka* (an administrative unit similar to a county) *pañcāyat* for each *taluka* (called a *tahsil* in some states); and (3) a district *pañcāyat* for each district. There were 182 *taluka pañcāyat*s in the early 1970s, 59 *nāgar pañcāyat*s, and almost 12,000 *grām pañcāyat*s covering about 19,000 villages. In addition, 2,000 *nyay* (judicial) *pañcāyat*s and 12,000 conciliation *pañca*s (committee members) had been constituted to administer justice in certain matters in rural areas. There were 59 municipalities and three municipal corporations (in Ahmadābād, Baroda, and Surat). These worked under the supervision of the director of municipalities.

Social conditions. *Health.* In the early 1970s health and medical services were still being extended to remote

and backward areas. There were programs for the eradication of malaria, tuberculosis, leprosy, and other communicable diseases as well as for improving supplies of drinking water and preventing food adulteration. Steps had also been taken to expand hospitals and medical colleges. There were 250 primary health centres, five hospitals, and a dental college.

Education. Primary schools for all children between seven and 11 had been opened in the early 1970s in nearly all villages with 500 or more inhabitants. Besides these, special schools were teaching arts and crafts. There were 2,000 secondary schools with 730,000 pupils, as well as seven universities and more than 250 institutions for higher education, with an enrollment of 125,000 students. Technical education was provided by seven engineering colleges, three junior technical, and 42 technical schools. Research institutions included the Physical Research Laboratory, the Ahmadābād Textile Industry Research Association, the Bholabhai Jesinghbhai Research Institute for Oriental Studies, the Indian Institute of Management, the National Institute of Design, and the Sardar Patel Institute for Economic and Social Research. There were also institutes doing research in agriculture and fisheries.

Welfare. At the same time there were state institutions under the director of social welfare for the welfare of children, women, the physically handicapped, the aged, infirm, and destitute, for delinquents, beggars, orphans, and released prisoners. There is also a Backward Class Welfare Department for the education, economic uplift, health, and housing of the backward class, which constitutes about 20 percent of the population.

The role of agriculture

The economy. About two-thirds of the population were engaged in agriculture in the early 1970s, the gross crop area amounting to about 50 percent of the total land area. Wheat and millets were the staple food crops, with rice production being concentrated in the wetter areas. Sugar production was increasing, while cotton, tobacco, and oil seeds (especially peanuts) were profitable cash crops. Cash crops were, indeed, a characteristic feature of the state's agricultural economy, slightly exceeding food crops in value.

Food requirements of the state were 5,740,000 tons and food production 3,950,000 tons. Lack of irrigation was a major handicap, only 10 percent of the crop area being irrigated, as compared with an all-India average of 21 percent. There had, nevertheless, been a distinct improvement in agricultural yields in recent years.

Gujarāt occupied third place, after Mahārāshtra and West Bengal, in the industrial economy of India. Output of soda ash amounted to 94 percent of the national production, salt about 55 percent, clocks about 40 percent, and cotton textiles about 30 percent. There were 116 textile mills, of which 70 were in Ahmadābād, "the Manchester of India." The pharmaceutical industry, concentrated at Baroda, Ahmadābād, and Atul (Bulsār), manufactured one-third of the national product. The steady growth of small industries has been significant. The Gandhian approach to labour problems—strict reliance on the truth, nonviolence, settlement by arbitration, minimal demands, and the use of the strike only as a last resort—has had a great impact in the field of industrial relations in Gujarāt, which has remained free from labour unrest.

Lacking in hydroelectric power, the state had a 250-megawatt thermal-power station using coal, natural gas, and oil built at Dhuvaran, the capacity of which would be doubled by 1973. From July 1969 the state was receiving power from the Tarapore nuclear station in Mahārāshtra. The total generating power in the state was due to increase to 1,452 megawatts by 1972.

Transport and communications. There were 3,500 miles of railways in the state in the early 1970s and 20,000 miles of roads. Road passenger transport had been nationalized and operated with about 2,600 vehicles and a staff of 24,000.

Transport by land was supplemented by water transport, particularly by coastal shipping. There were 45 ports besides the major international port of Kandla. Of these, 25 were open to foreign ships and 20 to coastal shipping. Six of these were all-weather ports.

Gujarāt had nine airports, and there were more than 6,000 post offices and 500 telegraph offices.

Cultural life. The folklore and folk culture of Gujarāt can be traced to the mythology of Lord Kṛṣṇa (Krishna; an incarnation of the god Viṣṇu, or Vishnu), as told in the *Purāṇas*, a class of Hindu sacred literature. The *rāsnṛtya* and *rāslīlā* dances in honour of Lord Kṛṣṇa have survived in the form of the popular folk dance, the *garabā*. This dance is performed at the *Navarātrī* festival, which honours the goddess Durgā, the dancers moving in a circle and singing and keeping time by clapping their hands or clashing together sticks called *daṇḍa*. A folk drama, the *bhavai*, has also survived.

Śaivism (Shivaism), the cult of the god Śiva (Shiva), characterized by ritual devotion, has long flourished in Gujarāt; so, too, has Vaiṣṇavism (Vishnuism, the worship of the god Viṣṇu), the devotional practices of which, with the emergence of the cult of *bhakti* ("devotion"), inspired its devotees to compose songs to be sung in temples or other religious gatherings. Vaiṣṇavism has encouraged the emergence of a number of saints, poets, and musicians, such as Narsiṃha, who composed *padas* (verses) in the 15th century; Mīrā Bāī, a 16th-century Rājput princess who renounced her royal home and composed *bhajanas* ("devotional songs"); Premanand, an 18th-century poet and writer; and Dayarām, an 18th-century composer of songs who popularized the *bhakti* cult. Jainism, with its nonviolence and vegetarianism, gained a stronger hold in Gujarāt than in any other part of India.

The architectural style of Gujarāt is well-known for its luxuriousness and perfection. The style is preserved in such monuments and temples as those at Somnāth, Modhera, Thān, Ghumli, Girnār, and Pālitāna. A happy synthesis of Hindu-Muslim culture has also been preserved in many ancient monuments. Gujarāt is famous, too, for some of its art and craft products. These include the *jari* (gold and silver embroidery) of Surat; the *bāndhaṇī* (a sari) of Jāmnagar; the *paṭolā*, a fine silk fabric of Pātan, woven with hereditary skill; the toys of Idar; the perfumes of Pālanpur; the hand-loomed products of Konodar; and the decorative woodwork of miniature temples and mythological figures at Ahmadābād and Surat.

The *mahājans*

Among the most durable and effective of the state's cultural institutions are the trade and craft guilds known as the *mahājans*. Often coterminous with castes and largely autonomous, the guilds have in the past solved disputes, acted as channels of philanthropy, and encouraged arts and culture.

BIBLIOGRAPHY

General: GUJARAT, INDIA, DIRECTORATE OF INFORMATION, *Gujarat: A Reference Annual*, a yearly survey with statistical details of various state activities.

History: H.D. SANKALIA, *Investigations in Prehistoric Archaeology of Gujarat* (1946), an account of the riverside settlements of prehistoric times; A.K. MAJUMDAR, *Chaulukyas of Gujarat* (1956), a history of the Solaṅki Vāghelā period, with a discussion of the administrative system and the social conditions of the time; M.S. COMMISSARIAT, *A History of Gujarat*, 2 vol. (1938–57), dealing with the history of the state from the Muslim sultanate to the break-up of the Mughal Empire, including an account of the socio-economic and cultural life.

Physical geography: K.R. DIKSHIT, *Geography of Gujarat* (1970), a brief outline of all relevant topics; W.T. SAXTON and L.J. SEDGWICK, "Plants of Northern Gujarat," *Rec. Bot. Surv. India*, vol. 6 (1918), pp. 207–323.

Economy: NATIONAL COUNCIL OF APPLIED ECONOMIC RESEARCH, *Techno-Economic Survey of Gujarat;* GUJARAT CHAMBER OF COMMERCE, *Gujarat at a Glance—Statistical Data.*

Population and language: GOVERNMENT OF INDIA, *Census of India, 1961*, vol. 5, pt. 9 (1965), a census atlas of Gujarāt, showing physical aspects and demographic trends; K.M. MUNSHI, *Gujarāt and Its Literature, from Early Times to 1852*, 3rd ed. (1967), tracing the evolution of Gujarati literature; T.N. DAVE, *Language of Gujarat* (1964), a broad outline of the evolution of Gujarati and Indo-European through Old and Middle Indo-Aryan down to old and modern Gujarati.

(D.N.P.)

Gulfs and Bays

Terminology and configuration

Any concavity of a coastline or re-entrant of the sea, regardless of size, depth, configuration, and geological structure, may be called a gulf or bay. Though the two terms are not strictly defined, the term gulf usually refers to large water bodies. Some large gulfs and bays preserve local names such as bight (Australia), channel (England), firth (Scotland), sound (U.S.), fjord (Norway), fjörd (Sweden), fjördhur or *floi* (Iceland), *zaliv* or *guba* (U.S.S.R.), ria or *río* (Iberian peninsula and South America), and *wan* (China and Japan). A number of pronounced concavities of oceanic margins have no proper name at all.

The problem of terminology extends to the difference between gulfs and seas. There are many small seas, such as the Sea of Marmara (0.011×10^6 square kilometres) and the Sea of Azov (0.038×10^6 sq km), which, strictly speaking, are really gulfs of the ocean or other seas (the Sea of Azov is a gulf of the Black Sea). The Gulf of Aden (about 0.270×10^6 sq km), another example, is part of the Arabian Sea and these water bodies have a common regime (similar tides, precipitation, evaporation, etc.). The narrow sound of Bab el-Mandeb connects the gulf with the vast Red Sea (0.438100×10^6 sq km) and exhibits a number of specific geomorphic features. The Red Sea, in turn, has two small gulfs to the north, namely, those of Suez and Aqaba.

The Bay of Bengal and the Arabian Sea are approximately the same size and have the same monsoonal water circulation. The Bay of Bengal is, in fact, the largest of the gulfs and bays, with a surface area of 2.172×10^6 sq km, volume of 5.616×10^6 cu km, and a length of 1,850 km (Table). The width of a gulf may exceed its length. The Great Australian Bight has the widest mouth (2,800 km). The Gulf of Guinea is deepest; its maximum depth (6,363 metres) exceeds that of the Bay of Bengal by more than 1,000 metres.

The shape and bottom topography of gulfs and bays are amazingly diverse. They depend on the geological structure and development of the region. Homogeneous bedrock of low strength or resistance results in simple shapes and shallow depths. The Gulf of Riga (at the Baltic Sea) is a possible example of the type. Long narrow arms with approximately parallel shores of the south Kara Sea extend inland for about 800 kilometres. They occupy troughs that originated by erosion during a period of lower sea level (Baidaratskaya Guba, Obskaya Guba with Tazovskaya Guba tributary, Yenisei Bay, Gydanskaya Guba). Deep, angular gulfs, on the other hand, are created along fractures, faults, and rifts (*e.g.*, Varanger Fjord); they usually have irregular bottom topography. Parallel fractures form very deep, narrow gulfs with parallel shores, such as the Gulf of California. Genuine fjord-gulfs are notable for their very high length to width ratios (up to 50:1). In regions that have undergone nonuniform deformation and uplift, gulfs and bays of complicated and irregular shape and bottom topography are consequently formed; the Gulf of St. Lawrence is an example.

Gulfs are connected with the sea by means of one or more straits. Sometimes there may be an archipelago in the mouth of the gulf as in the Gulf of Bothnia. There are some gulfs that open into the sea or into another gulf on opposite sides (Baffin Bay, Gulf of Aden, and the Gulf of Oman).

Single gulfs usually are formed along linear shores of the continent. If the shoreline is irregular and has a complex geological structure, groups of gulfs of a similar nature may occur. Most shorelines have small reentrants of various size that are called bays. These features are strongly influenced by local conditions, and they are not described or classified within the context of this article, which treats major water bodies of the world. For additional information on the dynamics of water within gulfs and bays, see WATER WAVES; TIDES; and OCEAN CURRENTS. For a further discussion of coastal morphology and shoreline features in general, see COASTAL FEATURES; RIVER DELTAS; and BEACHES. See also HOLOCENE EPOCH for a discussion of recent changes of sea level.

FACTORS THAT INFLUENCE THE NATURE OF GULFS AND BAYS

Climates, sills, and river runoff

Gulfs and bays may differ from the adjacent sea by virtue of water properties and dynamics and processes of sedimentation. Such differences are determined by the size and the shape of a given gulf, by the depth and bottom topography, and to a considerable extent by the degree of isolation from the ocean. Climatic conditions also are important. Isolation from an adjacent sea or ocean depends on the ratio of width of mouth to total surface area of a gulf or on the cross section of the mouth to total water volume. If there is a sill, the ratio of depth above the sill to the depth of the gulf is of great importance. No extensive comparisons of these ratios have been made to date; hence analysis of controlling variables must remain somewhat qualitative.

A high sill hampers the water exchange between an ocean and gulf and may lead to stagnation (oxygen deficiency) as is found in some fjords of Norway, in the Red Sea, and, particularly, in the Black Sea. Also, the presence of a sill causes independent circulation of gulf waters, created by local winds and the runoff of rivers. Sills are not indispensable for the formation of an independent circulation, however. A narrow mouth, as in the Gulf of Bothnia, leads to the same result.

In humid climates, the waters of gulfs are freshened by river runoff. Salinity is particularly low in the gulfs of the Baltic Sea and along the southern coast of the Kara Sea. Water becomes almost fresh in their heads, especially in the spring when snow begins to thaw. Gulfs of the arid zone suffer from intensive evaporation and receive little river runoff. Thus, salinity increases markedly in this climatic regime—up to 60 parts per thousand in the Persian Gulf and up to 350 parts per thousand in Kara-Bogaz-Gol (Caspian Sea). In addition to its effect on salinity, river runoff delivers organic matter and nutrient salts that may determine the specific features of life in the gulfs. The number of genera and species of organisms is small, but the organisms present tend to develop in quantities. That is why shrimp, oyster, and other fisheries are concentrated in many gulfs.

Funnel-shaped gulfs, in which the depth gradually decreases headward, usually have resonant tides. The tidal range at the head of such gulfs is several times greater than that in the open sea (Bristol Channel, Río de la Plata, Mezenskaya Guba, Zaliv Shelikhova, etc.). The world maximum tidal range has been registered in the Bay of Fundy (18.0 metres). The regularity (magnitude and frequency) of the flood tide may be distorted in such instances and the duration of the flood tide may become much shorter than that of the ebb tide. This may cause the phenomenon of bore, or mascaret, in which a steep wave will move rapidly upstream for dozens of kilometres.

Gulfs of simple shape with narrow mouth and a high degree of isolation from the ocean commonly are subject to seiches. These free oscillations can result from rapid changes of atmospheric pressure and, of course, from tectonic movements such as earthquakes. Seiches gradually decrease but some oscillation continues long after their cause disappears. A high rise of the water (storm surge) occurs in long and shallow gulfs if winds from the sea are protracted. Such phenomena are difficult to predict, and the high water levels may cause floods. Seiches often occur at the heads of Helgoländer Bay in the North Sea and in the Gulf of Finland.

Certain aspects of sedimentation are affected by the isolation of gulfs from the sea and river runoff. The rate of sediment accumulation in gulfs of limited area may be very high. This, of course, is a function of river discharge; sediment composition is usually similar to that of the load transported by entering rivers. Deposition of calcium carbonate often occurs in shallow gulfs in the arid zones where few if any perennial streams exist. The

Physical-Geographical Features of Some Gulfs and Bays*

names of gulfs and bays	surface area (10^6 sq km)	volume (10^6 cu km)	length in kilometres	width in kilometres		depth in metres			tidal range in metres	surface water temperature (Celsius)		surface salinity (parts per thousand)		river runoff
				max	mouth	max	mean	sill		max	min	max	min	
A_1 Group														
Gulf of Alaska	1.327	3.226	325	1,650	1,650	5,659	2,431	none	12.0†	12	<0	33	32	small
Bay of Bengal	2.172	5.616	1,850	1,720	1,720	5,258	2,586	none	10.7	27	25	34	18	large
Bay of Biscay	0.194	0.332	400	500	500	5,120	1,715	none	6.7	20	5	35.5	34	medium
Gulf of Guinea	1.533	4.592	540	1,900	1,900	6,363	2,996	none	2.7	27	25	35	31	large
A_2 Group														
Baffin Bay	0.689	0.593	>1,000	600	340‡	>2,300	861	466‡	4.2	5	<0	33.5	30	none
Gulf of Mexico	1.543	2.332	1,330	1,780	445	4,029	1,512	800	1.7	29	17	36.7	33	large
A_3 Group														
Gulf of Aden	...	...	900	335	335	3,328	...	none	2.9	>30	25	36.5	36	none
Gulf of California	0.177	0.132	1,200	200	200	3,660	813	none	5.2§	30	16	35.5	35	medium
Gulf of Oman	...	...	450	330	325	3,474	...	none	3.5	32	22	38	37	none
B Group														
Bay of Fundy	...	...	300‖	100	100	214	75	none	18.0	17	2	32	30	medium
Hudson Bay	0.819	0.092	1,560	1,140	190	274	112	none	7.9	14	<0	28.5	23	large
Río de la Plata	...	...	220	95	95	10	5–7	6	1.0	21	11	33	20	large
Gulf of St. Lawrence	0.238	0.030	¶	¶	¶	530	127	none	5.9	20	−1.8	32	26	large
C_1 Group														
Gulf of Aqaba	...	...	180	28	6	1,828	...	462	0.7♀	26	24	42	41	none
Sirt Gulf	...	...	200	450	450	1,627	...	none	0.3	27	14	38	38	none
C_2 Group														
Anadyrsky Zaliv	...	...	350	460	460	110	60–70	none	3.0	10	<0	30	28	medium
Persian Gulf	0.241	0.010	1,000	350	56	170	40	71	4.7	33	15	60♂	30	small
Gulf of Suez	...	...	325	58	58	82	40–60	none	1.8	28	23	43	41	none
Zaliv Shelikhova	...	...	750	300	190	495	100–150	none	12.9	14	<0	33	31	small
Gulf of Thailand	...	...	830	550	370	83	45.5	58	0.8□	31	27	32.5	30.5	large
D Group														
Gulf of Bothnia	0.117	...	668	240	155◇	294	21	none	0.6	14	0	5.5	1	medium
Gulf of Chihli	0.0827	0.0017	480▲	285	105	38	15–20	none	4.4	28	<0	31	22	large
Gulf of Carpentaria	0.4116	...	675	650	530	70	40–50	none	3.6	29	23	35.5	35	small
Mezenskaya Guba	...	...	105	97	97	31	10–20	none	10.0	16	<0	32	15	large
Obskaya Guba	...	...	800	90	60	18	10–12	7	0.7	14	0	15	1	large
Gulf of Finland	0.030	...	420	125	70	110	50–60	86	0.1	17	0	5	2	medium

*Data in this table may differ from those given elsewhere in set for some features because of differing definitions of geographical limits of each feature. Data adapted in part from Kossina (1921, 1933) and Lyman (1961). †Cook Inlet's head. ‡Davis Strait. §Up to ten metres at the Colorado River mouth. ‖Up to the Minas Bay head. ¶Not given because of the complicated outlines. ♀The value for the Red Sea's head. ♂In shallow parts along the south coast. □Up to 4 metres at the Mae Nam Chao Phraya. ◇South Quarken Strait—40 kilometres. ▲From the Laichow Wan head to Liao-tung Wan's head.

bottoms of long gulfs (or gulfs having sills) are usually covered with silt even at the shallowest depths (*e.g.*, Hudson Bay, the Gulf of Chihli, the inlets or *gubas* of the Kara Sea, the Gulf of Riga). Only strong tidal currents can prevent this siltation and, in some cases, cause the opposite phenomenon of bottom erosion. Currents maintain the existence of or actively deepen bottom troughs in narrow-mouthed gulfs whose depths are over 200 metres, whereas depths of adjacent parts of the open sea are only on the order of some dozens of metres.

Waves of the open sea either do not penetrate into comparatively isolated gulfs or, if they do, they become greatly reduced after entry. Small local waves that are related to gulf size prevail there. This tends to make gulfs quite navigable, and seaports and harbours have generally been situated on them.

CLASSIFICATION OF GULFS AND BAYS

The waters of gulfs and bays fill any depression in the earth's surface. Their geological structure and history of development are as varied as are those of the continents proper. The factors discussed above influence the morphological peculiarities of gulfs, and the latter, in turn, permit some general division or classification of these features to be made. The several groups in one possible scheme will be discussed here using typical gulfs and bays of each group as examples (Table).

Open concavities and isolated areas

Areas situated in open concavities of the continental coast (Gulf of Alaska, Bay of Biscay, Gulf of Guinea, Great Australian Bight, Bay of Bengal, Gulf of Tehuantepec etc.) are classified as the A_1 group. The depth of these gulfs in the region of the mouth usually is on the order of kilometres. The continental shelf and slope are generally pronounced. The general shape of such gulfs is simple; width of mouth usually exceeds its length. Water circulation and its physical properties are similar to those of the ocean. The character of the marine faunas does not differ from that of oceanic areas.

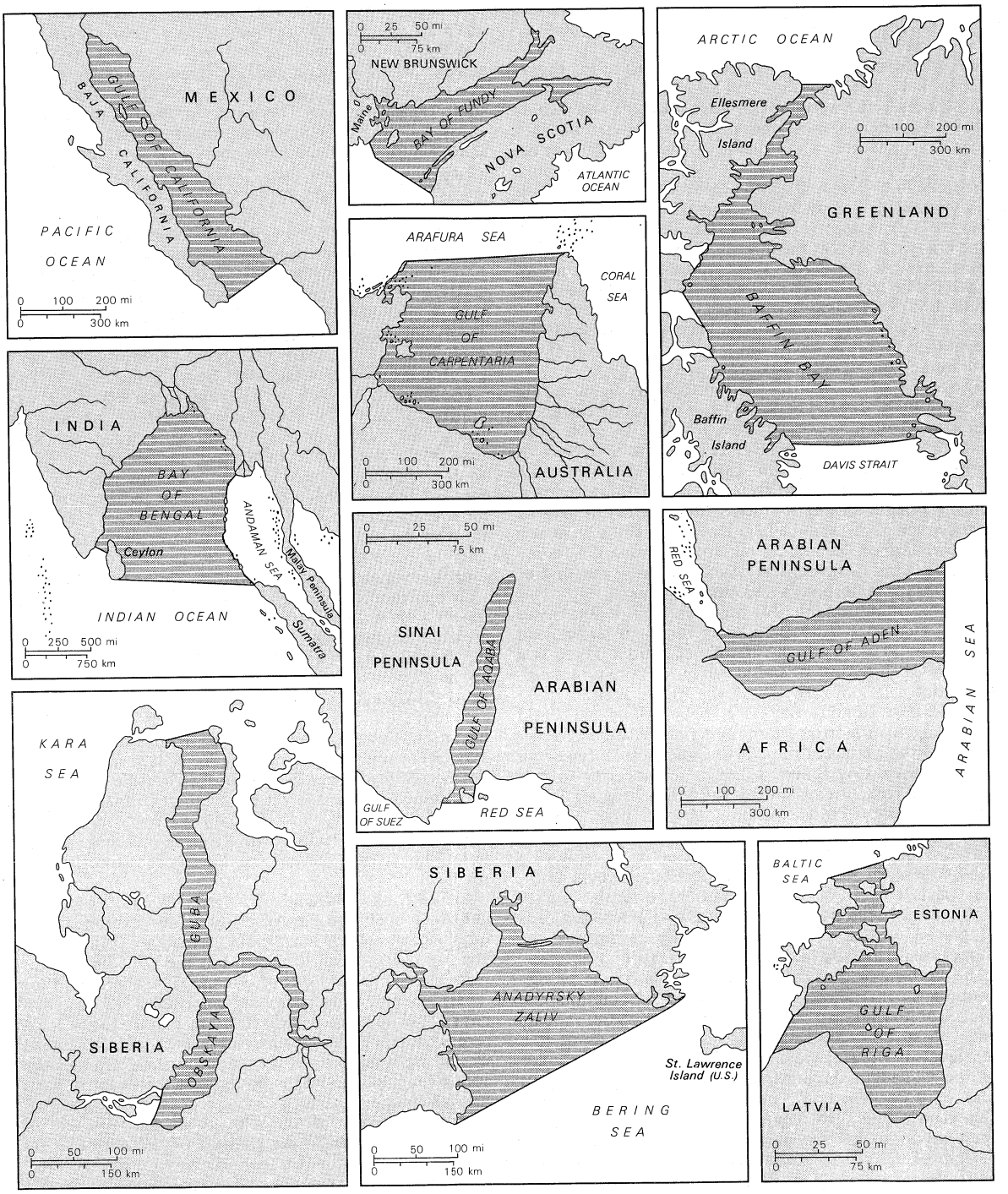
Large areas considerably isolated from the ocean, such as the Gulf of Mexico and Baffin Bay, are designated as group A_2 (Table). The former includes a geosynclinal hollow, founded in Mesozoic time and finally shaped during the Tertiary Period (2,500,000 to 65,000,000 years ago). It is connected with the ocean by the narrow and relatively shallow straits of Florida and Yucatan. Baffin Bay is a rift hollow that is connected by straits with the Atlantic and the Arctic oceans.

Ocean gulfs, such as the Gulfs of Oman, California, Aden, and some others have smaller areas and are isolated to a lesser degree. These gulfs and bays, in group A_3, have shapes that are determined by young faults and fractures. Depths in these gulfs generally exceed one kilometre. Unlike the previous group, in which gulfs might be of composite geological structure, these occupy areas that have undergone only a single episode of deformation.

Gulfs situated on the oceanic shelf, such as the Bay of Fundy, Hudson Bay, Río de la Plata, Golfo-St.-Mathias, and others, are in group B. The depth of such gulfs is up to 200 metres or more, and their configuration is determined by geological conditions. Because shelf areas repeatedly became dry land when the sea level fell during the ice ages, these gulfs received their final shape during the Pleistocene Epoch. The Gulf of St. Lawrence is included in this group (Table), though it is really intermediate between groups A_3 and B. It contains both a pronounced shelf and a long trough up to 530 metres deep.

Marginal and shelf areas

Gulfs of intercontinental and marginal seas are considered to be a third category. These may be divided into group C_1, which consists of gulfs of basin seas, including the deepwater part only (Aqaba) or both the deepwater and the shelf parts (the Gulf of Honduras), and group C_2, the shelf gulfs of the same seas (the Persian Gulf, Gulf of Suez, Anadyrsky Zaliv, Bristol and Norton channels, Zaliv Shelikhova, etc.).



Plans of typical gulfs and bays.

Finally, there are the gulfs of the shelf seas (*gubas* of the Arctic seas of the U.S.S.R., gulfs of the Baltic and the White Seas, the Gulf of Carpentaria, the Gulf of Chihli, and many others), which are placed in group D (Table). The shallow character of the shelf seas influences the water dynamics of the gulfs. Water exchange is weakened, and sediments may accumulate in the gulf mouths, thus forming submarine barriers and further reducing exchange.

Certain characteristic features of these groups of gulfs and bays are enumerated in the Table. Some additional group characteristics can best be seen by describing in greater detail selected typical gulfs and bays exhibiting common and distinctive features.

TYPICAL GULFS AND BAYS OF THE WORLD

Gulf of Alaska. The Gulf of Alaska is a wide concavity of the North American Pacific coast. Its water properties and circulation are those of the ocean. Along its entire shore there are mountains and fractures that form fjord type bays and mountainous islands. The continental shelf in this area is rather broad to the north and west of the gulf, but it is dissected by troughs and contains other depressions. The main elements of coastal relief and continental shelf are of tectonic origin (related to earth deformation), but they received a final molding during the last glaciation. Glaciers still flow down into the heads of some bays and form icebergs. The shore of the gulf has high seismicity and is part of the Pacific earthquake belt.

Diatomaceous ooze (muds composed of unicellular organisms termed diatoms) is widespread in the deeper part of the gulf. An important peculiarity of the bottom topography of the Gulf of Alaska is the occurrence of submarine mountains with flat tops (guyots). These are ancient volcanoes that were abraded and planed off at or

near sea level. Subsequently, these volcanoes were submerged and their present depth is as much as 1,000 metres or more.

Bay of Bengal. The Bay of Bengal is part of the Indian Ocean. It washes the Indian and Sri Lanka (formerly Ceylon) shores in the west and borders on Burma and a chain of the Andaman and Nicobar islands in the east. Its basin occupies a tectonically stable area. It has a flat bottom and steep continental slopes. The shelf is narrow though developed all along the continental coast. It is fairly wide at the head of the bay where a broad delta of the Ganges and the Brahmaputra rivers is situated. A large submarine canyon begins at the mouth of the Ganges, which almost crosses the entire shelf. Submarine canyons of smaller size were discovered recently along the coasts of India and Sri Lanka. The Indian coast is relatively flat and is bordered with broad sand beaches. Coral reefs are developed along the Sri Lanka shores and those of the Andaman and Nicobar Islands.

Deltas, coral reefs, and submarine canyons

The water of the bay has certain oceanic features, but its salinity is somewhat lower. It exhibits a seasonally fluctuating circulation that changes with the monsoons. In spring the current moves clockwise with a velocity of three to five knots along the Indian coast. Subsequently, a counterclockwise and weaker circulation appears. During this period the salinity of the water falls sharply because the rainy season begins. During the dry season an upwelling of deep water takes place along the Indian coast (Waltair), the content of oxygen falls, and that of nutrient salts (phosphates and nitrites) rises.

The prevailing sediment of the deep part of the bay is Globigerina ooze. Alluvial mud accumulates in the head of the bay at a very high rate of deposition. There is a zone of shell-sediment and calcium carbonate concretions along the outward part of the Indian shelf.

Baffin Bay. Baffin Bay is a deep, elongated basin stretching in a northwestern direction between Greenland and the islands of the North Canadian Archipelago. It is connected by Smith, Jones, and Lancaster sounds with the Arctic Ocean and by the much broader Davis Strait with the Atlantic Ocean.

Ice molding and Holocene terraces

The flat bottom of Baffin Bay originated as a result of sinking (possibly along faults) that took place during pre-Quaternary time (more than 2,500,000 years ago). Its coasts consist of igneous and metamorphic rocks. They are mountainous and are cut by numerous fjords of glacial origin. In the northern part of the bay the shores are formed by the Greenland Icecap.

The bottom of Baffin Bay near its margins is of discontinuous character; broad troughs that continue the trends of the largest fjord systems occur there. Apparently these were molded by ice during periods of low sea level of Pleistocene time. It is possible that sounds of the Canadian Archipelago are ancient river valleys that were deepened by as much as 600 metres in pre-Quaternary time. Holocene terraces on the coasts have been raised to heights of more than 240 metres. The bottom of the central hollow is covered by land-derived mud; ice-rafted sediment is ubiquitous. Coarse sediment is deposited on the shelf and the continental slope.

Along the east shore of Baffin Bay the comparatively warm, northward flowing West Greenland Current occurs. Cold arctic water penetrates through Smith sound, over a sill that is less than 300 metres deep. Mixing with Atlantic water, these currents form the southerly flowing Baffin Island Current. Hence the general water circulation is counterclockwise. The waters of Baffin Bay also exhibit vertical stratification (layering), which results in an oxygen deficiency at great depths. The open part of the gulf is ice-free from July to September.

Gulf of Mexico. The Gulf of Mexico has a comparatively simple oval shape. Its shores lack large bays, but lagoons are abundant. There are two stretches of broad shelf to the north; these are off western Florida and Texas-Louisiana. The shelf areas are divided by the Mississippi Delta, in front of which a wide submarine trough is developed. There is a still broader area to the south that is termed the Yucatan (Campeche) Shelf. The shelf is very narrow in the west. A significant element of the bottom topography is a submarine cone descending to abyssal depth, which is derived from recent deposits of the Mississippi River. Steep continental slopes border the abyssal Sigsbee Plain, one of the flattest regions of the sea floor. There is a group of knolls, with a relative height of 370 metres, and smaller knolls have been discovered at the outer shelf zone and continental slope region of the northern Gulf of Mexico. Both groups of knolls are thought to be related to the intrusion of salt domes (*q.v.*).

Knolls of the Sigsbee Plain

In the middle of the gulf, the Earth's crust is of oceanic type, and a comparatively thin layer of continental type is present. In the north, there is a zone of crustal sinking that contains thicknesses of about 15 kilometres of Cretaceous and Tertiary sandy-clay sediments. To the east these alluvial sediments are replaced by a calcareous facies (lateral carbonate gradation). This elongated depression is called the Gulf Coast Geosynclinal.

The waters of the gulf have normal ocean salinity. They enter through the Yucatan Channel (1,500–1,900 metres deep) and form a mighty current. The latter makes a loop in the eastern part of the gulf at velocities as high as two metres per second and flows out through the Strait of Florida. This produces the Gulf Stream Current; an average of 25×10^6 cubic metres of water passes through the strait per second. In the western part of the gulf the currents have temporary character; their speed and direction depend on the winds.

Though the tidal range is small, the water level along the western and northern shores of the gulf may rise by as much as five metres during hurricanes. Destructive floods take place when this occurs. The waves of the gulf are much weaker than those of the ocean and their height generally does not exceed five metres.

Bottom sediments in the gulf are largely of terrigenous origin; silt occurs on the continental shelf and slope, and clay in the central parts of the gulf. A thin layer of carbonate sediments has been deposited on the shelves of Florida and Campeche. These rest upon a limestone surface of Quaternary and Tertiary age, which exhibits karst topography (pitted with caverns and sinks). The latter was formed during the Pleistocene lowerings of sea level.

Gulf of Oman and Gulf of Aden. The gulfs of Oman and Aden open into the Arabian Sea and exhibit similar physicogeographical conditions. They are fairly deep and are nearly equal in width, but the Gulf of Aden is twice as long as the Gulf of Oman. The salinity and the temperature of their waters are very high. The basins of both are of tectonic origin; they are related to the fracture system of the northwestern part of the Indian Ocean. Both gulfs have very narrow shelves and steep continental slopes, and the bottoms of their central parts are irregular.

Water circulation of the gulfs, connected with that of the ocean, depends on the changes of the monsoon regime. In the Gulf of Aden there is an established surface-water circulation. In summer it is clockwise and in winter counterclockwise. The same regularity may be observed in the Gulf of Oman, but it is less pronounced.

Gulf of California. The bottom of the Gulf of California consists of oceanic type crust with a thin (10–11 kilometre) overlying sialic (continental type) layer. It is part of a long trough that connects the San Andreas continental fault system and the mid-American deep-water trench. A number of recent, northwest-trending faults have broken the bottom into several isolated deepwater hollows; the gulf area is one of high seismicity. The continental slopes are very steep and the western slope is cut by a number of submarine canyons. There is a narrow shelf along the eastern side, and some rivers deposit their sediment loads on it.

Relation to the San Andreas fault

Three different areas of sedimentation occur from north to south. Alluvium of the Colorado River, principally silts and sands, is followed by diatomic silts, and oceanic clays occur farthest south. Carbonate sands are spread along the steep western shore.

The deep water of the gulf is oceanic in character. Thermal stratification (temperature zonation) causes an oxygen deficiency in intermediate layers. Surface-water circulation depends on the winds. Winter winds from the north produce an outflowing current, which is compensated for by upwelling of waters at the head of the gulf. Deep-sea water is rich in nutrient salts, and this upwelling leads to extensive plankton development. In summer, south winds direct the surface current toward the head, and water from the intermediate layer goes out into the ocean.

Bay of Fundy. The Bay of Fundy is remarkable for its high tides, which reach 18 metres in the head of Minas Bay (a world record). It is caused by the funnel-like shape of the bay and by a gradual rise of the bottom. This configuration results in resonance of the tidal wave, which is less than five metres high in the open sea. The same phenomenon raises tides sharply in the bays of Shelikhova, Cook Inlet, in the Mezenskaya Guba, and elsewhere.

High current velocities (two to four knots in the bay proper and up to 11 knots in Minas) are related to the coarse character of the bay sediments, which are mainly gravel. In the mouth of the bay these currents maintain a furrow that is more than 200 metres deep. Fine muds are deposited in protected areas near shore; broad tidal flats form in these parts of the bay.

Hudson Bay. Hudson Bay is situated within the Canadian Shield of Precambrian (more than 570,000,000 years ago) age. It has a simple rounded shape, but James Bay extends farther inland at the south. The shallowest part of the bay with a flat bottom is to the southwest. This is in the region of horizontally bedded Paleozoic rocks. The eastern part, which consists of Precambrian metamorphic rocks of greater surface irregularity, has a deeper and more irregular bottom, and a number of islands appear on the surface of the bay. Some irregularly shaped depressions about 200 metres deep exist in the central and the northern part of the bay. Recent research has shown that these hollows are the remnants of an ancient drainage net that led to the ancestral Hudson Strait. This strait connects Hudson Bay to the Atlantic Ocean today.

Drainage and Quaternary history

Coarse sand and gravel and much ice-rafted material are deposited in shallow places near shore. In the open part of the bay, to the west, the sediment is coarser (silt) in comparison to that in the east (silty clay). Sediment cores of the deep part of the bay reveal the presence of varved clays and till. This proves that the bottom of the bay was not submerged during the Quaternary regressions of the sea. During late phases of the glacial period, the area of Hudson Bay was one of the centres from which ice spread radially in all directions. Recently (during the last 8,000 years, approximately) the shores of the bay were subjected to isostatic uplift as a consequence of removal of the overburden of ice.

Hudson Bay receives a number of rivers, and its character is somewhat estuarine; pronounced water stratification exists in summer. Water coming through the Hudson Strait is of about 33.5 parts per thousand salinity, whereas the surface water is comparatively fresher. General water circulation is counterclockwise but this changes during periods of intensive river runoff. Hudson Bay is ice-free for only a few months a year, and the surface water is always very cool.

Gulf of St. Lawrence. The Gulf of St. Lawrence is situated between the crystalline rocks of the Canadian Shield and the folded Paleozoic rocks of the Appalachians. Its complicated shape and bottom topography are strongly determined by faults. The gulf is connected with the ocean by the Cabot Strait. Another narrow strait is called Belle Isle and is situated to the north of Newfoundland. A comparatively deep and smoothly curved trench leads from Cabot Strait to the mouth of the St. Lawrence River. It has some branches to the north of Anticosti Island and in the direction of Belle Isle. The southern part of the gulf (between Cape Breton and the Gaspé Peninsula) is quite flat and shallow.

The Earth's crust is up to 45 kilometres thick in the Gulf of St. Lawrence. The bottom topography of the gulf and its hilly shores received their modern shape through the influence of Quaternary glaciers. It is thought that a glacier excavated the comparatively deep furrows along the axis of the gulf. The northern shores of the gulf have undergone isostatic uplift at a rate of 40 centimetres per 100 years. The rate of uplift is zero in the direction of the southern shore, however.

The water of the cold Labrador Current penetrates the gulf, and St. Lawrence River water is mixed with this, reducing its salinity and raising its temperature (in summer). The general surface circulation is counterclockwise. Oceanic water forms the bottom layer with a salinity of 34 parts per thousand. The water stratification is subjected to seasonal changes. Ice cover is present from November through April in the western part of the gulf and until June in the eastern part.

Persian Gulf. The Persian Gulf is situated in a pronounced depression and is connected with the Gulf of Oman through the Hormuz Strait. The Arabian shore on the south consists of crystalline rocks of the Arabian Shield, upon which sedimentary rocks rest. This shore is fairly high but is bordered by a low-lying plain. Lagoons and small coral islands are present.

The Zagros Mountains extend along the northern shore. They were formed during Tertiary episodes of folding, and the thickness of rocks in the Zagros is over 12 kilometres. The Persian Gulf depression was created as a result of mountain-building processes. Salt plugs may be observed at the bottom and on the shores of the gulf, and there is some evidence of recent submergence in the deltaic area, while, to the south, there are uplifted terraces on both shores.

Topography, currents, and salinity

The bottom topography of the gulf is rather discontinuous, save for its northern and southern parts. Bottom deposits form three different regions, or zones. There is a thick layer of sand and clay with up to 20 percent calcium carbonate in the head of the gulf. In the centre, calcareous sediments of chemical and biogenic origins prevail. Coarse carbonate sands are deposited in shallow parts of the gulf, especially along its southern shore. These sands contain shell fragments, corals, and organic remains in general. Windblown material from the adjoining deserts is present everywhere.

Due to strong evaporation in the Persian Gulf, the surface waters of Oman penetrate the gulf waters as a compensation current. As their salinity increases, they become denser and flow back to the Gulf of Oman along the bottom. Strong tidal (up to four knots in the Hormuz Strait) and wind currents are superimposed on these salinity currents. Distinct vertical stratification probably does not exist because of strong currents and the shallowness of the gulf.

Zaliv Shelikhova. The Zaliv Shelikhova is an enlarged copy of the Bay of Fundy. The resonance of its tidal wave raises water levels in its forked head (Penzhinskaya Guba) by almost 13 metres. Tidal currents determine the coarseness of the bottom sediments, and a trough over 400 metres deep, stretching from the Sea of Okhotsk into the gulf, also is related to these tidal currents.

Zaliv Shelikhova is frozen for some months. In summer the upwelling of its deep cold waters takes place near the mouth, and the surface circulation is counterclockwise.

The Gulf of Thailand and the Gulf of Carpentaria. These gulfs have much in common. Both are vast shelf areas with rectangular contours (determined by tectonics) that open widely into the sea. The Gulf of Thailand is in the humid tropical zone, however, and the Gulf of Carpentaria, in the arid tropical zone.

The water of the South China Sea penetrates into the Gulf of Thailand at its bottom, and the freshwater of the Mae Nam Chao Phraya (Chao Phraya River) is spread on the surface. This forms stratification of estuarine character. The surface water circulation, depending on the monsoons, is clockwise in autumn and winter and coun-

Water circulation and swamp occurrences

terclockwise in summer. Upwelling of bottom water of high salinity (up to 34 parts per thousand) takes place in areas of divergent currents. Muddy alluvium of the Mae Nam Chao Phraya covers the entire gulf bottom, and a considerable part of its low shores with mangrove swamps were similarly derived.

The Gulf of Carpentaria extends into the northern shore of Australia. Its flat basin is bordered by faults that separate Precambrian rocks of the western coast from Mesozoic and Tertiary rocks of the east. These rocks are not visible everywhere because the low shore is bordered by mangrove swamps. The gulf was not filled by seawater until Holocene time.

Mezenskaya Guba. This bay is distinguished from other gulfs and bays of the White Sea by the high tides at its head. In this respect it is similar to the Bay of Fundy. Because of the strong currents, the flat bottom of the bay is covered by coarse gravel and exposed outcrops of Permian carbonate rocks. Clay and peat shores of the bay retreat at a rate of up to ten metres a year. This is due to the influence of drift ice (the bay is frozen for some months) and melting of the permafrost within the shores. Muddy tidal flats, many kilometres in width, exist in estuaries of Mezenskaya Guba and in some open areas. These flats also constitute a characteristic physical feature of the area.

Obskaya Guba. The longest inlet of the southern shore of the Kara Sea is the Obskaya Guba. It receives the runoff of the Ob and Taz rivers, and its currents tend to have runoff characteristics. In summer, water emerging from the *guba* is six to eight degrees warmer than that in the adjoining part of the sea.

The shores of the *guba* consist of loose Quaternary alluvium. They are precipitous on the west, where they are being attacked and destroyed by the sea. On the east they are mainly low-lying.

Gulf of Chihli. This gulf is noteworthy because it receives the water of the Yellow River, which has the largest sediment discharge in the world (up to $1{,}380 \times 10^6$ tons a year). The flat bottom of the gulf is covered by yellowish mud, which is derived from erosion of the loess of China. The southern and the western shores of the gulf are also muddy because the mouth of the Yellow River has repeatedly migrated through time.

The turbid water of the gulf is warmed down to the bottom by strong tidal currents, and productivity (fish, shrimp, and oyster) is high. The Río de la Plata of South America is quite similar; this gulf is in a basin that is completely filled by alluvium of the Paraná River.

Gulf of Bothnia and Gulf of Finland. These gulfs of the tideless Baltic Sea have much in common. They receive the runoff of many rivers and their turbid waters are especially freshened at the basin heads. Local, unstable circulations that depend on the winds and river floods are characteristic of both of them. Fluctuations of water level due to seiches are peculiar. In the Gulf of Finland these reach a height of 120 centimetres, and high storm surges (up to 375 cm) at its head can cause disastrous floods. In winter both gulfs are frozen for some months.

Ice retreat and isostatic uplift

During the last glaciation the basins of both gulfs were filled with ice. When glacial retreat and ice melting occurred, fresh glacial lakes were formed in this region. Subsequently, as the world sea level rose and the isostatic uplift of Fennoscandia (Norway, Sweden, and Finland) occurred, the shape, water composition, and marine population of the gulfs changed several times. A complicated history has been deciphered from study of terrace altitudes and ancient deposits of the shores and the bottom of the gulfs. The shores are still rising today. In the head of the Gulf of Bothnia the rate of uplift is over one metre a century. This diminishes to the south and is essentially zero in the head of the Gulf of Finland.

The mouth of the Gulf of Bothnia is partitioned off by the Aland Skerries Archipelago, and it is only to the west that a deep strait remains. The hollows of the two gulfs occupy ancient depressions of the crystalline Baltic Shield. There also are pre-Paleozoic fractures along some parts of the shores that may have undergone renewed movement during the Tertiary Period. The southern shore of the Gulf of Finland is formed by horizontal layers of Paleozoic rocks. The basins of the gulfs were considerably influenced by glaciers, and moraine ridges are still preserved at the bottom of the Gulf of Finland. The northern shore of the Gulf of Finland and the whole of the southern part of the Gulf of Bothnia are strongly dissected by *fjärds* and skerries islands.

The sediments of the central part of the Gulf of Bothnia consist of terrigenous muds that overlie varved glacial clays and till (see VARVED DEPOSITS). Sediments of the Gulf of Finland are coarser, and in many places sands are deposited on its bottom. There is an admixture of ice-rafted stones in the sediments of both gulfs.

BIBLIOGRAPHY. Information on the properties and general characteristics of large gulfs and bays of the world may be found in R.W. FAIRBRIDGE (ed.), *The Encyclopedia of Oceanography* (1966). A.J. HUXLEY (ed.), *Standard Encyclopedia of the World's Oceans and Islands* (1962), contains abbreviated information on numerous gulfs and bays, including some of the smaller ones. For morphometric and physical-geographical data on some of the larger gulfs and bays, see E. BRUNS, *Oceanologie*, vol. 1 (1958). Recent monographs on particular gulfs and bays include: F.P. SHEPARD *et al.* (eds.), *Recent Sediments, Northwest Gulf of Mexico* (1960), a summary of all aspects of the Gulf of Mexico as known through the late 1950s; T.H. VAN ANDEL and G.G. SHOR (eds.), *Marine Geology of the Gulf of California* (1964); and V.P. ZENKOVICH (ed.), *Tikhiy okean:berega Tikhogo okeana* (1967), a fairly detailed description of geological and geomorphological conditions in most of the gulfs and bays of the Pacific Ocean region (in Russian).

(V.P.Z.)

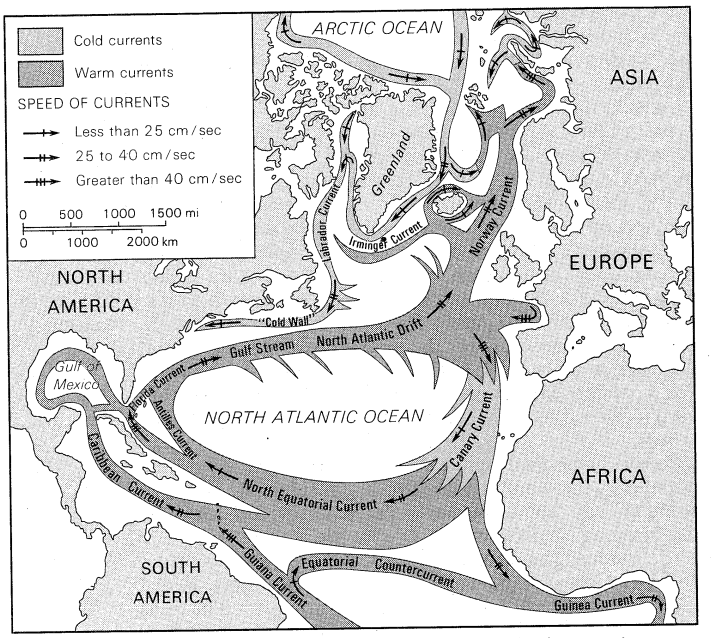
Gulf Stream

The name Gulf Stream often is used in reference to the entire warm ocean current flowing northeastward from the Straits of Florida to the coast of northwestern Europe. It refers properly, however, only to that portion of the western Atlantic current extending from the vicinity of Cape Hatteras, North Carolina, to the area southeast of the Grand Bank off Newfoundland. In deference to popular usage and to the importance of the continuous current, this article will cover the larger, though less correctly named, phenomenon. (For information on related topics, see the articles ATLANTIC OCEAN; OCEAN CURRENTS.)

Origins and components

The Gulf Stream is part of a general clockwise-rotating system of currents in the North Atlantic. It is fed by the westward-flowing North Equatorial Current moving from North Africa to the West Indies. Off the northeastern coast of South America, this current splits into the Caribbean (Florida) Current, which passes into the Caribbean Sea and through the Yucatán Channel into the Gulf of Mexico; and into the Antilles Current, which flows to the north and east of the West Indies. The Caribbean Current re-emerges into the Atlantic through the 95-mile-wide Straits of Florida between the Florida Keys and Cuba. It is then deflected to the northeast by the submerged Great Bahama Bank southeast of the Florida Peninsula. North of the Bahamas it is joined by the Antilles Current and flows roughly parallel to the eastern coast of the United States to about Cape Hatteras. Here it veers more to the east and passes close to the Grand Bank, south of Newfoundland, before turning eastward as the North Atlantic Drift (or current). The drift later subdivides, one branch moving southeast and south as the relatively cool Canary Current, while the warm North Atlantic Drift continues northeast off the British Isles and into the North and Norwegian seas.

History of scientific study. The Gulf Stream was first described by the Spanish navigator and explorer Juan Ponce de León early in the 16th century. Later it was studied by such scientists as Benjamin Franklin and the American naval officer Matthew Fontaine Maury, and in 1844 systematic surveying of the stream was begun by the United States Coast and Geodetic Survey. Concentrated modern efforts were inaugurated only in the early 1930s by the ketch "Atlantis" of the Woods Hole Oceanographic Institution in Massachusetts. Since then a great



The Gulf Stream in relation to other North Atlantic currents.

many surveys have been made, particularly by a number of scientists from Woods Hole.

One of the difficulties of scientific study of the Gulf Stream is its extremely complex makeup. It is not a simple ribbon of moving water but rather a complicated network of currents that tend to shift course over time, to disappear and then reappear, and to develop eddies along the margins. To supplement the work of individual research vessels, two multiple-ship surveys have been carried out. Seven ships took part in "Operation Cabot" in 1950; and, a decade later, in "Gulf Stream '60," four ships obtained hydrographic data over a three-month period in the general area east and southeast of Cape Cod, Massachusetts.

Movement and physical features. Most of the waters that enter the Gulf Stream system first have been driven westward across the Atlantic by the Northeast Trade Winds. In the Caribbean and the Gulf of Mexico the current is gradually narrowed, and its velocity increases to about 3.5 knots (four miles [6.5 kilometres] per hour) as it passes through the Straits of Florida. The volume of flow here has been measured at 920,000,000 cubic feet per second, or about 1,000 times that of the Mississippi River. As it turns north between Florida and the Bahamas, the Florida Current flows at a depth of some 2,600 feet and then follows the continental slope beyond the edge of the shelf. Velocities gradually decrease to about one knot (one mile [two kilometres] per hour) off Cape Hatteras. In the vicinity of St. Augustine, Florida, the current is about 30 miles off the coast; at Savannah, Georgia, 85 miles; and at Cape Hatteras, less than 20 miles.

Salinity and temperature

In the western Atlantic, the current's deep-blue water, with its higher temperature and salinity, is readily distinguishable from surrounding waters, particularly along its well-defined western margin. This edge corresponds closely with the 650-foot depth, whereas the eastern edge gradually moves seaward as the current moves northward. The water between the current and the mainland, with its lower salinity and temperature, forms a boundary known as the "Cold Wall." This water, overlying the continental shelf, frequently has a southerly flow, counter to that of the Florida Current.

Water in the eastern portions of the current is warmer than that in the western part. The lateral mixing of current and ocean waters results in eddy formations along the margin. Off the coast of the United States, the Gulf Stream system separates the relatively warm waters of the Sargasso Sea in the mid-Atlantic region from the colder waters to the west and north. In winter, for example, average surface temperatures of the Gulf Stream off New England may be 20° F (11° C) higher than those of surface waters only 150 miles to the northwest, although there is less than a 10° F (6° C) change in surface-water temperatures over a 1,000-mile distance to the southeast.

Beyond Cape Hatteras the Gulf Stream broadens and moves into deeper water. Here it crosses the Western Boundary Undercurrent, which consists of cold, southward-flowing water that sinks to considerable depths in the vicinity of Greenland. About 1,500 miles northeast of Cape Hatteras, in the area of the Grand Banks, the warm Gulf Stream waters come in close proximity with the cold, southward-flowing Labrador Current. The contact of cold, humid air moving over the Labrador Current with the warm surface waters of the Gulf Stream causes widespread condensation. This climatic condition causes the region to have one of the highest incidences of fog in the world.

Easterly and southerly currents

Moving out into the North Atlantic, the current becomes shallower and begins to break down into a meandering pattern of disconnected filaments flowing in the same general direction. The pattern itself often changes, and marginal eddies may detach themselves completely from the main body of flow. Much of the initial force of the current has been dissipated by this time, and momentum is afforded primarily by the westerly winds. Part of the water here is diverted southward into the Sargasso Sea area. Near the middle of the ocean, the North Atlantic Drift divides into the Canary Current, which flows southward past the Iberian Peninsula and northwestern Africa, and the North Atlantic Drift, which moves toward northwestern Europe. One branch of the drift, the Irminger Current, is diverted northwestward toward the south and west coasts of Iceland.

Effects on marine and human life. The marine organisms of the Gulf Stream system are not of great commercial value. Principal species include the bluefin tuna, the Atlantic salmon, and the flying fish. Its warm waters, however, in mingling with the colder waters both on the Grand Bank and off northwestern Europe, contribute to turbulence and the availability of nutrient salts that have made these regions among the most productive commercial fishing grounds in the world.

Warming of northwestern Europe

A major contribution of the Gulf Stream system is its warming effect upon the climates of adjacent land areas. In winter the air over the ocean west of Norway is more than 40° F (22° C) warmer than the average for that latitude, probably the greatest temperature anomaly in the world. The prevailing westerly winds carry the warmth and moisture of the ocean to northwestern Europe, giving Bergen, Norway, at 60° north latitude, the high average temperature for its coldest month of 34° F (1° C), while Reykjavík, Iceland, 4° of latitude farther north, has a 30° F (−1° C) average for its coldest month. In southwestern England the climatic modification produced by the current is reflected in the extraordinary mildness of the winters at this northern latitude, including the growing of winter vegetables and flowers and the presence of subtropical vegetation and lemon trees in southern Devonshire. Along the western margins of the North Atlantic, however, where the winds are predominantly from the shore, the Gulf Stream has little effect. Halifax, Nova Scotia, nearly 1,000 miles south of Bergen, averages only 23° F (−5° C) during its coldest month.

BIBLIOGRAPHY. HENRY STOMMEL, *The Gulf Stream: A Physical and Dynamical Description*, 2nd ed. (1965), one of the standard works in English on the Gulf Stream system, includes considerable oceanographic data and detailed maps; F.C. FUGLISTER, *Atlantic Ocean Atlas of Temperature and Salinity Profiles and Data from the International Geophysical Year of 1957–58* (1960), the most significant comprehensive mapping of Gulf Stream data available; D.F. BUMPUS and L.M. LAUZIER, *Serial Atlas of the Marine Environment #7: Surface Circulation on the Continental Shelf Off Eastern North America Between Newfoundland and Florida* (1965), an excellent background source for knowledge about the ocean environment of the Gulf Stream; HENRY CHAPIN and F.G. WALTON-SMITH, *The Ocean River* (1952), a popularized description of the North Atlantic Ocean with particular attention paid to the Gulf Stream.

(L.M.A.)

Gunnery

Gunnery is the technique of developing, operating, and maintaining all weapons of war classed as guns; that is, all weapons that fire large-calibre projectiles. (For a description of small arms, see SMALL ARMS, MILITARY.) Modern guns employ explosive charges as propellants, but the earliest missile-throwing weapons, such as ballistas and catapults, used torsion, tension, and counterweights. A few modern small-calibre weapons use the word gun in their nomenclature; *e.g.*, machine gun and shotgun.

From the introduction in the 14th century of explosives for propelling projectiles until about the second decade of the 20th century, classification was not difficult. Land-based weapons were known as artillery and those used at sea as naval guns. All these were surface-to-surface weapons. The advent of aircraft, long-range rockets, and guided missiles has vastly complicated classification, and there are now four dimensions: surface-to-surface, surface-to-air, air-to-surface, and air-to-air.

The most modern types are highly sophisticated, as are their projectiles, means of movement, and ancillary equipment. Many are closely guarded secrets, and all are subject to changes and short life because of supersession by improved types. For every weapon in service there is likely to be another in an experimental stage and yet another on a drawing board. In these circumstances, a description of contemporary artillery weapons can be made only in general terms.

History of gunnery

"ENGINES OF WAR" AND MEDIEVAL GUNS

Objectives of the attacker and defender

Throughout history, battles and engagements have almost invariably been fought with one side on the offensive, the other side defending. The attacker's object is to bring his assault troops—infantry, cavalry, and in modern times tanks—to close quarters as quickly and as economically as possible. To accomplish this he has sought means to inflict casualties and cause confusion before delivering his assault—a process called "softening up" in modern military parlance. Similarly, the defender has sought means to inflict losses on the attacker and cause the assault to break down before reaching its objective.

Among the earliest ancients this was done by means of light, short-range weapons such as slings and bows and arrows. Later, more sophisticated weapons were introduced with longer ranges and capable of discharging heavy stones and spears by means of springs and balancing devices. Little is known of the detailed design or methods of employing these weapons. It would appear that their chief value was in siege warfare against fortifications; but it is known that both the Greeks and Romans used them as field weapons. Their material value in the field must have been quite limited, though their large size and awe-inspiring appearance were no doubt a morale factor. Although their missiles were large and formidable in appearance, they did not fragment like modern shells, and there must have been a long interval between shots.

As with modern weapons a wide variety of names have been given to similar types—catapult, ballista, mangonel, scorpion, springal, falarica, and many others—all with the common characteristic of being able to discharge a heavy missile a longer distance than was possible with a hand-operated weapon. (See also WEAPONS AND DELIVERY SYSTEMS.)

In the earliest types, two methods of propulsion were used; namely, torsion, which was the twisting of thick rope, and tension, the bending of a spring or piece of wood. Later, in medieval times, a third method was employed: the trebuchet (and others of similar type) propelled a stone by means of a balance or counterweight.

With the introduction of gunpowder, or black powder, the whole conception of long-range weapons changed. Gunpowder set in motion a process of development that over the centuries produced more and more massive bombardments. As has remained the case throughout history, the older weapons continued to be employed after the introduction of the new. The "engines of war" operated concurrently with guns using explosives for several decades.

Considering the lack of communications in the 14th century and the revolutionary nature of the invention, it is remarkable how quickly gunpowder was adapted to military purposes. It is quite certain that it was not introduced into Europe before the first decade of the 14th century, although it, or some similar explosive, seems to have existed in China before then. Doubts exist as to when guns were first employed in battle. Some historians give credit for making the first gun to a German monk, Berthold Schwarz. Claims for the first use of guns in battle have been made for the siege of Metz (1324) and at Cividale in Italy (1331). There is little doubt that after 1325 cannon existed all over western Europe and were certainly used by the English under Edward III at Crécy (1346).

First use of guns in battle

The probable inefficiency of these early guns would account for the continuance for a considerable time of some of the older weapons, particularly the longbow used by English archers.

The missiles for these early guns were originally in the form of a spear, somewhat like a harpoon; later, shaped stones, and later still, iron balls were employed. The guns themselves were mostly made of cast bronze or brass. These materials were followed by wrought-iron guns, but it was not until the 15th century that cast iron was used.

DEVELOPMENT OF GUNPOWDER ARTILLERY

One of the difficulties of early gunnery was the erratic behaviour of gunpowder, the ingredients of which did not readily mix and which tended to deteriorate after mixing, particularly when transported. This made it unpredictable in use. Early in the 15th century, however, a new process, known as "corning," was invented; this improved method of mixing went a long way toward producing a standardized powder.

Nevertheless, the progress made in the development of artillery during the next 300 years, although steady, was far from sensational. As with some new weapons in modern times, guns did not at first fulfill the expectations of their protagonists. This was not altogether the result of the extremely crude structure of the first types; the high efficiency of some of the older weapons, the paucity of communications, and the general conservatism of the times contributed. As a result, as late as the reign of King Henry VIII of England (1509–47), 200 years after the introduction of gunpowder, weapons of the old "engines of war" type were still in use.

During the 15th century a number of very large guns, known as bombards, were manufactured in widely different countries. Edinburgh Castle's famous "Mons Meg" weighed some five tons and was said to throw a 19.5-inch iron ball nearly a mile; a 13-ton bombard, "Dulle Griete" of Ghent, had a 25-inch calibre and threw a 700-pound missile; an even larger weapon was the "Tsar Cannon" of Moscow, which had a 36-inch bore and weighed some 40 tons. Details are lacking of the formidable pieces known to have been used by the Turks when they captured Constantinople in 1453. All of the above were smoothbore weapons. It is said that rifling (*i.e.*, grooving the inside of the gun barrel to impart spin to the projectile) was attempted in the 15th century and again in the 17th, but with little success.

It must be appreciated that these heavy and cumbersome weapons, using ammunition difficult to produce and handle, were of little value except as part of the armament of a fortress or for siege operations planned well in advance. During the 15th century a good deal of experimental work took place with a view to producing guns sufficiently mobile to accompany field armies and take part in encounter battles. Early attempts by mounting guns on sledges drawn by oxen proved unsatisfactory. Under Charles VIII of France (1483–98) guns were mounted on wheeled vehicles drawn by especially trained horses. About the same time trunnions were introduced —a pair of wooden devices forming a cradle to balance

Efforts to produce mobile weapons

the gun on the carriage, by which the muzzle could be elevated or lowered for range adjustment, and through which some of the shock of discharge was transferred to the carriage. These measures succeeded in producing weapons that could keep pace with marching infantry on reasonably good roads over flat country. The Italians also produced field guns, and an Italian condottiere, Bartolomeo Colleoni (died 1475), introduced a special form of tactics for field artillery in which the guns fired from the rear through gaps left by the infantry.

There is evidence that the 15th century also saw the introduction of what has become known as "case shot," in which, instead of a single missile, a large number of round bullet-sized shots were fired, which scattered widely like the pellets of a shotgun, for use at short ranges.

The 16th century did not mark any revolutionary step forward in artillery techniques but rather an improvement on existing types of guns and their better classification and organization.

Weapon standardization

Hitherto standardization had been unknown, but in 1544 Charles V of Spain decided on seven different types of cannon for use in his armies. These included a 40-pounder (firing projectiles weighing 40 pounds, or 18 kilograms), a 34-pounder, two types of 12-pounder, two 6-pounders, and a 3-pounder. In France in 1550 Henry II standardized with six types, the heaviest a 33-pounder (drawn by 21 horses) and the lightest a 2-pounder (drawn by 2 horses). Later (in 1584) a 12-pounder and a 24-pounder were added. In the German states, standardization, although attempted, did not progress to the same extent. In the period 1550–1600 there were said to be 11 different types of gun, from a 1-pounder to a 94-pounder, which with variations in each type brought the total to around 40.

In England, Henry VIII took a personal interest in the provision of guns, but in the absence of a home industry was forced to rely on overseas products. By employing a Fleming, Hans Poppenruyter, he acquired some 150 guns of varying calibres, including the celebrated bombards known as the "Twelve Apostles." About 1515 he imported a number of foreign armourers and established schools of instruction for English craftsmen. Largely because of Henry's encouragement, two crude types of shell were introduced, which can be recognized as the forerunners of the explosive shell and the incendiary missile of modern times.

Perhaps the most important development of the 16th century was the beginning, in Italy, of the science of ballistics, which in time was to promote gunnery from an empirical rule-of-thumb operation to one of scientific precision.

By the first half of the 17th century artillery was taken seriously. In an attempt to improve mobility, Gustavus II Adolphus of Sweden introduced the "leather gun" in 1626. By making the external casing of the barrel of leather, and the bore of copper tubes, he reduced the weight of the gun to 90 pounds (40 kilograms), with a corresponding reduction in the weight of the carriage. Its mobility was excellent, but it generated excessive heat and was dangerous to operate. In 1631 it was discarded in favour of a less mobile but safer weapon.

Although the organization and standardization of the artillery of most countries became less haphazard than in the previous century, ammunition supply remained a problem, and there appears to have been no arrangement for the supply of spare parts for either guns or carriages.

Siege warfare, and operations in mountainous country, demanded a weapon with a high trajectory. This requirement was met by the mortar, sometimes called a bombard or howitzer, which could lob a missile into a fortress or castle or over a small hill. These were smoothbore weapons with very short barrels; the length of the barrel was sometimes no more than its calibre.

Not long after the introduction of artillery its potential in sea warfare was recognized. As early as 1509 a Venetian fleet, equipped with guns adapted to the galleys of the time, rowed up the Po to within a few miles of the capital of the Duke of Ferrara. The guns of the Venetians, however, proved no match for the Duke's land artillery, and the Venetian fleet was destroyed. In addition to greatly increasing the prestige of artillery, this action gave birth to a principle that lasted for more than 400 years; namely, that, unless in overwhelming strength, naval guns are always at a disadvantage against shore-based artillery.

The progress of naval gunnery, and the differing doctrines for its employment during the Middle Ages, is well illustrated by the series of actions in the English Channel in July and August 1588 between the Spanish Armada and the English fleet. In these engagements the Spanish fleet relied on large ships equipped with heavy but comparatively short-range guns (average weight of shot 17 pounds). The English, on the other hand, relied on much smaller but more mobile ships with light guns (average weight of shot 7.5 pounds) with a longer range than those of the Spaniards. The English tactics paid a good dividend as they were mostly able to keep out of range of their enemy's big guns while inflicting damage with their own weapons. Moreover, the heavy guns of the Spaniards soon ran out of ammunition. The extent to which guns had come to play an active part in naval warfare by the end of the 16th century is demonstrated by the number judged to have been available to the two sides; namely, Spain 1,124 (44 percent heavy, short-range and 56 percent lighter, medium-range); England 1,972 (5 percent heavy, short-range and 95 percent light, long-range).

Defeat of the Armada

THE 18TH-CENTURY IMPROVEMENTS

In the nearly 400 years from the introduction of gunpowder until the end of the 17th century there had been a considerable increase in the range of guns, and this in turn had produced changes in the tactical handling of artillery. There had also been a steady increase in the number of pieces available. Apart from this, however, little progress had been made in the guns or their projectiles. Gunnery remained haphazard, uninfluenced by ballistics and other sciences still in the experimental stage.

This slow pace quickened during the 18th century. In France measures were introduced for the better classification of guns, and in 1776 the organization of the French artillery was greatly simplified by a reduction in the number of calibres and improvements in the design, and particularly the uniformity, of ammunition. It was not until their Revolutionary Wars, however, that the French introduced horse artillery—guns light and mobile enough to accompany cavalry in the field.

The Seven Years' War in Europe (1756–63) found Prussia in the process of making a series of artillery experiments, as was Austria, particularly with the more precise weighing and measuring of ammunition with a view to achieving uniformity of flight.

Perhaps the greatest progress was made in Great Britain, where Benjamin Robins published his *New Principles of Gunnery* in 1742. This work exposed the fallacy of many of the old theories and methods and, for the first time, brought science into the field of practical gunnery. Robins' invention of the ballistic pendulum enabled the velocity of missiles to be judged accurately at any stage of their flight. The year 1784 marked the invention by Henry Shrapnel, a British officer, of the form of ammunition that bears his name—an antipersonnel projectile timed to burst in the air toward the end of its flight and discharge a large number of small, bulletlike balls over a wide area. Shrapnel remained a major artillery missile until World War I.

Invention of shrapnel

An important 18th-century advance was the introduction of limbers, vehicles towed behind the gun carriage for the carriage of ammunition, tools, and spare parts and for carrying some members of the gun team. In most European armies artillery was divided into three main categories: horse, for use with cavalry; field, for use with infantry; garrison, for coast defense and other static roles.

Despite the increasing attention devoted to artillery and its increased prestige as a battle-winning factor, the number of guns deployed was still, by modern standards, modest. In one of the greatest battles of the 18th century, Blenheim (1704), the French and Bavarians had approxi-

mately 120 guns and the British and their Allies under Marlborough about 60. During the whole of Marlborough's campaign in Europe the average proportion of guns to infantry in his Allied army was about nine guns to 10,000 men.

THE 19TH-CENTURY ADVANCES

The 19th century saw remarkable developments in every aspect of gunnery: the pieces, the projectiles, and the propellants.

There were three main reasons for this rapid progress. The first was the influence of Napoleon, whose sensational victories, in many people's view, were the result of his use of artillery. This was probably an exaggeration, but it undoubtedly added to the prestige of the arm. Napoleon introduced the method of using massed artillery at what he judged was the vital point of battle; he also raised the status of his artillery men by making them a *corps d'élite*. Second, the advance of science made possible improvements in both guns and ammunition. Third, the number of wars—the Napoleonic Wars ending in 1815, the Crimean War (1853–56), the American Civil War (1861–65), the Franco-Prussian War (1870–71), and many colonial wars in Africa and Asia—provided an incentive to produce better guns. The better organization of armed forces, including the general staff system initiated by Prussia, promoted greater and more efficient use of artillery.

Introduction of rifled barrels

As a result of scientific progress and war experience, several improvements came about. Rifled gun barrels were introduced. These imparted spin to the missile, resulting in greater accuracy, less deflection by wind, and a heavier and shaped missile in place of the mostly round balls previously used.

Improved powder to propel the heavier ammunition was made available by the researches of Capt. (later Gen.) Thomas J. Rodman (U.S. Army); but this in turn (in the 1860s) was superseded by the introduction of guncotton by Baron General von Lenk of the Austrian Army. This explosive had the advantage of being smokeless. It was, however, dangerous in storage until a French chemist, Paul Vieille, in 1887 produced a process that made it safe.

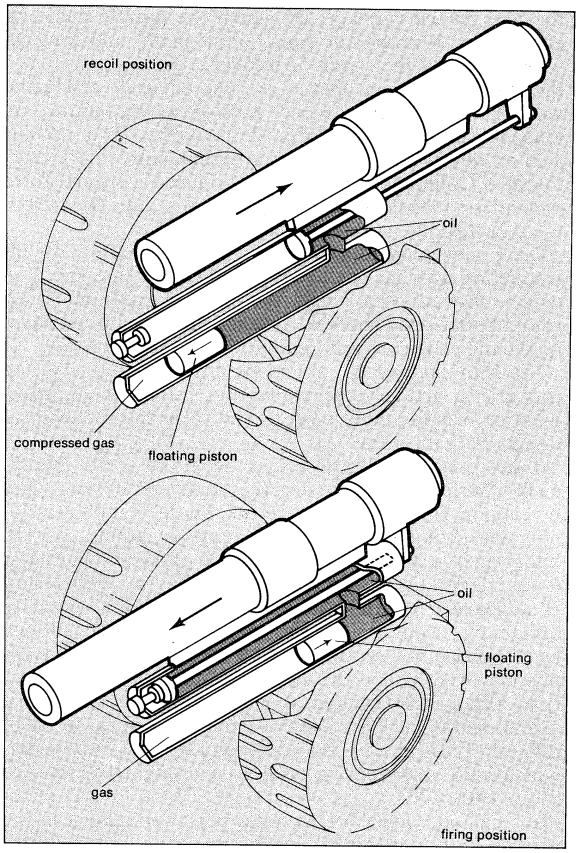
The provision of guns sufficiently robust to withstand the increased charges that had become available posed a serious problem. Various attempts to solve this problem were made in France, in the United States, and in Britain. The answer came in 1851, when Alfred Krupp of Germany displayed in London an all-steel gun drilled out of a single block of cast metal. This weapon was only a six-pounder, but within ten years Krupp was producing guns of cast steel of more than eight-inch bore.

Toward the end of the century two problems were solved that had defied the efforts of inventors; namely, a satisfactory breech-loading mechanism and a satisfactory recoil system. Breech loading was achieved in the 1860s and 1870s almost simultaneously in a number of countries—France, Germany, Spain, and the United States. About the same time, the recoil problem was overcome by the introduction of a system for absorbing the shock of discharge and leaving the gun in approximately the same position after firing as before (see illustration).

Introduction of pack guns

Consideration was also given to other types of artillery. In the second half of the century pack guns were introduced; these were light weapons broken down into convenient loads for carriage on packhorses or mules, for use in mountain warfare. These were employed extensively by the British on the northwest frontier of India and survived in World War II when they were used in Italy and other mountainous theatres. Attention was also given to guns of larger calibre than had hitherto accompanied field armies, with the use of railroad transport. One of the surprises of the early weeks of World War I was the German deployment of heavy guns of Austrian manufacture to bombard the forts around Liège.

The 19th century saw a radical change in artillery tactics. In the Napoleonic Wars the short range of guns made it necessary to site them well forward. The increasing range of artillery gradually made this unnecessary; it became the custom to deploy the guns well in rear of the infantry to shoot over their heads.



Simple recoil system used on field gun. When the gun is fired, oil is forced into the cylinders, compressing the gas and absorbing the recoil.
Adapted from I.V. Hogg, *The Guns 1939–45*.

Finally, there was a brief revival of interest in rockets, with which British warships were equipped during the Napoleonic Wars. The development of artillery temporarily put an end to rocketry as a military weapon.

WORLD WAR I GUNNERY

When in August 1914 World War I began, conventional artillery and the techniques for its employment had approached their zenith. Improvements made during the war did not touch fundamentals. Heavier types of gun were developed, and quantities of both guns and ammunition startled the military world, but basically the artillery pattern remained quite similar to that in the Franco-Prussian War of 1870–71. Light (or field) artillery, plus a few heavier guns, formed part of divisional artillery in all major armies, with more of the heavy pieces under higher direction. As the war progressed it became common practice for all light artillery plus a few medium guns to remain with divisions, most of the medium artillery (including counterbattery pieces) under corps control, and heavy, superheavy, and antiaircraft under army headquarters control.

In the opening stages of the war the artillery of a British division comprised three brigades each of three batteries of 18-pounder guns, and one brigade of three batteries of 4.5-inch howitzers and one heavy battery of 60-pounder (5-inch) guns. The divisions of other major belligerents were equipped on much the same lines with only minor differences. French batteries were four-gun; British, six-gun. The French medium and heavy pieces within the division were 4.8-inch guns and 5.9-inch howitzers; the Germans employed 5.9-inch and 8-inch howitzers.

Light guns of World War I

Practically all the light guns in service in 1914 stood the test of war. Special mention, however, must be made of the French 75-millimetre guns, first brought into use in 1897, that not only remained the French Army's chief artillery weapon throughout the war but also was the

main equipment for the American divisions during the latter stages.

To understand the vastly increased ammunition requirements of World War I it is necessary to understand the change in the pattern of fighting. The armies of the great powers had been equipped in peace in the belief that the next war would be a short one and the fighting open and fluid in character. These conditions required highly mobile light artillery firing mainly antipersonnel projectiles (*i.e.*, shrapnel). By the end of 1914 the opposing armies on the Western Front had settled down to siege warfare —continuous lines of trenches with deep dugouts and concrete machine-gun emplacements, protected by barbed wire. This setting required a higher percentage of heavy guns and high-explosive shells, and a vastly greater number of all types of guns and projectiles, if either side was to carry out a successful offensive. On the Eastern Front the same trend toward trench warfare was present but in a less marked degree than in France and Belgium.

Production of artillery

In the continental countries of Europe, with their large conscript armies, the problem was difficult but not desperate. They had a good proportion of heavy guns, and their industries had been geared in peace to the production of large quantities of guns and ammunition. In Britain, with its small volunteer army and original Expeditionary Force of no more than six infantry and one cavalry division, the position was very different. It was necessary to create the industries for war production at the same time as the nation expanded its army to many times its peace strength. The United States was to experience similar difficulties when it entered the war in April 1917.

Some idea of the immense increases in artillery as the war progressed can be seen in statistics: In August 1914 the first six divisions of the British Expeditionary Force (BEF) in France had a total of about 486 guns, all but about 24 being light fieldpieces. By May 1915 the total of guns with the BEF had risen to 700 18-pounders; 200 15-pounders; 125 13-pounders; 80 4.7-inch guns; 28 60-pounders; 50 5-inch howitzers; 130 4.5-inch howitzers; 40 6-inch howitzers; and 12 9.2-inch howitzers, for a total of 1,365. At the Armistice in November 1918 the total of all types was 6,437.

On the opening day of the Battle of the Somme, July 1, 1916, the number of heavy guns on the 18-mile front of the British attack was 455, or one heavy gun every 57 yards. In addition there was the whole of the light artillery of some 19 divisions (about 72 per division).

By the Battle of Messines (Mesen) in 1917 the number of guns supporting the attack was 2,266. Of these 756 were heavy pieces—one to every 20 yards. The preliminary bombardment began on May 21, and the attack was made on June 7. Altogether 3,500,000 shells were fired at an estimated cost of £17,500,000.

In the attack on the Saint-Mihiel salient by United States troops in September 1918, on a front of about 12 miles, the attacking divisions were supported by 3,010 guns of all calibres, or one every seven yards.

Technical improvements in guns

In addition to the greatly increased size of guns, there were several areas of technical improvement. The first was range. All light artillery achieved ranges of at least 10,000 yards (9,000 metres), and the German 21-centimetre gun, known as "Big Bertha" and used for shelling Paris in 1918, attained a range of 76 miles (122 kilometres). Night firing became a matter of normal routine, and defensive fire or counterpreparation, usually a static barrage or line of bursting shells, was introduced. Later, the creeping barrage, a line of fire that moved forward slowly, and behind which the infantry advanced, became normal practice in attack. Finally, artillery communications were greatly improved by field telephones and radio, and the location of targets by means of aircraft and captive balloons.

The greatly increased use of aircraft brought into use special guns, firing shrapnel, for anti-aircraft purposes. In the absence of adequate fire-control methods, which came later, these weapons were not very effective in World War I.

Along with the developments in guns came improvements in ammunition and increases in the variety of shells. Except with a few heavy guns and howitzers, in which the propelling charge remained separate, all artillery projectiles used in World War I were self-contained, with propelling charge, bursting charge, shrapnel (if any) all in one piece, commonly called a "shell." Many different types of shell were used in addition to the shrapnel and high explosive of 1914. Smoke and gas shells were used by most armies.

Conditions of trench warfare on the main battlefronts, and technical improvements, resulted in changes in artillery tactics. Guns had to be more dispersed and concealed and protected by earthworks and sandbags and their crews by deep dugouts. With the introduction of defensive fire and creeping barrages, guns had to be deployed differently and more attention devoted in defense to counterpreparation, to counterbattery fire and fire to disperse enemy troops during the process of concentration and forming up for attack.

In some of the secondary theatres, such as the Middle East, the new trends did not apply, and artillery was used more in the manner prescribed in peace for mobile warfare.

The introduction of tanks by the Allies on the Western Front in 1916 eased the role of their artillery, as two of the main tasks of the latter, crushing and making gaps in the wire and silencing machine-gun posts, were carried out increasingly by tanks. The reverse was the case for the Germans, who, until they produced their own tanks in the closing months of the war, had to deploy a proportion of their light artillery in an antitank role.

Trench warfare increased the importance of high-angle fire, leading to a revival of the use of mortars. Mostly small and of smooth bore, these were easy to produce. They were, however, very different in design from those of an earlier age. The Germans were the first to use these weapons extensively in World War I, but they were quickly followed by the British, whose three-inch Stokes mortar was the archetype of several produced later. In comparison with its small bore it had a relatively long (three feet, or one metre) barrel. The missiles were dropped down the barrel from the muzzle end, the fall of each round on the base of the barrel exploding the propellant charge. This enabled a high rate of fire for a short period; as many as six rounds could be in the air at one time.

During the early decades of the 19th century, naval gunnery had lagged behind land-based artillery in design of guns and gunnery technique. By the beginning of the 20th century this gap had been closed. It was realized in all major countries that the formidable battleships of the times made ideal platforms for modern guns. By the beginning of World War I, naval guns of up to 13.5-inch calibre existed, firing shells at ranges exceeding 10,000 yards. The fire of these weapons was controlled by what was known as "director fire," a device that calculated movement of target, movement of own ship, effect of wind and temperature, and even effect of wear-and-tear of gun barrels.

Battleships as gun platforms

World War I also saw major developments in the employment, particularly by the Germans, of torpedoes fired from submarines, and occasionally from destroyers.

WORLD WAR II GUNNERY

Except in anti-aircraft guns the interwar period saw little advance in artillery pieces or their projectiles; some advance was made in gunnery technique, and aids such as radar were introduced.

As a generalization it can be said that artillery and naval gunnery played a less important part than they did in 1914–18. The tank had partly relieved the gun of its destructive role of crushing wire and earthworks, and tactical bomber aircraft often relieved it of its bombardment role. At sea, aircraft played an increasingly important role, and in some naval operations—particularly in the Pacific in the latter stages of the war—the decision was reached by aircraft action before the opposing fleets had come within gun range or sight of each other.

Even on the Russian front the density of artillery did

not reach the heights of World War I. In some of the bigger battles the concentration of guns was considerable, but it was never equal to the vast arrays of 1916–18 with their huge expenditure of ammunition. The demand was for light, mobile guns and a consequent reduction in the heavier types. There were, however, a greater variety of artillery weapons, particularly in the field of anti-aircraft and antitank guns.

Except for two new types, few important changes were made in ammunition. The two exceptions were proximity-fuzed ammunition, an American product used by American and British forces, which, by means of a small electronic device, exploded in the air just before reaching the target; and armour-piercing ammunition, used by all major belligerents against tanks.

Toward the end of the war the Germans used two new types of long-range weapons, mostly against Antwerp, London, and southern England—these being known as V-1, flying bombs or pilotless aircraft, and V-2, high-angle rockets. These weapons were fired from static launching sites mainly in Belgium and The Netherlands.

By the outbreak of World War II the horse had been almost entirely replaced by motor vehicles (wheeled or tracked) for towing artillery. This, and the introduction of self-propelled guns (guns mounted on vehicles similar to tanks but less heavily armoured), greatly increased the mobility of all artillery.

Improved anti-aircraft guns

During the latter years of the interwar period, considerable progress was made in all major countries in the design of anti-aircraft guns, and this progress was greatly accelerated during World War II. For European countries, including Britain, there were two aspects of anti-aircraft defense—the defense of their homeland cities, towns, and installations and the protection of their warships by special anti-aircraft guns and their field armies by mobile anti-aircraft artillery. The weapons used for these purposes increased in size, velocity, range, and efficiency as the war progressed. For these roles the British used special guns of mostly 3-inch, 3.7-inch, and 4.5-inch calibre; the United States used the 90-millimetre and—the biggest of all—the 120-millimetre (4.7-inch) "stratosphere" gun firing a 50-pound shell to a height of 50,000 feet (15 kilometres). The Germans relied very largely on their multirole 88-millimetre gun—probably the outstanding gun of the war.

Other specialized forms of gun were those mounted on tanks, of small calibre when the war began but of much greater size and range by 1945.

Mortars, as an adjunct to artillery and mostly in support of infantry, were employed extensively by all countries in a variety of calibres. The Soviet Army used several heavy types, including one of 240 millimetres. The Germans developed a variety of types and tended to replace guns with mortars, particularly among their airborne troops.

Improvements in gun operation

Although the basic artillery weapons did not greatly change in World War II, great advances were made in the technique and methods of operating guns and gunlike weapons. The technique known to artillerymen as "survey" (employing map grid coordinates), enabled selected targets to be located in relation to the guns with great accuracy. This made the old method of actually firing trial shots at selected targets (known as ranging by registration) unnecessary, which in turn greatly increased the chances of surprise in offensive operations. The invention of radar enabled moving objects to be located, and their movements followed, long before human senses could detect them, and in the case of aircraft it was possible to distinguish friend from foe. This was of special assistance to anti-aircraft gunners. In addition, there were marked increases in the efficiency of radio communication.

These inventions, together with increased ranges and mobility, improved the flexibility of artillery. Whereas in World War I gunfire was largely confined to the particular sector behind which the guns were deployed, it now became possible to concentrate the artillery of one, or even more, divisions on a single small target area in a matter of minutes. It was common practice for a single battalion performing some special task, or in difficulties, to receive the support of the whole divisional artillery, without the necessity of redeploying the guns.

The heavier types of artillery, now much more mobile than in previous wars, tended to be concentrated under higher command but ready at short notice to proceed to any part of the battle area where they might be required. Typically a unit of this kind attached to an army or army group headquarters contained proportions of heavy, medium, and anti-aircraft guns and howitzers.

DEVELOPMENTS SINCE 1945

The explosion of two nuclear bombs on Japanese cities in August 1945 changed the whole conception of war in almost a flash. It was clear that any nation, or group of nations, with a monopoly of this weapon could impose its will on those not in possession of it. By the 1960s this situation had changed, with the two giants—the United States and the Soviet Union—reaching near parity in nuclear weapons and thereby creating a deterrent to nuclear war. Meanwhile, wars have taken place in many parts of the world—in Korea, Vietnam, and the Middle East on a large scale, and in many other places on a smaller scale —with conventional weapons mostly of World War II pattern.

Strategic and tactical nuclear weapons

Nuclear weapons. By 1970 there existed a situation in which all defense matters involving gunnery in its widest concept had to be considered in two tiers. First was the nuclear aspect, in which East and West faced each other with massive armouries of strategic nuclear weapons; some land-based, others seaborne, with ranges up to 8,000 miles (13,000 kilometres), any one of which was capable of causing devastation over a wide area. Strategic bomber aircraft, although still in service, were of diminishing importance and on the way out. There were tactical nuclear weapons with shorter ranges, up to about 410 miles (660 kilometres), which could be used for the bombardment of rear areas with low-yield missiles.

Neither strategic nor tactical nuclear weapons were used in any of the conflicts mentioned because of their two-edged nature and the realization of the indiscriminate destruction they would cause.

Conventional weapons. Conventional artillery and gunlike weapons, although overshadowed by the publicity given to the nuclear types, have remained in widespread use and have undergone considerable development either as new weapons or in improved versions of World War II types.

The principal major developments outside nuclear technology have been in guided missiles: surface-to-surface, surface-to-air, air-to-surface, and air-to-air.

Modern artillery and gunlike weapons

Up to the end of World War II, guns, and gunlike weapons, were one of many agencies that influenced the course of military and naval operations. In the 1970s the weapons at the top of the scale, the intercontinental ballistic missiles with nuclear warheads, do much more than that. The threat of these weapons has produced a nuclear deterrent; and not only do they influence the general policy of the greatest nations but their cost also affects national economies and the way of life of all mankind. This form of gunnery has greater influence on human affairs than almost any other single factor in the modern world.

At the other end of the scale, conventional guns and howitzers are being gradually superseded by short-range rockets, usually combined with a guidance system and known as guided missiles. This applies particularly to the heavier types of artillery, whose role is particularly suitable for the longer range missiles, with or without nuclear warheads.

Influence of guerrilla fighting

To a certain degree the advent during the past two or three decades of the well-organized and well-trained guerrilla fighter has contributed to the decline in the usefulness of conventional artillery. Even the lightest traditional artillery pieces are too heavy for the guerrilla fighter, and in close country he usually offers a target too small and too difficult for the regular gunners to locate. For both types of fighter the light and inconspicuous

mortar is often a better weapon. Thus although traditional surface-to-surface artillery still plays a part in modern war and will continue to do so into the foreseeable future, many of its roles are better suited to the new weapons. This is even truer in the surface-to-air dimension, where the guided missile has very largely superseded the anti-aircraft guns of the two World Wars.

Two points should be noted in connection with the survey of contemporary guns and gunlike weapons that follows. First, the various types of weapons mentioned are those in use in mid-1970; second, the names given to weapons of the Soviet Union are the code names used by the North Atlantic Treaty Organization (NATO) alliance.

TECHNOLOGY

Gun design and manufacture. The process by which sophisticated war material, including guns, is produced involves a number of stages, of which the following is a simplification:

1. General staff specification. This states the military requirement in some detail—weight, calibre, range, type of projectile, etc.
2. Drawing-board design.
3. Manufacture of prototypes.
4. Trial of prototypes with troops under varying conditions of terrain and climate.
5. Mass production (if trials prove satisfactory).
6. Issue to troops.

In most Western countries the production of a high proportion of artillery weapons is carried out by private firms, under close government supervision and sometimes with the government as a major shareholder. In the Communist countries all arms production is, of course, carried out in nationally owned factories. Some countries, of which Sweden is a notable example, make arms for export as well as for their own defense forces. Indeed, the export trade in arms is almost universal, partly to help allies and partly to reduce the cost by producing in economic quantities.

The number of private firms making guns and other war material is too numerous to list; some, such as Krupp in Germany and Vickers in Britain, have been designing and manufacturing guns and other war equipment for generations and have become household names.

Special problems in gun production

The production of arms, and gunlike weapons in particular, involves problems that are not usually encountered in most other manufacturing industries. In addition to being meticulously accurate, a gun must be simple to operate and maintain; it must be robust against rough treatment and severe weather conditions, but light enough to ensure maximum mobility.

But perhaps the most important factor is the quality of the material of which the gun is made. Armed forces continually demand longer and longer ranges and, in most categories of gun, ever-greater velocity, in order to reduce the missile's time of flight between the gun and a moving target. These improvements nearly always result in an increase in the propelling charge, which in turn makes it necessary to ensure that the metal of which the firing mechanism is made can withstand the additional shock of discharge. Because of the rough treatment to which war material is subjected, a good margin of safety must be allowed.

In practice under modern conditions the design and planning of a weapon is, almost invariably, entrusted to a team of experts who combine the necessary skills required for the project. The most important of these skills in the case of a gun, or gunlike weapon, are metallurgy and ballistics. Metallurgy is the science and technology of metals. Ballistics is the branch of applied physics concerned with the propulsion, flight, and effect on the target of missiles of all kinds (see BALLISTICS).

Although these two skills are the most important, there are many others, including the production and operation of computers to analyze and store data, and, in the case of weapons for which training with live ammunition is difficult or impossible, the production of simulators for training purposes.

It is apparent that the problems connected with rockets —in which the propulsion charge is contained within the missile—are more difficult and exacting than with a shell from a conventional gun. If a guiding device is added to the rocket the problem becomes even more complicated.

Projectiles. Until the time of the Napoleonic Wars the missile was comparatively simple compared with the gun itself. Today it is highly complex, and in some cases it far exceeds the launcher in intricacy of design and manufacture.

For the purpose of this article the term projectiles includes all missiles fired from guns or gunlike weapons.

There are two main classes of projectile—shells and rockets. A shell is a missile propelled entirely by an explosive charge fired within the launching agency (*i.e.*, the gun, howitzer, or mortar) and not repeated during flight. The characteristic of a rocket is that the propelling mechanism is contained within the missile itself and continues to operate during flight. This characteristic gives rockets much longer ranges than shells.

Each of these two groups of projectile is subdivided into many different kinds. Shells are high-explosive shrapnel (rarely used today), antitank, anti-aircraft, canister or case shot (still used occasionally in mid-20th century), chemical, smoke, nuclear, etc. Rockets include intercontinental ballistic missiles, medium-range and submarine-launched; and many types of short-range, or tactical, rockets.

There are also a few varieties of projectile that do not fall conveniently within the two above definitions, among them torpedoes and recoilless rifle ammunition.

In nontechnical language a shell consists of two parts—one containing the propelling charge and the other a bursting charge. A rocket contains an additional mechanism for boosting its propulsion during flight.

The explosives used in shells are of two kinds, the propellant charge and the bursting charge.

The ideal propellant charge should be smokeless and flashless to avoid detection and should have no tendency to absorb moisture. An explosive with all these qualities has been difficult to obtain; the one in most common use is known as "smokeless powder."

The requirement for the bursting charge is an explosive with great shattering effect that at the same time can withstand the shock of the propellant charge without exploding.

During the first half of the 20th century the most common fillers for high-explosive shells were TNT (trinitrotoluene) and amatol. The TNT was particularly satisfactory, but ordnance scientists continued to seek improvements. Among those produced were that called cyclonite by the United States and RDX by the British, which, with some special processing, was used extensively in World War II; haleite, named after a U.S. chemist; PETN (pentaerythritol tetranitrate); and pentolite, half TNT and half PETN, used for antitank shells.

Nuclear weapons have, of course, completely eclipsed in destructive effect all other explosives. As early as 1953 the United States demonstrated a 280-millimetre atomic shell with a diameter of about 11 inches and a range of about 20 miles (32 kilometres). The tactical atomic shell, with limited and manageable effect, for battlefield use, is now a part of the armament of several major armies, although its use is restricted by the deterrent effect of nuclear weapons in general.

Range-finding, guidance, and related systems. The object of gunnery is to propel a missile onto a selected target—enemy personnel or material objects such as buildings, earthworks, equipment, or industrial centres. The main technical consideration is the determination of the distance and direction from gun to target, the technique known as range-finding.

Early range-finding procedure

Until the second half of the 19th century, range-finding was done by direct and quite simple methods. The range was either estimated or established by means of an optical instrument known as a range finder; the sighting equipment of the gun was then adjusted to the range and aligned on the target, and the missile was sent on its way. Certain adjustments had, of course, to be made for wind, the known error of particular guns, and other factors.

After the first one or two rounds, corrections were made. The technique was much the same in naval gunnery as on land. The essential was for the gunner to see his target. If it was hidden by a hill, woods, or buildings, shooting was little more than guesswork. Except at close range, firing by night or in fog was impracticable and shooting at moving targets at long range very difficult.

The position is quite different today. Modern science and technology not only offer a variety of methods of range-finding but also provide missiles that can be guided to the target during flight or are attracted to it in the latter part of their flight.

In World War I improved maps and aerial photographs often made possible reasonably accurate shooting at unseen static targets. The lack of reliable air-to-ground radio communication made spotting from the air ineffective, however. The system known as registration ensured great accuracy in the conditions of trench warfare. By this means one gun could, by firing trial shots, ascertain the range of a number of prominent landmarks in a given area. From this information other targets could be judged with considerable accuracy, and other guns in the vicinity of the registering guns could make use of the same data. The drawback of this system was its slowness and forfeiture of surprise. Field telephones facilitated the employment of forward observing officers, and observers in captive balloons were used, but both were in danger from enemy fire.

During the interwar years (1918–39) three factors increased the efficiency of gunnery. First, the rapid development of air-to-ground radio, improving spotting from aircraft; second, the system known as "survey," enabling certain points in the target area to be plotted in relation to the gun positions; and, finally, improved line and wireless communications that, combined with survey methods, made it possible to concentrate quickly the fire of many guns on a single target.

All these methods for increasing the accuracy, speed in action, and general effectiveness of artillery fire were further developed in World War II, whose most important innovation, introduced late in the war, was the proximity fuze, which caused the charge to explode when the projectile reached the vicinity of the target. Since 1945 science has provided the gunner with many new devices, enabling targets to be engaged at vastly longer ranges and with much greater accuracy than formerly. Nevertheless, operations in the Middle East and Asia show that in the absence of sophisticated equipment, or in conditions in which it is impracticable to use it, the methods developed and employed so successfully in World War II are still practiced extensively.

The increased range of traditional artillery and gunlike weapons, and the introduction of rockets of much greater ranges, accentuated the difficulties connected with the identification and selection of targets and the methods of directing projectiles onto them. To a very great extent these problems have been solved by a series of new inventions and by the adaptation of older ones to new purposes. There has also arisen a classification of the types of target best suited to the increased variety of weapons. These are:

1. Strategic targets, long-range targets outside the operational area that can be engaged by weapons in the intercontinental ballistic missile (ICBM) class.

2. Battlefield rear-area or interdiction targets, up to a range of some 500 miles (800 kilometres) behind the battle area, such as centres of rail and road communications, areas of known troop concentrations, airfields, military depots, and other installations.

3. Battlefield support targets, engaged by traditional artillery types that include mortars and also some antitank guided missiles of post-World War II design.

In the case of (1) and (2) the question of "extreme" accuracy does not arise; the target is an area, not a point. The longer the range and more powerful the projectile the less important its accuracy. A strategic rocket with a nuclear warhead, fired at a range of 5,000 miles, may be a mile or more off the centre of the target and still have a shattering effect. There is no difficulty in calculating the exact range of targets for these weapons: the distance from any prominent place on earth to another is known, or easily obtainable by well-established methods, and the necessary data can be recorded in advance. Some doubt must exist, however, as to the accuracy of the missiles themselves, which, because of their power, complexity, and high cost, cannot be as readily tested as less sophisticated types. Their reliability must be assessed largely by the scientist's opinion and will depend largely on the accuracy of calculations relating to factors other than range and direction.

Accuracy of the strategic rocket

Similarly, weapons used against targets classified under (2) above, with ranges between about 85 and 500 miles, such as the United States Sergeant and Pershing missiles, present little problem as regards range-finding. With modern computing equipment, it is now routine practice to integrate quickly all the factors—range, wind and other atmospheric conditions, error of the weapon—to direct the missile onto what are mostly area targets.

Battlefield targets present a different problem: a high proportion demand pinpoint accuracy. An enemy tank, machine-gun position, or occupied building requires a direct hit or a very near miss. If the target can be seen, the range is quickly obtainable by established methods in the case of traditional guns and howitzers, or the target may be engaged by one of the wire-controlled guided missiles such as Swingfire (Britain), Dragon (U.S.), or Snapper (U.S.S.R.).

In addition to wire-controlled guided missiles, systems based on radar are employed to locate unseen targets at longer ranges and direct the missile toward them. "Homing" devices within missiles also exist; these attract the missiles to the target when they get within close range.

Carriages and mobility systems. Until World War I the problem of conveying artillery pieces to the battlefield quickly remained difficult. The march of armies was frequently delayed and operations postponed because of the slow forward movement of guns and ammunition. Horses, mules, oxen, and even camels and elephants were used to haul cumbersome pieces along unmetalled roads, or as pack animals for mountain artillery and other light guns capable of being broken down into pack loads.

During the 19th century the mobility of artillery appreciably increased, but mobility remained a problem and continued so until World War I. Such progress as was made was the result of the improved design of the mount or carriage on which the piece was fixed and which in reality became part of the gun itself. Teams of horses were able to move light pieces and, if the terrain was favourable, bring them into action at a gallop. But even in the early stages of World War I the movement of heavier artillery pieces was a slow and difficult matter.

By World War II the developments in mechanical vehicles in general, and tracked vehicles in particular, together with increases in the number of hard-surfaced roads, revolutionized the movement of all artillery. With modifications, and minor improvements, the position is much the same today. Guns and howitzers rely on two main systems for movement: mechanical towing vehicles, which usually also carry the ammunition and crew; and as self-propelled weapons—that is, mounted on the vehicle that propels them. Sometimes, but not always, the gun team and ammunition are also carried on the vehicle. The vehicle may be wheeled or tracked, the modern tendency favouring the latter. The earliest example of the self-propelled gun was one mounted on a tank, although its main purpose was, and still is, to protect the tank against enemy tanks while it carries out its assault role. In contrast, self-propelled guns perform the normal supporting role of artillery; the vehicles on which they are mounted are designed to enable them to perform this role and are less heavily armoured than tanks.

Systems for gun movement

Although self-propelled guns are becoming increasingly popular, and have many obvious advantages, they also have a higher silhouette and are consequently more difficult to conceal in action. Also, a mechanical breakdown immobilizes a self-propelled gun, not necessarily the case with a towed gun, whose vehicle can usually be replaced.

The strategic movement of artillery has been greatly facilitated by the use of modern aircraft. All but the heaviest pieces, including some self-propelled weapons, can be transported by air, and artillery is now part of the equipment of all modern airborne forces. Helicopters are capable of carrying light guns.

Despite modern mechanization there are occasions in mountain and jungle country when the only means of providing artillery support is with mountain artillery, broken down into light pack loads for horses or mules. Indian troops on India's northeast frontier are equipped in this manner, as are other armies likely to be involved in difficult terrain.

Essential as it is for modern armies to take advantage of mechanical progress, the advent of large numbers of mechanical vehicles, of which towed and self-propelled artillery and its supporting administrative vehicles form a considerable part, provides major staff problems. In large-scale operations, columns of mechanical vehicles many miles long are normal. All vehicles travelling in such columns must be capable of moving at approximately the same speed. A distance of 100 miles or more in a day is not uncommon, and at the end all personnel must be fed, vehicles replenished with fuel, and running repairs carried out. Staff work of a very high order is required if movements of this kind are to run smoothly. Failure in this respect caused considerable confusion in the German forces when Hitler's troops moved into Austria in March 1938; the lesson was noted by the general staffs of all the nations that went to war 18 months later, when the technique of modern road movement had been greatly improved.

Problems of movement

CLASSIFICATION AND FUNCTIONS

There are some gunlike weapons that were first developed for the close support of infantry and therefore classed as infantry weapons. In recent years, however, circumstances have arisen that leave the validity of this classification in doubt. These circumstances are partly technical and partly the result of changes in the pattern of warfare since 1945. The weapons concerned fall into two categories: recoilless rifles and mortars.

Recoilless rifles. At a very early stage in World War II it became apparent that a short-range, light, manhandled and one-man or two-man operated, antitank weapon was required. The British began the war with the Boys .55-inch antitank rifle, a weapon of the elephant-gun type. It was effective against lightly armoured vehicles but made little impression on German battle tanks. The first truly recoilless rifles were produced in the United States, first as a 57-millimetre bore weapon, later in 75-millimetre and 105-millimetre versions.

Today, one of the most notable weapons of this kind is the 84-millimetre infantry antitank gun that is in common use in a number of NATO countries. In Sweden, where it originated, it is known as the Carl Gustav. It is shoulder-controlled and capable of destroying any known tank at a range of 500 metres (1,600 feet). Its weight is 36 pounds (16 kilograms), and it can be carried and operated by one man; a second man is required to carry the ammunition and act as loader.

Guerrilla fighters in the Middle East and Asian theatres of war have used Soviet or Chinese weapons of this kind in recent years. The Israeli Army is equipped with 106-millimetre recoilless rifles mounted on jeeps.

Mortars. During the trench warfare of World War I, greatly improved mortars came into use, with much longer ranges, greater accuracy, and very much greater rates of fire. Improved methods of transporting these weapons and their ammunition resulted in their increased use in open warfare in World War II.

Conditions since 1945 have increased still further the usefulness and popularity of mortars and also have improved their performance. In the fluid conditions of European warfare, modern mortars often provide more mobile and less vulnerable weapons than guns or howitzers. In mountain and jungle country their high-angle fire enables them to reach targets inaccessible to other supporting weapons, and their light weight and inconspicuousness enable them to be brought into action quickly in difficult terrain. They are easily transportable by air, and they provide useful and early close support for troops carrying out an opposed landing from the sea. In some cases they can be fired from a vehicle. Moreover, they are inexpensive to manufacture, simple to operate, and robust and easy to maintain. With these advantages it is not surprising that sophisticated armies, particularly that of the Soviet Union and other Warsaw Treaty countries, should turn to mortars as an alternative to some of their light artillery. In the Soviet Army a considerable proportion of the divisional artillery has been replaced by mortars.

Advantages of mortars

The types of mortar in service in the various armies and guerrilla forces throughout the world are legion. All are improved models of the original Stokes mortar invented by the British in World War I. Typical of modern mortars is the United States's 81-millimetre mortar.

Light and medium artillery. The classification of "light" and "medium" artillery is basically the same everywhere, varying only slightly in detail from country to country.

Today, surface-to-surface guns and howitzers, classified as light, form the bulk of the artillery of a division or similar mixed force, such as a brigade group, combat group, or army task force. The improvement in the mobility of artillery pieces, by reason of tracked towing vehicles and the introduction of self-propelled guns, has, however, made it possible for weapons classified as medium to be added to the establishment of divisions and similar mixed forces in greater numbers than formerly.

The position today is that both light and medium artillery is usually available to support the close combat arms (armour and infantry) at very short notice. Additional medium artillery is normally retained under higher control as corps or army artillery, but readily available as a reinforcement in the combat area. Not all countries use the terms light and medium (*e.g.*, the British term for light normally is "field"). As a general rule, however, it can be said that pieces of a calibre up to about 105 millimetres (about four inches) are classified as light and those over 105 millimetres, but not more than about 155 millimetres (about six inches), as medium. All major armies have guns in these categories. Soviet medium artillery includes truck-mounted multibarrelled rocket launchers.

Some light and medium artillery pieces rely for movement on a towing vehicle, others are self-propelled.

There are a few categories of specialized artillery that come within the light classification; *i.e.*, artillery used solely in an antitank role, mountain artillery pieces that can be broken down into loads for pack animals or light aircraft, and artillery pieces especially designed for the use of airborne troops and conveyance by air.

The function of light artillery is to support combat troops in attack and defense, including antitank defense. That of medium artillery is the same, with the additional roles of counterbattery fire and bombardment of troop concentrations, installations, and communications in the immediate rear of the combat area.

Recently improved techniques and changes in the pattern of warfare have resulted in an increasing tendency to substitute mortars for light artillery as the close support of infantry.

Heavy artillery. Until the end of World War II the term heavy artillery applied to pieces of large calibre that were too immobile to accompany and support the leading troops in fluid operations. Heavy guns were reserved for static operations, such as those on the Western Front in 1914–18; for siege operations; and as fixed armaments in fortresses and coastal-defense installations. Since 1945, however, the term has been given a wider implication and may now be taken to include the various types of long-range strategic rockets.

The pattern of World War I, combined with the limited range and weight of shells of light artillery, gave rise to demands by all belligerent armies for heavy and extra-heavy artillery. Since 1918 there has been a progressive decline in this demand, for a variety of reasons, includ-

ing the development of bomber aircraft and rockets to take on bombardment roles and improvements in towing vehicles and self-propulsion that have made many heavy pieces classifiable as light from the point of view of mobility. In the 1960s it was found that the armed helicopter was often able to perform the harassing role that heavy artillery alone could carry out in the past.

Decline of heavy artillery

The defense of ports and fortresses by conventional guns, known variously as fortress, coastal, and garrison artillery, has now been almost entirely discontinued by the leading powers. Gone too are the cumbersome heavy guns and howitzers that took days to move a few miles and could only fire a few shots per hour. There are, however, a number of guns, howitzers, and rockets of large bore, and firing projectiles of considerable weight, that still form part of the equipment of the armies of both West and East. The conventional artillery pieces are either self-propelled or have mobile towing vehicles; the rocket launchers have been made reasonably mobile by similar means or have been given sufficient range to compensate for lack of mobility.

Anti-aircraft artillery. In modern military parlance the term anti-aircraft artillery covers a wide variety of weapons, all products of the 20th century and mostly of post-World War II origin. They range from rockets with a ceiling of over 100,000 feet (30 kilometres) down to highly mobile weapons for dealing with aircraft flying at treetop level.

Innovations in anti-aircraft defense

Two developments since 1945 have changed the basic pattern of anti-aircraft defense. These are the increased height at which modern aircraft can fly and the development of intercontinental ballistic missiles (ICBM's).

These two innovations have resulted in the production of the surface-to-air guided missile for dealing with aircraft at the higher altitudes, and the search for a practical antimissile missile.

It appears likely that in the foreseeable future the rocket missile will entirely replace aircraft as a means of long-range bombardment, and in consequence there will be an acceleration in the development of antimissile missiles. A foreseeable trend is the development of two distinct types of antimissile/anti-aircraft weapons: one to intercept and destroy long-range missiles, the other to deal with enemy aircraft in the operational or battle area.

There are many factors that make it impossible to classify or describe these weapons with precision. In addition to the secrecy that shrouds operational characteristics, new types are being produced at bewildering speed.

The most powerful and sophisticated weapons are deployed in the strategic-air-defense forces of the two major powers: the joint U.S.-Canadian North American Air Defense Command (NORAD) and the Soviet Union's Air Defense Command (PVO-Strany). These two organizations exist to defend their countries' homelands against attack by aircraft and to give warning and, to a limited degree, protection against attack by ballistic missiles.

Among the antimissile missiles are the Nike-Zeus (U.S.) and the Galosh (U.S.S.R.). A difficulty in the development of this type of weapon is the immense cost and the fact that, in order to be effective against a saturation (all-out) attack, they must be deployed thickly on the ground.

Air artillery. These weapons include two distinct classes, air-to-surface and air-to-air, the latter being in reality anti-aircraft weapons.

The projectiles usually associated with aircraft are the small-calibre machine gun for air-to-air and air-to-surface use, and the bomb for air-to-surface. During and after World War II a series of rockets was developed for use by normal aircraft and helicopters.

The short-range air-to-surface rocket is still in service, in both winged aircraft and helicopters, for harassing convoys of vehicles, troop concentrations, and other targets. The longer range, or "stand-off," type is in service for use with or without a nuclear warhead. Its purpose is to attack heavily defended targets from outside the zone of enemy defensive fire.

Rockets and guided missiles. The technique of guiding projectiles to a target is so closely associated with rockets as to make it convenient to consider them together.

From a technical gunnery point of view a rocket is a missile that is propelled by a rearward flow of hot gases generated in the rocket itself during flight, as distinct from the propellant charge of a conventional gun or howitzer shell, which ceases to operate after the initial explosion within the launching weapon.

A guided missile is one capable of receiving postlaunching guidance during flight. There are three main methods of guidance. Wire control is used for short-range surface-to-surface weapons within sight of the firer, a thin wire being trailed out to connect the missile and the operator. By this means the firer, who can see the missile in flight, directs it onto the target. Radio guidance can be used for longer range missiles. Finally, "homing" devices based on heat sensing attract the missile to the target.

Techniques of missile guidance

The classification of rockets and guided missiles is complicated by a number of factors. The design, technique for use, and even the existence of some is highly secret; some are strategic weapons with ranges of thousands of miles; some with shorter ranges are for battlefield use; most can use either nuclear or conventional warheads, but a few are confined to one or the other; they are in service with all combat services—land, sea, and air; and their organization and method of employment vary from country to country. (For information on specific rockets and missiles, see ROCKETS AND MISSILE SYSTEMS.)

Modern surface-to-surface rockets have three functions. The first is strategic, for the long-range bombardment of cities, military installations, docks, and similar targets at ranges from about 500 up to 8,000 miles (800 to 13,000 kilometres). Equipped with nuclear warheads, they are part of the nuclear deterrent maintained by East and West, to be used only in conditions of all-out nuclear war.

The second function is tactical, for bombardment within the operational area, which may be taken as up to some 500 miles (800 kilometres) behind the battlefront. They can be used with or without nuclear warheads.

The third is for battle support; for use, almost invariably in an antitank role, in support of front-line combat troops, and operated by a wire control system.

The primary purpose of surface-to-air rockets is to destroy enemy aircraft flying at heights beyond the range of conventional anti-aircraft artillery. Another role, of increasing importance, is the interception and destruction of enemy intercontinental ballistic missiles while they are still outside harmful range and altitude.

Surface-to-air rockets combined with guiding devices, particularly those designed for anti-ICBM purposes, are highly secret and subject to constant improvements, a process likely to continue indefinitely. One of the main limitations is the enormous cost of providing protection against ICBM's, even for a small area such as a port or military installation.

During World War II short-range air-to-surface missiles were introduced for action against such targets as columns of vehicles and troop concentrations. Since 1945 the emphasis has been on long-range missiles. The purpose of these latter weapons is to enable aircraft to bombard an important target with missiles, with or without nuclear warheads, from a distance outside the range of the enemy's defense system. These weapons are commonly known as "stand-off" missiles because the aircraft launching them "stands off" from the target.

Partly because of the high speed at which the launching aircraft is travelling, considerable difficulty has been experienced in devising satisfactory guiding systems for long-range air-to-surface missiles. Programmed systems have not proved feasible, and reliance has had to be placed on highly technical and delicate systems, such as the inertial-guidance system (see also GYROSCOPE).

The main purpose of air-to-air rockets and guided missiles is to engage enemy aircraft. A possible future role is against ICBM's, the aircraft engaging the missiles before they come within effective range of their targets; but much research and experiment will be necessary before this becomes a viable method.

The main characteristics of rockets and guided missiles, when used for war purposes, can be summarized briefly. The limit of range of the most powerful rocket (the ICBM) was in 1970 about 8,000 miles, but increases were anticipated for the near future.

Limitations of missile defense

The cost of rockets, especially the guided-missile variety, is prodigious. The cost of an anti-ICBM system protecting only a small area amounts to hundreds of millions of dollars. The high cost, combined with the long range of some types and their destructive powers, means that rockets and guided missiles undergo less rigid preservice tests than most other weapons. Although simulator training devices have been produced for many, personnel get little training with live missiles of the larger type. These limitations for training and testing would almost certainly reduce the expected effectiveness of many of these weapons in war. But the potentialities of the rocket and guided missile are immense and by no means exhausted. Development is likely to continue indefinitely.

(C.N.B.)

MODERN NAVAL GUNNERY

Data processing. The amount of information provided by the sensors carried by modern warships (such as radar, radio, sonar, and passive devices) is so extensive, and the armament and its associated fire-control equipment is so complex, that computers must be employed to sift the information and present it to the command in a way enabling prompt action by the appropriate weapons system. This computer system calculates the course and speed of any air, surface, or subsurface target encountered, and indicates the threat it poses to the force or ship. It determines the course to be steered for interception by the ship or by the aircraft she controls, and by means of converters known as decoders it can control the operation of missiles, guns, and torpedoes.

Missiles. The main armament of the modern warship is the guided missile, four types of which are mounted in ships: strategic, surface-to-surface, surface-to-air, and surface-(or submarine)-to-submarine. There are also the air-to-surface missiles with which carrier-borne aircraft are armed.

Submarine-launched ballistic missiles

The most formidable weapon in the naval armoury today is the submarine-launched ballistic missile. Developed initially by the United States as Polaris, the latest model, Poseidon, has a range of about 3,000 nautical miles and carries a multiple independent re-entry vehicle containing ten nuclear warheads targeted hundreds of miles apart. Each submarine carries 16 missiles fitted with inertial-guidance systems (see GYROSCOPE) linked through a computer, up to the moment of firing, to the submarine's inertial-navigation equipment, thus ensuring that it is continuously fed the correct coordinates of the target at which it is aimed. Some Soviet submarines are armed with ballistic missiles having a range of 750 miles (1,200 kilometres), but these are being superseded by those equipped with missiles of the Polaris type.

Until recently, intership combat, except through the intermediary of aircraft, appeared unlikely, but Soviet development of ship-to-ship missiles has recreated the possibility. Antiship missiles with a range extending beyond the radar horizon, which is 6 percent greater than that of the eye, require the cooperation of a satellite or aircraft to provide the necessary geographical coordinates or, alternatively, radar illumination of the target. The technique of using such missiles has not yet been fully developed, though doubtless such a system is employed for the ship-to-ship missiles with which many Soviet ships are equipped.

There are about ten types of close-range surface-to-surface missiles at sea, primarily designed for mounting in fast patrol boats. A typical missile of this type is the French Exocet, weighing 1,600 pounds (725 kilograms), which carries a conventional warhead with a shaped charge weighing 330 pounds (150 kilograms) with an impact equal to that of a 15-inch (380-millimetre) shell. It attains a speed of Mach 0.96 (just under supersonic; one Mach is about 700 miles per hour at sea level) and is guided by an inertial system to the vicinity of the target; then a nose-fitted radar takes over and ensures impact. This missile has a low trajectory, which makes it difficult to detect and intercept.

The ability of a warship to defend herself against air attack is of primary importance, but it is too costly to arm every ship with sufficient ship-to-air missiles in the way in which they were formerly equipped with anti-aircraft guns. A system of area defense is therefore employed. The first line of defense of a naval force is still interceptor aircraft controlled by the ships, and behind the zone in which they operate is that in which ship-to-air missiles are used. Although these are generally capable of dealing with any high-flying aircraft that elude the combat air patrol, against low-flying aircraft and missiles, close-range missiles and gunfire are desirable. Typical of the ship-to-air missiles in use is the British Sea Dart, in which the first stage is a rocket motor using a solid propellant and the second is a ramjet motor. The guidance system is semi-active homing; a radar receiver in the nose picks up the pulses reflected by the target when illuminated by the firing ship's radar, and it includes a means of detecting small changes in target movement. The missile also has a ship-to-ship capability. The whole weapons system comprises a search radar, two tracking radars, a twin launcher and automatic handling system, and a computer linked with the central computer system.

The close-range missile and gun defense zones are closely integrated. Short-range ship-to-air missiles are generally fitted with radio command or semi-active homing guidance. Those in service include the British Seacat (shortly to be replaced by Sea Wolf), the Swiss-made Sea Indigo, and the United States Sea Sparrow. An essential feature of their equipment is a stabilized target indication and control sight. These missiles are also used in an antimissile role against nonballistic and cruise-type missiles like the Soviet Styx and its successors.

Antisubmarine missiles

Surface ships and submarines designed to attack enemy submarines are armed with special antisubmarine missiles such as the Australian Ikara, the Norwegian Terne, the French Malafon, and the United States Astor, Asroc, and Subroc. Ikara and Malafon are rocket-assisted torpedoes using radio-command guidance. Terne is a rocket-propelled depth charge in which detonation is achieved by a combination of acoustic proximity, time fuse, and impact. Astor is a wire-guided torpedo with a range of 11 miles, while Asroc is an antisubmarine missile that can carry either a homing torpedo or a nuclear depth charge to a distance of about 6 miles. Subroc is also an antisubmarine missile with a nuclear warhead and a range of about 26 miles (42 kilometres) especially designed for launching from the torpedo tubes of a submerged submarine. On being fired it rises to the surface and, after an air trajectory to the vicinity of the target as determined by sonar, the propulsion and bomb units separate, the latter re-entering the sea and detonating.

There are a number of air-to-surface weapons with which naval aircraft are equipped, one of the most successful being the United States Bullpup, which carries a 1,000-pound (450-kilogram) warhead a distance of nine miles (14 kilometres). Longer range missiles are under development, among them the U.S. Condor, with a range of 35 miles (56 kilometres). Employing radio-command, visual, and television guidance systems, these missiles offer a serious threat to surface ships.

Guns. Guns are still needed in ships for shore bombardment in support of land forces, for close-range air defense, and for such peacetime duties as fishery protection. Apart from the 16-inch guns in four U.S. battleships now on reserve ("mothballed"), the largest gun afloat today is the eight-inch (203-millimetre) with which some U.S. cruisers are armed, a calibre considered to be the minimum for effective shore bombardment. But the high rate of fire necessary for anti-aircraft defense cannot be attained with a calibre greater than 5-inch (127-millimetre); the majority of modern ships are armed either with such guns or with 4.5-inch (115-millimetre) or 3-inch (76-millimetre) types. The essential function of target acquisition is assured by new types of stabilized radar, protected against electronic countermeasures and capable of

detecting low-flying aircraft. Tracking is done by radar assisted by computers. A fire-control system widely employed is the N.V. Hollandse Signaalapparaten of The Netherlands; another offering a wide choice of missiles and guns is Contraves Sea Hunter 4 of Switzerland. The efficiency of the modern gun as measured by the firepower-to-weight ratio is twice the level of World War II, and crew requirements have been greatly reduced. The Italian Oto Melara three-inch (76-millimetre) gun has a rate of fire of 90 rounds per minute, weighs 6.3 tons, and has a crew of three. It features fully automatic loading to ensure instant readiness, and reaction time is less than ten seconds.

Torpedoes. The torpedo, still a very effective antiship weapon, is today mainly used against submarines. The diameter varies from 22 inches (559 millimetres) to 12.75 inches (324 millimetres). As mentioned, various means are employed for airlifting torpedoes to the vicinity of their targets. Propulsion systems used include compressed air, electric cells, solid propellant, and liquid monopropellant. Guidance is active acoustic, passive acoustic, wire, or terminal acoustic, or a combination. In active acoustic a sonar transmitter in the head seeks out the target and homes the torpedo on to it; in passive acoustic a listening device in the head picks up the propeller noise of the target and homes the torpedo on to it; in terminal acoustic the torpedo follows a preset path to the target area and on reaching it circles or zigzags to and fro, the charge being triggered by an acoustic proximity fuse when the target comes within range. Guidance by means of a wire is also employed. This system enables the controlling ship to direct the torpedo towards the target. The torpedoes are launched directly from ships and submarines, by rockets, aircraft, and helicopters (see also NAVAL SHIPS AND CRAFT). (B.B.S.)

Propulsion systems

BIBLIOGRAPHY

Histories: H.W.L. HIME, *Origin of Artillery* (1915), a standard historical work; CHARLES OMAN, *History of the Art of War in the Middle Ages* (1924), useful for background information; A. MANUCY, *Artillery Through the Ages* (1949), a comprehensive history of artillery; B.P. HUGHES, *British Smooth-Bore Artillery: The Muzzle Loading Artillery of the 18th and 19th Centuries* (1969), a well-illustrated and comprehensive work.

Modern gunnery: Brassey's Annual: Defense and the Armed Forces 1970, ch. 17–19 (1970), an up-to-date description of developments in the sea, land and air equipment (including gunnery) of all major countries; INSTITUTE FOR STRATEGIC STUDIES, *The Military Balance* (annual), contains reliable data concerning the latest guns and gunlike weapons of all major countries.

Naval gunnery: B.B. SCHOFIELD, *The Royal Navy Today* (1960), background information, and "Developments in Maritime Forces," *Brassey's Annual* (1965–70), a yearly annual progress survey; W.T. GUNSTON, "Developments in Aircraft and Missiles," *Brassey's Annual* (1968–70); W.D. O'NEILL III, "Gun Systems? for Air Defense?" *Proc. U.S. Nav. Inst.*, 97:44–55 (1971).

(C.N.B./B.B.S.)

Guns, Sporting and Target

Sporting and target guns are the firearms used in hunting game and in target-shooting competitions and practice. Compressed air and gas weapons are fairly numerous in some places; firearms are much more popular. For hunting in general, see HUNTING, SPORT; for competitive target, trap, and skeet shooting, see SHOOTING. For general details about firearms, see SMALL ARMS, MILITARY.

HISTORICAL DEVELOPMENT

Firearms, introduced in the Western world in the 14th century, were originally military in purpose, but even the earliest small arms were undoubtedly used for informal matches and perhaps for taking game. Within 200 years weapons were being made specifically for hunting and target shooting.

Fine early muzzle-loading wheel-lock rifles can still be shot accurately by those who understand proper loading procedures. Some that survive were produced more than 300 years ago.

Flintlock fowling pieces, which made wing shooting (shooting at game birds in flight) practical, appeared in the 17th century, firing either small projectiles or shot pellets in fairly even patterns. Skillful hunters could kill game birds in the air, though range was limited. Many improvements were introduced to increase both range and accuracy. The Industrial Revolution brought factory production, followed by such 19th-century refinements as cartridge breechloaders; chokebores in shotguns; small calibre high-velocity rifles; and smokeless powder. Sporting and target arms are today among the most efficient and advanced of mechanical devices.

FIREARMS DESIGN AND OPERATION

Rifles. A rifle is a shoulder firearm that fires a single projectile or bullet and is usually classified on the basis of the type of action it employs and on the size or calibre of ammunition it fires.

Types of actions. Bolt-action rifles similar to military weapons of the 1890–1940 era are still the most common type for hunting throughout the world. Bolt action is efficient, reliable, and easy to manufacture. Most weapons of this type have magazines for quick reloading after each shot. When a bolt-action rifle is properly assembled, fitted to the stock, and supplied with appropriate ammunition, it can be extremely accurate. The disadvantage of all bolt-action weapons is that the total overall length for any given barrel is greater than "break-open" actions with one or more barrels. A double-barrelled sporting rifle can also deliver a second shot more quickly than any manually operated repeater.

There are three other types of sporting rifles. Semi-automatic hunting rifles are rare, save in Cal. 22 rimfire, but are gaining in popularity. Lever-action rifles were more common half a century ago than today. Slide or pump-action weapons are used largely by sportsmen who have grown accustomed to this type in their shotgun shooting.

Calibres. The diameter of the bore in inches or millimetres is generally called calibre. The full title of a rifle ammunition generally contains other information. For example, Calibre (or Cal.) .30–30 means a rifle with a bore diameter of .30 inch and a cartridge case designed to hold roughly 30 grains of black powder. Cal. 30–'06 indicates a .30 inch bore and a cartridge originally standardized in 1906. The new Calibre 6 x 47 means a bore 6 millimetres in diameter (.236 inch) and a cartridge case 47 millimetres long.

Bore diameter gives only a partial indication of power. The performance of a rifle depends on the weight and shape of the bullet and its velocity. A larger bore does not necessarily fire a heavier bullet; velocity usually decreases with increased diameter. For instance, a Cal. 257 Weatherby (the name of the inventor of the rifle and the cartridge) is considerably more powerful than weapons with larger bore diameters like the Cal. 30–30 because the Weatherby bullet travels faster.

Shotguns. A shotgun is a shoulder firearm that fires a cluster of shots or pellets, as opposed to a single bullet for a rifle.

Types of actions. In general, shotguns have actions similar to those of rifles, but "break-open" shotguns are far more numerous than "break-open" rifles. These are probably still the most popular type worldwide and can be made with one, two, or more barrels. A single barrel with this type of action is a single-shot weapon. Although cheap to produce and reliable, it has the disadvantage of an extremely slow second shot. Double-barrelled shotguns are most common. These are made in both side-by-side and over-and-under configurations. Three-barrelled and even four-barrelled "break-open" weapons are rare; they usually include one or two rifle barrels.

Double-barrelled shotgun

Popular in America for many years, slide or pump-action shotguns have gained favour in Europe. To chamber a new round after each shot, the fore-end is pulled smartly to the rear as far as it will go and then pushed forward again. A hunter experienced with a pump shotgun is not conscious of operating this mechanism and can give his entire attention to shooting. In general, weap-

ons of this type are the cheapest and most reliable of the popular repeaters.

Semi-automatic shotguns have been made since the early 20th century; early weapons of this type were the long recoil Brownings. Since 1945, gas-operated shotguns have become more popular. In these weapons, the pressure of the gas generated by the fired shell ejects the spent shell, cocks the action, and moves a fresh shell into the firing chamber. Gas-operated shotguns are now considered reliable and effective on both game and clay targets. All semi-automatic shotguns require more careful maintenance and cleaning than similar pumps. "Automatic" shotguns tend to absorb part of the recoil rather than communicating it to the shoulder of the shooter in the form of "kick."

Lever-action and bolt-action shotguns have been produced, but neither type is popular. The inherent defects of all bolt-action weapons are excessive length and slow second shots. The lever-action shotgun never became popular because it was clumsy in appearance and operation.

Gauges. Shotguns are made in several standard barrel or bore diameters. The bore or gauge number designation originally indicated the number of spherical lead balls of bore diameter that weighed one pound. The derivation of the term accounts for the paradox of having a smaller number indicate a larger weapon. There are five standard-gauge bores in use today: 10-, 12-, 16-, 20-, and 28-gauge. There is also the so-called .410, which borrows its designation from rifle nomenclature; it is a calibre .410, often described as .410-gauge. The 12-gauge shotgun is the most popular in America and Britain; elsewhere the 16-gauge is often preferred. Other sizes are far behind everywhere. The 10-gauge is now obsolescent.

A shotgun usually fires a charge of spherical pellets of uniform size. By trial and error, sportsmen and target shooters have established which sizes are best for various types of shooting. Though large shot carry farther, their pattern is less dense because there are fewer pellets per given weight of charge.

Chokes and patterns. A shotgun's effectiveness depends on the uniformity of its pattern. A good weapon places its charge evenly in a circular pattern on a surface perpendicular to the line of flight. Density of the pattern depends on the number of pellets in the charge and the distance of the muzzle from the target. The farther the patterning surface from the muzzle, the less dense the pattern. For quick shooting at close range, a pattern of maximum size and minimum density is desired. Normally, a shotgun with a uniform cylindrical bore (the same size from breech to muzzle) is used to provide such a pattern.

Shot patterns

For longer ranges, a tighter pattern is desirable. Such concentrations of shot are obtained by reducing the diameter of the bore close to the muzzle (choking) in various ways. Full chokes effect maximum concentration of normal charges concentrating 70 percent or more of the shot in a 30-inch circle at 40 yards. A modified choke is about halfway between a full choke and a cylindrical bore; ¼ and ¾ chokes exist. Great skill is required to hit fast-flying birds at long range regardless of the choke of the shotgun. Any reduction in pattern size handicaps a shooter at close range; full chokes hinder more often than they help.

Any single-barrelled weapon can be equipped with a device attached to the muzzle that allows a shooter to change his choke at will, either by attaching a new tube or adjusting the muzzle diameter. These can be combined with recoil-reducing buffers that turn some of the muzzle gases to hold the weapon where it is, rather than allowing it to recoil.

Handguns. A handgun is a small firearm that can be held in one hand. It usually fires a single projectile or bullet.

Types of action or operation. Single-shot pistols are still used, though they are less popular than revolvers or semi-automatic pistols. Full-automatic pistols have been made but are illegal in many countries. In fact, private ownership of pistols of any sort is restricted throughout a good portion of the world.

Target. Where more than one shot must be delivered within a time limit, revolvers or "automatics" are preferred. A top-quality target pistol can place five shots in a 2-inch (50 millimetre) group in 20 seconds at 25 yards (24 metres), although only a few competitors have the skill to utilize this accuracy.

Sport. Many sportsmen throughout the world carry a small pistol with them when hunting and fishing, generally Cal. 22 rimfire weapons, not practical for use on game larger than grouse and rabbits. On the other hand, powerful pistols have been manufactured for centuries for use by hunters as secondary weapons, even against dangerous animals. Typical are the heavy revolvers that take Cal. 44 Magnum ammunition. Weapons and cartridges of this type are more powerful than many rifles. Some men hunt with pistols only and equip their handguns with telescope sights. Such pistols are satisfactory for medium-sized game, including deer, if the hunter has sufficient skill to place his bullets accurately.

Hunting with pistols

Self-defense. Occasionally, a hunter in primitive territory carries a pistol for self-defense against other human beings. Revolvers of medium to high power are popular for this type of service. Though a pistol is more likely to be at hand when required than a rifle, it takes far more skill for effective use. Pistols are also purchased throughout the world by homeowners and others for personal defense, despite legal and other risks.

SIGHTS AND STOCKS

Rifle. The portion of a rifle normally made of wood is known as the stock. It usually consists of a single piece extending from the butt to the fore-end in bolt-action types or of two pieces in most other weapons. The stock supports the barrel and action and temporarily joins the entire mechanism to the shooter. A rifle stock is designed so that when held against the shoulder it permits the shooter to sight along or through the rifle sights. Stocks vary widely, however, depending on the use for which the rifle is intended, the type of sights installed, the size of the user, and the desire for ornamentation. Though not often embellished with gold, silver, or ivory today, stocks are frequently carved in various ways and finished with resins and plastics of unusual variety and beauty. There are even some laminated stocks made of thin pieces of different woods chemically bonded to each other.

Iron sights are sights in which no glass is used. The front sight can be a blade, a bead, or an aperture surrounded by metal. It can be seen through or over a rear sight that may have an open "V" section, a "U" section, or a small, round hole called a peep. Various combinations are used for both hunting and target shooting.

Special telescopes are also installed on rifles to replace or supplement iron sights. Scope sights magnify the target and indicate precise location of the bullet's point of impact by "cross hairs" or a similar device. Telescope sights add to the effectiveness of semiskilled marksmen and provide additional minutes of effective shooting because of their light-gathering power early in the morning and late in the evening.

Telescope sights

Shotguns. Shotgun stocks resemble those of rifles; fit is even more important. Shotgun sighting equipment is less complicated, however, because aiming is less precise. A shotgun normally has only a bead front sight, though a few now have additional small beads located about the middle of the barrel. Sighting is accomplished by looking along the barrel and placing the front bead on or under the target. This is usually done quickly and to some extent intuitively. Both normal iron sights and scope sights have been tried on shotguns, but without much success.

Pistol. Pistol stocks or grips are designed to fit men with normal-sized hands. Target shooters often adapt them to their individual requirements. Pistols have a bead or rectangular front sight and an open rear sight with either a "U" or a rectangular opening. Most target weapons have adjustable rear sights. Techniques have been worked out for precise shooting in which the front sight

is seen through the opening of the rear sight with the tops of both level. The target is then "placed on" the front sight. Since the human eye can see only one thing at a time clearly, both the rear sight and the target are slightly blurred. Telescope sights are occasionally mounted on pistols but are not effective for target shooting. They are useful in hunting, especially with powerful weapons.

AMMUNITION

Rifle and pistol. Many different types of ammunition have been developed, each for specific purposes. Rifle and pistol ammunition is frequently rated on the basis of effectiveness, accuracy, and range.

Game effectiveness. The effectiveness of a hunting rifle depends on where the bullet hits the game and its power or energy at the time it strikes. Both are important. A bullet's power or energy varies directly with its weight and the square of its velocity. Weight remains constant throughout flight, but velocity decreases rapidly with increasing range.

Big-game ammunition

To kill large game normally requires at least 2,500 foot-pounds of bullet energy. This can be achieved by a 200-grain bullet travelling at 2,500 feet per second (fps), by a 300-grain bullet travelling at 1,960 fps, or by a 500-grain bullet travelling at 1,700 fps (a grain is $\frac{1}{7000}$ of a pound). These velocities and bullet weights correspond to projectiles from standard Cal. 30 rifles out to about 100 yards, Cal. 375 rifles out to 175 yards, and Cal. 45 rifles out to about 300 yards. High-intensity "Magnum" cartridges like those in the Weatherby series extend power for any given calibre at least one class; the Cal. 300 Weatherby Magnum loaded with a 180-grain bullet is almost as powerful as a Cal. 458 Winchester. The 460 Weatherby Magnum has more power than the Cal. 577 British Nitro Express cartridges.

Large bullet energy increases hunter security when going after dangerous game on foot at close range because of its greater knockdown capability. More bullet power, however, means more recoil to be absorbed by the firer and can lead to inaccurate and less effective shooting than having less power more accurately placed.

Accuracy. High accuracy is essential for hunting at long range and for target competition. Cartridge-case design plays an important part; so does the shape and weight of the bullet. Both the ammunition and the rifle must be manufactured carefully, in accordance with principles of quality worked out over the years. Bullets must be uniform in weight with solid concentric jackets and core. The powder charge from round to round must not vary more than $\frac{1}{10}$th of a grain. Contact between the action of a rifle and the stock should be even and solid. Heavyweight barrels appear to shoot better when free-floating; *i.e.*, anchored firmly in the action but not touching the fore-end.

Ammunition accuracy is equally important for pistol competition. Top-flight target shooters continually try new combinations of weapons and ammunition, varying bullet shapes, weights, and materials. Powder and primers are also important.

Range and noise. Maximum effective range is determined for rifles and pistols by the velocity of the bullets they fire. The higher the initial velocity the farther the bullet travels before it loses accuracy and target effectiveness. Maximum power rifles like the Cal. 460 Weatherby and the Cal. 577 Nitro Express are seldom fired at long range, whereas some weapons designed primarily for varmint (small animal pest) shooting, especially the high-velocity Cal. 22 centre-fires, are rarely shot below 50 yards. The latter type are accurate out to at least 200 yards and are often fired at targets much farther away.

Noise is also directly related to rifle power. Basically, the more power the rifle has, the more noise it makes. Many small-game hunters in the more heavily populated areas of the world use Cal. 22 rimfire rifles because they fire light bullets at low velocities and make less noise.

Shotgun. Shotgun ammunition varies in effectiveness, maximum range, and evenness of pattern formed by the pellets.

Killing power and range. Some of the factors in effective killing power of rifles apply to shotguns as well. The weight of projectiles hitting a target determines effectiveness. To some extent, velocity at the time the game is struck also plays a part, although shotgun pellet velocity does not vary widely in standard loadings.

The smaller the shot size, the more shot in a given charge and the thicker the pattern at a given range. But small shot have less power than larger pellets and do not penetrate the feathers of larger game birds as well. In general, smaller shot are used for smaller game birds. Small shot loses velocity and effectiveness more quickly with range than larger shot. Ducks and geese, which require more penetration and are normally shot at medium to long range, require larger shot. Shot sizes vary from country to country, but larger numbers indicate smaller pellets. In the U.S., #8 and #9 are for quail, #6 and #7½ for pheasant, and #2s, #4s, and #5s for waterfowl.

As a practical consideration, shooting at what a hunter considers moderate range is more rewarding than when he believes he is just inside maximum effective range. There are two reasons for this. First, the shooter usually underestimates the distance to birds in flight. Second, almost all hunters overestimate the effective range of their combination of weapon and ammunition.

Shooting at moderate range

Even patterns. Patterns vary with chokes as discussed above. They also vary with ammunition. Distribution of the total shot charge varies slightly from round to round in any given gun even with the same type cartridges. Recent improvements in ammunition, including elimination of the overshot wad (the wad that is placed over the shot), and surrounding of the shot charge with a plastic cup while in the barrel, have increased the number of undamaged shot that fly truly out to maximum effectiveness. This means a thicker pattern.

USES OF GUNS

Local hunting. Only a few areas in the world still have wild game available close to large concentrations of people. Even when game exists in populated areas, hunting is frequently illegal because of the danger to humans and domestic animals. In some parts of western Europe, wild game hunting is still possible close to cities because of controlled conditions that generally include feeding the game during a portion of the year and careful protection of them from their natural enemies. In these areas, hunters must take courses in game recognition, firearms, field safety, and other similar subjects. This preparation greatly reduces the incidence of unsafe shooting.

In other areas like the Rocky Mountains of the U.S. and Canada, and in Australia and New Zealand, human population density is low and game population high. A farmer or rancher may even kill animals because they destroy his crops or take grass from his cattle. Game birds have been so plentiful in New Zealand that normal hunting did not keep their numbers within bounds. Many wild geese were destroyed with dynamite.

Game farm and distant hunting. Hunters who do not wish to travel long distances often pay for the privilege of shooting in nearby areas where game is preserved. Guides and outfitters in the Western U.S., Canada, and Alaska frequently charge high fees but may guarantee success against big game or at least guarantee a fair shot at an animal of a desired species.

Preserved hunting of game birds is increasingly popular in America and Europe. Game on preserves is usually partially wild, supplemented by pen-raised birds. They may or may not be varieties that were once wild in the particular area. Chukars, or Hungarian partridges, are now found in many countries besides Hungary.

Competitive shooting. A major area of gun use is in competitive or target shooting. Rifles, pistols, and shotguns are so used.

Rifle. Man has engaged in target competition with military and sporting weapons since the dawn of history. As archery declined, firearms became important. Olympic rifle shooting is only one of at least a dozen types currently popular throughout the world. Rules, weapons,

time limits, ranges, positions, and other details vary widely.

Rules of rifle competition

Rifle competition usually resembles either military or big-game shooting or a combination of the two. In Britain and America, the major competitions are conducted with rifles similar to those used in the Armed Forces at different ranges. Firing positions resemble those used in the military. Rapid fire, that is, the firing of more than one shot within a limited time, is included. Olympic rifle competition, on the other hand, includes no rapid fire.

A new type of competition that tests the ability of a competitor to collect accurate equipment more than his marksmanship has recently come into being. Matches are fired from shooting benches where the gun is clamped under conditions that approximate laboratory testing of weapons and ammunition. Some of the rifles used in these matches are not even practical for game hunting or normal target competition. Ten shots at 100 yards may be fired into an area smaller than the U.S. ten-cent piece.

Pistol. Pistol matches are always fired without support for the weapon or the shooter's arm. A complete match generally includes deliberate or slow fire and one or two forms of timed or rapid fire. Weapons used, trigger pulls, sighting radius, and other details are sometimes, but not always, controlled.

Shotgun. In the two worldwide popular forms of shotgun matches (trapshooting and skeetshooting), shots are fired at artificial or "clay" targets thrown by machine in flight patterns designed to resemble those of birds. These targets break when hit by shot and are scored as "dead."

Live pigeons and occasionally other game birds are also shot competitively in some areas of the world. Live bird competition rules normally require a shotgun that delivers two shots quickly in pattern of maximum tightness. Since a released bird may fly directly away from the shooter, a competitor needs a weapon effective at long range. Kills at 45 yards and beyond are not unusual.

GUNSMITHING

Guns, like most things made by man, require maintenance. Small parts break; complicated mechanisms need adjustment. Even more important, commercially available weapons may not give maximum effectiveness for specialized types of shooting. A gunsmith familiar with rifles can generally improve the accuracy of a weapon.

Competent gunsmithing can be even more important for competitive pistol shooters. Military semi-automatic pistols vary widely in accuracy as manufactured, but when worked on by a knowledgeable mechanic, shot group size can be reduced dramatically. Gunsmiths can also install special sights and stocks.

Shotguns often need changes in stock, triggers, and shot patterns, in addition to normal adjustment, repair, and cleaning. Considerable skill is required here, but competent gunsmithing can improve a shooter's chances of hitting his target.

Gunsmithing at home

Many shooters throughout the world perform some or all of their gunsmithing chores for themselves. They may even make weapons or parts of them to their own specifications. There is ample literature in this connection, as well as tools, including specially designed machine tools.

Most amateur gunsmiths and many other shooters handload some or all of their ammunition, usually by refilling used cartridge cases, the most expensive component of a complete round. Equipment varies greatly in simplicity, cost, and effectiveness, but skillful handloading can improve on factory ammunition accuracy for both rifles and pistols. Reloading reduces the cost of trapshooting by about 40 percent.

COMPRESSED AIR AND GAS GUNS

Guns resembling firearms but using compressed air to project either a single bullet or a charge of shot relatively short distances have been used for hundreds of years. Originally, a tank in the stock of the gun was pumped full of air. A trigger released this suddenly into the barrel behind the bullet or charge of shot. Range, accuracy, and power were always limited, but the weapon was nearly ideal for a hunter who wanted to avoid making a loud noise and was able to creep up on game, especially at night.

The compressing of air into a reservoir was always a problem. This form of air gun was good for only one discharge and was costly to produce. During the 16th century a spring was substituted for the air reservoir. When the trigger released the spring, the latter actuated a piston that compressed air that in turn drove a pellet through the bore. This is the principle used by most air "rifles." It can also be used in air pistols.

More recently, weapons have been constructed that use cylinders of compressed gas, usually carbon dioxide. A single cylinder will give a number of shots before replacement is necessary.

Air or gas weapons are sometimes used in military training and for informal indoor shooting, especially by family groups. An air rifle can stun or even kill small game, but accuracy, range, and power are all limited. On the other hand, an air rifle is not a safe plaything. A pellet that hits a person in a spot covered by two or three thicknesses of clothing will sting only, but a shot in the eye or some other sensitive part can have tragic results. Weapons of this type are illegal in some areas.

BIBLIOGRAPHY. Information on the small arms of most major nations may be found in W.H.B. and J.E. SMITH, *Small Arms of the World*, 7th rev. ed. (1962). Works dealing with specific types of guns include: SIR GERALD BURRARD, *The Modern Shotgun*, 3 vol. (1950); J.S. HATCHER, *Textbook of Pistols and Revolvers* (1936); J. O'CONNOR, *Complete Book of Rifles and Shotguns* (1961); and P.B. SHARPE, *The Rifle in America*, rev. ed. (1958). General information may be found in J.V. HOWE, *The Modern Gunsmith*, rev. ed., 2 vol. (1941). See also *The American Rifleman* (monthly), published by the National Rifle Association; and *Shooter's Bible*, an annual devoted to guns and hunting.

(J.We.)

Gustav I Vasa, of Sweden

Gustav I Vasa (known before his coronation as Gustav Eriksson), one of the great rulers of the 16th century, made Sweden an independent state; gave his country, for the first time in a century, nearly 40 years of stable and intelligent government; ensured the triumph of Lutheranism; established the first truly national standing army of modern times; and founded the Swedish navy.

By courtesy of the Svenska Portrattarkivet, Stockholm



Gustav I, portrait after J. Binck, 1542. In the University of Uppsala, Sweden.

Early life. Gustav was born in 1496 (?), the son of a Swedish senator and of a noble family whose members had played a prominent part in the factious aristocratic politics of 15th-century Scandinavia and that was connected by marriage with the family of Sture, which had supplied Sweden with three regents. Gustav fought in the

army of Sten Sture the Younger against Christian II of Denmark in 1517–18 and was one of the hostages sent by Sten to Christian in 1518 as part of the terms of an armistice. Christian violated the agreement and carried Gustav off to Denmark. In 1519 Gustav fled from his captivity to Lübeck, where he made friends who were to be of great importance later. On May 31, 1520, he returned to Sweden. Sten Sture had meanwhile died of wounds, and Christian was master of almost all Sweden save Stockholm. In November, by the "Stockholm bloodbath," Christian removed the most dangerous of his opponents, including Gustav's father and two of his uncles.

Rise to prominence

Faced with the alternative of rebellion or flight, Gustav chose the former. He succeeded in rousing the midland province of Dalarna to resist, purchased by judicious concessions the support of lay and ecclesiastical magnates to whom a union of the three Scandinavian kingdoms under Christian had become unwelcome, and was able (since Sten Sture's son was a mere boy) to pass as leader of the surviving Sture party. A considerable body of folk legend deals with his real and supposed adventures at this period. For the eviction of the Danes, as he soon found, outside help was necessary; and he obtained it from the rich free city of Lübeck, whose merchants felt themselves threatened by Christian's aggressive economic policies. This aid enabled Gustav to establish Sweden's independence and may have been responsible for his election as king (June 6, 1523). In return for it, Lübeck extorted far-reaching commercial privileges; and it was to be one of Gustav's main concerns to emancipate his country from its dependence on his former backers.

His monarchy. Gustav's crown continued for some years to be precarious. Christian II had been driven out of Denmark by his uncle, who succeeded him as Frederick I, and a common fear of Christian's restoration soon drew Frederick and Gustav together, so that despite recurrent periods of tension the threat from Christian, and afterward from his heirs, enforced a measure of harmony between Sweden and Denmark. But Gustav had to face serious internal dangers: from aggrieved members of the old Sture party who resented his favour to some of their former enemies; from the men of Dalarna, who added to this grievance complaints on economic and religious grounds; and from great nobles, who found Gustav a more formidable ruler than they had expected. Indeed, Gustav proved to be a harsh master and an exigent lord; he became known for being suspicious, mendacious, cruel, vengeful, demagogic, and capricious; and to his enemies, he seemed to have most of the attributes of a tyrant.

Role in the Swedish Reformation

The need to pay his debts to Lübeck, and to strengthen the royal authority, forced Gustav to impose heavy taxes; and it was essentially with a view to tapping the church's wealth that he embarked on the measures that led to the Reformation in Sweden. The Diet at Västerås in 1527 put the church's property at his mercy. Gustav had few theological interests or preferences; but he resented the presence in Sweden of any authority that challenged his own, and he had some sympathy with the idea of services in Swedish, for he was an indifferent Latinist himself. The move toward Lutheranism, however, was both accelerated and retarded by purely political considerations. Sweden did not become irrevocably a Lutheran country until 1544 at the earliest, and it was a long time before Protestantism was popular outside Stockholm.

The last great revolt of the reign, in 1542–43, had a strong anti-Protestant strain. Gustav's vain attempts to become a member of the Schmalkaldic League, formed by the German Protestants, were dictated by a desire to provide himself with allies rather than by religious convictions. In foreign policy, indeed, he inclined always to caution and a husbanding of resources. If he intervened in the so-called Count's War between pretenders to the Danish crown (1534–36), it was because he saw at last a chance of liberating Sweden from Lübeck's tutelage; and his only other adventure was the later war with Muscovy (1555–57).

Gustav's greatest achievement was the creation of a strong monarchy. He based his power on a massive agglomeration of crown and family lands, acquired for the most part by confiscation from the church, which put him beyond the rivalry of any other noble house. The supervision and exploitation of these lands was his personal concern, and with it went an infinite solicitude for the least detail of fiscal policy. In the 1540s, due to a lack of trained Swedes, he imported German administrators. This was a brief episode, but their work had a lasting effect in Sweden. It enabled Gustav to maintain his personal supervision and to combine it with high efficiency.

As the political heir of a faction, he found it expedient to bribe his nobility with church lands, and he was successful in many policies that the Stures had only attempted. In 1544, for instance, he induced the Diet to declare the monarchy hereditary rather than elective. He summoned the estates frequently in the uncertain years at the beginning of his reign, though less often thereafter; and his use of them to endorse his policies undoubtedly aided their development as an effective parliamentary body. On the other hand, he reduced to a position of relative insignificance the aristocratic council of state, which had played the leading part in the constitutional struggles of the preceding century.

He died on September 29, 1560, and was succeeded by his son, who became Eric XIV.

BIBLIOGRAPHY. CARL GRIMBERG, *Svenska folkets underbara öden*, new ed. (1959), a colourful but uncritical national history; *Gustav Vasa minnen* (1938), a splendid photographic collection of the personal remains of the King; INGRID HAMMARSTROM, *Finansförvaltning och varuhandel, 1504–1540: Studier i de yngre Sturarnas och Gustav Vasas statshushållning* (1956), an epoch-making study of how a rustic society achieved military and political prominence; SVEN LUNDKVIST, *Gustav Vasa och Europe: Svensk handels- och utrikespolitik, 1534–1557* (1960), a reliable work; INGVAR ANDERSSON, *Sveriges historia* (1943; Eng. trans., *A History of Sweden*, 1956), ch. 13–14 on Gustav Vasa.

Gustavus II Adolphus, of Sweden

Gustavus II Adolphus, king of Sweden from 1611 to 1632, inherited a weak and divided country and transformed it into one of the greatest powers of all Europe, by creating a Swedish Empire of the Baltic that endured for almost a full century. His intervention in the Thirty Years' War (1618–1648), at a moment when the armies of the Habsburg emperor and the German princes of the Catholic League controlled almost the whole of Germany, ensured the survival of German Protestantism against the onslaughts of the Counter-Reformation. The consequences, for Germany and for Europe, extended far beyond the religious field. By supporting the German princes against the Emperor, Gustavus Adolphus defeated the attempts of the Habsburgs to make their imperial authority a reality, and thus played a part in delaying the emergence of a united Germany until the 19th century. As a military commander, he was responsible for military innovations that marked an epoch in the history of the art of war. But from the point of view of his own country, these achievements were less significant than his domestic labours—his extraordinarily wide-ranging creative work in the fields of administrative organization, economic development, and education. If Gustavus Adolphus had never won a battle, or conquered an inch of territory, he would still rank as the greatest of Swedish kings, and one of those who were largely responsible for laying the foundations of the modern Swedish state.

Problems of the monarchy

Gustavus was born in Stockholm on December 9, 1594, the eldest son of Charles IX and his second wife, Christina of Holstein. He was still some weeks short of his 17th birthday when he succeeded his father in 1611, and it was only in exchange for important constitutional concessions that the Swedish Estates (the Riksdag or Assembly) permitted him to assume full control of the government. He found himself in an extraordinarily difficult position. Charles IX had usurped the throne, having ejected his nephew Sigismund (who was also king of Poland) in 1600, and the resulting dynastic quarrel involved Sweden and Poland in a war that continued inter-



Gustavus II, portrait by Matthäus Merian the Elder, 1632. In Skokloster, Uppland, Sweden.
By courtesy of the Svenska Portrattarkivet, Stockholm

mittently for 60 years. Until 1629 Gustavus had always to reckon with the danger of a legitimist invasion from Poland and the attempted restoration of the elder Vasa line. Charles had also begun a war in Russia in an attempt to put forward a Swedish candidate for the vacant Russian throne, and then, when his armies were deeply committed in Russia, had rashly provoked war with Denmark. Not only had Charles placed Sweden in a calamitous situation internationally but he had left behind him a legacy of domestic troubles. His usurpation of the throne had meant not only expulsion of a Catholic sovereign whose rule seemed to threaten Sweden's Lutheranism, but also the defeat of the aristocratic constitutionalism of the Council of State, and it had been followed by the execution of five leading members of the high aristocracy. Charles's rule had been arbitrary and violent; his religious views (he was suspected of leaning toward Calvinism) had involved him in an incessant struggle with the Lutheran church. At his death the country was exhausted by constant warfare, the monarchy was generally unpopular, and the accession of a new king seemed to offer the opportunity to extort from the crown guarantees against a recurrence of misgovernment.

Resolution of foreign wars. Thus, in 1611 Gustavus had three foreign wars and a major constitutional crisis upon his hands. As the war with Denmark was as good as lost, he set about to end it on the best possible terms.

Peace of Knäred

By the Peace of Knäred (1613) Sweden was forced to leave its only North Sea port, Älvsborg, in Danish hands as security for the payment of an enormous war indemnity. That indemnity entailed crushing taxation and, even with the aid of last-minute loans by the Dutch, was not paid off until 1619. The war left bitter hatred behind it, and Gustavus never forgot that Denmark was the national enemy and might be expected to take advantage of any Swedish weakness. Meanwhile, the war with Poland remained largely in abeyance, although in 1617 Gustavus sent an abortive expedition to seize the fortification of Dünamünde outside Riga (in present Latvian S.S.R.). The main danger, however, seemed to be Sigismund's attempts to pursue his claims by fifth-column activities in Sweden and propaganda in Europe.

The war in Russia was much more serious, and it was here that Gustavus, in a succession of difficult and indecisive campaigns, learned the rudiments of warfare. It dragged on until ended by the Peace of Stolbova in 1617, by which time it had clearly changed its character. Charles IX had intervened in Russia to prevent the Poles from placing their own candidate on the Russian throne; the election of the Russian Michael Romanov in 1613 had ended that danger, and Gustavus continued the struggle with the deliberate intention of annexing as much of Russian territory as possible. He feared Russia's military and naval potential; he feared that once the country's position was stabilized, a new tsar might try to make Russia a Baltic maritime power. He was determined, therefore, to exploit Russia's momentary weakness to cut it off from direct maritime contact with the West and to channel Russian trade through Swedish middlemen, thus enriching his impoverished exchequer with tolls and duties. In this last respect the outcome proved disappointing, but politically and strategically Stolbova was a treaty of European importance. By annexing Ingria and Kexholm, Sweden came to possess a continuous belt of territory connecting Finland with the Swedish province of Estonia. It thus cut Russia off entirely from the Baltic, thrust it back toward Asia, and postponed its emergence as a major European power until the time of Peter the Great.

Annexation of Russian land

Resolution of internal problems. Meanwhile, the internal tensions that Gustavus Adolphus had inherited had been largely resolved. The charter that the Estates extorted from Gustavus when he became king in 1611 might well have entailed the virtual subjection of the monarchy to the council and the high aristocracy. This, however, did not happen; for the man who had drawn the charter, the chancellor Axel Oxenstierna, became, in fact, the King's closest collaborator and remained so for the whole of the reign—a great historic partnership in which the temperaments and gifts of each supplemented those of the other. The King observed the spirit of the charter, and the aristocracy did not always insist on the observance of its exact provisions. They found in Gustavus a king favourable to their interests. He enlisted the nobility in the service of the state and thus provided them with numerous economic benefits. It was one of the healthiest features of Swedish society during this period that the nobility served the state, prepared to sacrifice even its privileges in the interests of the country. Thus the long-standing constitutional struggle between crown and aristocracy was suspended during his reign, largely because of the personality of the sovereign and the unique collaboration between himself and Oxenstierna. In this improved climate it was possible to undertake measures of sweeping reform.

Administrative reforms

The first decade of the reign, therefore, saw the creation of a new Supreme Court (1614) and the establishment of the Treasury and the Chancery as permanent administrative boards (1618); and by the end of the reign an Admiralty and a War Office had been created—each presided over by one of the great officers of state. The Form of Government of 1634 summed up these reforms in a general statute giving Sweden a central administration more modern and efficient than that of any other European country. Stockholm became a true capital with a permanent population of civil servants, the most important of whom were noblemen. And in the 1620s a thorough reform professionalized local government and placed it securely under the control of the crown. The Council of State became, for the first time, a permanent organ of government, able to assume charge of affairs while the King was fighting overseas. An ordinance of 1617 fixed the number of Estates in the Riksdag at four (nobles, clergy, burghers, and peasants) and regulated its procedures on a basis that lasted until 1866. Both council and Riksdag were identified with the King's policies, not least because of Gustavus' brilliant gift for expounding them: his speeches reveal him as a master of debate and an orator of extraordinary eloquence and force. And the decisions were always his, though they were usually arrived at after intimate consultation with Axel Oxenstierna. His hesitations, his vacillation in the face of grave decisions (such as that of intervening in the Thirty Years' War in Germany), reflect his profound sense of responsibility to the nation. Of all these domestic reforms, however, none had a more enduring and more beneficial effect upon his country than his work for education: his creation of the *Gymnasia* in the 1620s gave Sweden, for the first time, an effective provision for secondary education; his splendid munificence to the University of Uppsala gave it the financial security that was essential to its development; and his foundation of the University of Dorpat (now Tartu State University) provided the first centre for higher learning in the Baltic languages.

Educational reform

In 1620 he married Maria Eleonora of Brandenburg. In 1621, taking advantage of a Turkish attack upon Poland, Gustavus renewed the war with Sigismund. His capture of Riga was followed by a gradual conquest of Livonia (present Latvian and Estonian S.S.R.'s). His object was to compel Sigismund to renounce his claims to Sweden; and he hoped to gain his end by the economic pressure which would result from Poland's loss of access to its main export routes to western Europe. It was in pursuit of this policy that, in 1626, he transferred the seat of war to Prussia: a stranglehold on the Vistula River, he hoped, would bring Poland to its knees. But already he was concerned with the larger question of the danger to German Protestantism entailed by the victorious campaigns of the Habsburg commanders, Tilly and Wallenstein. He saw his Polish campaigns as one aspect of the general struggle of Protestantism against the Counter-Reformation: if Sigismund were restored to the Swedish throne, the re-Catholicization of Scandinavia would follow soon after; the Habsburgs and their allies would be able to close the passage into the Baltic to Dutch shipping, and the United Netherlands might then be unable to continue their struggle against Spain.

Prologue to intervention in Germany

Thus, the fate of Europe was bound up with what happened in Livonia or Prussia. Protestant Europe was slow to appreciate the connection; but as the Protestant cause plunged to disaster in Germany, its leaders increasingly turned their eyes to Gustavus as a possible saviour. But before he was prepared to commit himself to any Protestant league, Gustavus required adequate assurance of support. The disastrous defeat (1626) of Christian IV of Denmark, who had intervened in Germany without such an assurance, justified his caution, but it also made Swedish intervention inevitable. The Imperialists' occupation of the German Baltic shore and their plans for a Habsburg-Polish navy seemed to pose a direct threat of invasion. In this emergency, Gustavus and Christian joined forces to send an expedition to Stralsund, the last remaining Protestant bastion in Pomerania, which arrived just in time to prevent its capture by Wallenstein (1628). From this moment, full-scale involvement in the German war became simply a question of time. The Polish war was resolved in 1629 by the Truce of Altmark, and Gustavus was at last free to turn his attention to Germany. In June 1630 the Swedish expeditionary force landed at Peenemünde.

Entrance into Thirty Years' War. The motives prompting his intervention have long been a subject of historical controversy. An older generation of historians saw him, as his contemporaries did, simply as the Protestant Hero, the "Lion of the North"; later, he was viewed as having been moved by purely political considerations; and in recent days he has been characterized as an economic imperialist who sought to remedy Sweden's poverty by seizing control of the whole Baltic coastline, and thus to monopolize trade between Russia and western Europe. The most probable explanation, however, is the one which he himself adduced: that he sought security from dangers which seemed to threaten the Swedish state and the Swedish church; that he considered his actions essentially defensive; and that he had no precise long-range plans, either economic or political, when he landed on German soil.

Military innovations

He had, however, an army of unusual quality, fighting in a style new to Germany, and he combined tactical innovations with a grander concept of strategy than Europe had seen for many years. By reducing the size of the tactical unit, by opposing a flexible linear formation to the cumbrous massive formations of his opponents, by solving (at least for his time) the perennial problem of combining infantry and cavalry, missile weapons and shock; and, lastly, by producing the first easily manoeuvrable light artillery, he completed the transformation of the art of war begun by the Dutch commander Maurice of Nassau, prince of Orange, earlier in the century. The vastness of his operations in Germany initiated a permanent increase in the size of European armies. The whole process had profound social effects on the history of Europe.

Gustavus landed in Germany without allies. Whatever the feelings of the Protestant populations, the Protestant princes resented Swedish interference; and the refusal of George William of Brandenburg to cooperate with the Swedes thwarted Gustavus' attempts to save Magdeburg from capture and sack at the hands of Tilly's armies. His position was strengthened, however, by the Treaty of Bärwalde in January 1631, an alliance with France in which almost all the advantage lay with Sweden; in June he extorted by force the reluctant collaboration of Brandenburg; in September John George of Saxony, provoked by violations of his neutrality, formally allied himself with Sweden. In September, at Breitenfeld, the Swedish-Saxon forces shattered Tilly's army in a battle that was a landmark in the art of war and a turning point in the history of Germany. In the ensuing months Gustavus swept triumphantly through central Germany, systematically consolidating his base areas as he advanced: by Christmas he had established himself at Mainz. It seemed that the fate of Germany lay in his hands.

These developments forced Gustavus to reassess the limited and vague plans with which he had embarked on the expedition. In 1630 he had defined his aims as security and indemnity, the indemnity to be a cash payment to cover his war expenses, the security to be provided by a permanent Swedish alliance with Pomerania. By the close of 1631, with most of north and central Germany under his control and the liberation of the south German Protestant states already in prospect, his plans had broadened. He had always insisted that the German Protestant princes must work for their own salvation, and he saw the best hope for their future preservation in the creation of a comprehensive, permanent Corpus Evangelicorum (or Protestant league). His experience of the feckless and selfish German princes convinced him that such a league could be effective only if it were organized and directed by himself, and military necessity in any case demanded a unified command that could not be directed by anyone other than himself. Security, then, was to be achieved by a Protestant league of which he would be patron, military director, and political head. For indemnity he no longer claimed monetary compensation but large territorial cessions, particularly, the transference of Pomerania to Sweden. Thus, the old security had become the new indemnity. Many Germans feared, and some Swedish diplomats now believed, that a final settlement must probably entail the deposition of the German emperor Ferdinand II and the election of Gustavus as emperor in his place. It was a solution he must certainly have contemplated, but there is no firm evidence of his attitude; probably he considered it only as a last resort. Certainly it would have alienated those German allies who had no wish to exchange a Habsburg domination for a Swedish one. They already resented Gustavus' dictatorial methods as well as the Swedish army's practice of making war support war. A Swedish administration was being organized in the occupied areas; Gustavus rewarded his generals and supporters by conferring the conquered lands on them; in some of the treaties he concluded with German princes there was more than a hint that he regarded them as his feudal inferiors. In October 1632 he did, indeed, lay the basis for a league of Protestant princes; but it was confined mainly to southern Germany, where the peril from an Imperialist reaction was greatest, and the two greatest Protestant states—Saxony and Brandenburg—never became part of it.

The Protestant league

Last phase of Gustavus' campaign. The prospect of success depended upon the outcome of the campaign of 1632, which was designed to cripple Bavaria as a preliminary to the conquest of Vienna in 1633. Up to a point, it was highly successful. The brilliant crossing of the Lech River in Bavaria, in the face of Tilly's armies, opened the way to the occupation of Munich. In this crisis, Wallenstein, whom the Emperor had dismissed from his service in 1630, was recalled to lead the imperial armies. His threat to Nürnberg forced Gustavus to leave Bavaria in order to relieve the city. His attack on Wallenstein's entrenchments on the Alte Veste—an operation that probably no other contemporary com-

mander would have attempted—was unsuccessful; and for the next few weeks there followed a tense war of manoeuvre that ended when Gustavus fell upon Wallenstein's army at Lützen (November 6, 1632) as it was dispersing to winter quarters. Morning mist robbed Gustavus of the advantage of surprise, and gave Wallenstein time to reunite his forces. The fight raged fiercely all day, but when night fell the Swedes had won a clear tactical victory and an important strategic success. It was, however, dearly bought; for while leading a cavalry charge Gustavus became separated from his men, and perished in the *mêlée*.

His death came at a moment when it had already begun to appear that the victory he believed to be essential to the stability of Germany and the security of Sweden might be more difficult to achieve than he had imagined. But he had lived long enough to deflect the course of German history; he had launched his country upon a career as a great power that it was really not strong enough to pursue; and by his domestic reforms he had bequeathed to Sweden things more durable and more important than all the captured standards of Breitenfeld and Lützen.

BIBLIOGRAPHY. The standard work in English is MICHAEL ROBERTS, *Gustavus Adolphus: A History of Sweden 1611–1632*, 2 vol. (1953–58), which contains a very full bibliography. NILS AHNLUND, *Gustav Adolf the Great* (Eng. trans. 1940), is a classic study of selected aspects of the reign by the leading Swedish authority. JOHANNES PAUL, *Gustav Adolf*, 3 vol. (1927–32), is the best account in German, with its main emphasis on the military and diplomatic history of the Thirty Years' War. For a discussion of the King's motives and objectives for intervention in the war, see MICHAEL ROBERTS, *Essays in Swedish History* (1967), which also includes studies of his significance for the art of war, and of the constitutional implications of the reign. Some illustrative documents are printed, in English translation, in MICHAEL ROBERTS, *Sweden as a Great Power, 1611–1697* (1968).

(M.Ro.)

Gutenberg, Johannes

Johannes Gutenberg, a German craftsman and inventor, originated a method of printing from movable type that was used without important change until the 20th century. The unique elements of his invention consisted of a mold, with punch-stamped matrices (metal prisms used to mold the face of the type) with which type could be cast precisely and in large quantities; a type-metal alloy; a new press, derived from those used in wine making, papermaking, and bookbinding; and an oil-based printing ink. None of these features existed in Chinese or Korean printing, or in the existing European technique of stamping letters on various surfaces; or in woodblock printing.

Gutenberg, whose full name was Johannes Gensfleisch zur Laden, alternatively called (zum) Gutenberg from a family estate, was born in the last decade of the 14th century, the son of a patrician of Mainz, Germany. What little information exists about Gutenberg, other than that he was associated with the goldsmith's guild and had acquired skill in metalwork, comes from documents of financial transactions. Exiled from Mainz in the course of a bitter struggle between the guilds of that city and the patricians, Gutenberg moved to Strassburg (now Strasbourg, France) in 1430. Records of his association with the goldsmith's guild put his presence there between March 14, 1434, and March 12, 1444. In partnership with business associates, he engaged in such crafts as gem cutting and the production of mirrors, and he also taught these crafts to a number of pupils.

Some of his partners, who became aware that Gutenberg was engaged in work that he kept secret from them, insisted that, since they had advanced him considerable sums, they should become partners in these activities as well. Thus, in 1438 a five-year contract was drawn up between him and three other men: Hans Riffe, Andreas Dritzehn, and Andreas Heilmann. It contained a clause whereby in case of the death of one of the partners, his heirs were not to enter the company, but were to be compensated financially. When Andreas Dritzehn died at Christmas 1438, his heirs, trying to circumvent the terms of the contract, began a lawsuit against Gutenberg in which they demanded to be made partners. They lost the suit, but the trial revealed that Gutenberg was working on a new invention. Witnesses testified that a carpenter named Conrad Saspach had advanced sums to Andreas Dritzehn for the building of a wooden press, and Hans Dünne, a goldsmith, declared that he had sold to Gutenberg, as early as 1436, 100 guilders' worth of printing materials. Gutenberg, apparently well along the way to completing his invention, was anxious to keep secret the nature of the enterprise.

Invention of the press

After March 12, 1444, Gutenberg's activities are undocumented for a number of years, but it is doubtful that he returned immediately to Mainz, for the quarrel between patricians and guilds had been renewed in that city. In October 1448, however, Gutenberg was back in Mainz to borrow more money. By this time his printing experiments had apparently reached a considerable degree of refinement, for he was able to persuade Johann Fust, a wealthy financier, to lend him 800 guilders—a very substantial capital investment, for which the tools and equipment for printing were to act as securities. Two years later Fust made an investment of an additional 800 guilders for a partnership in the enterprise. Fust and Gutenberg eventually became estranged, Fust, apparently, wanting a safe and quick return on his investment, while Gutenberg aimed at perfection rather than promptness.

Gutenberg's dream was to create a way of mechanically reproducing medieval liturgical manuscripts without losing any of their colour or beauty of design. His dream was destroyed, however, when Fust won a suit against him, the record of which is preserved, in part, in what is called the Helmaspergersches Notariatsinstrument (the Helmasperger notarial instrument), dated November 6, 1455, now in the Library of the University of Göttingen. Gutenberg was ordered to pay Fust the total sum of 2,026 guilders, which included the two loans and compound interest. Instead of the chance to perfect his invention, he faced complete financial ruin. There is no reason to doubt that the printing of certain books (*werck der bucher*, specifically mentioned in the record of the trial, refers to the 42-line Bible that was Gutenberg's masterpiece) was completed, according to Gutenberg's major biographers, in 1455 at the latest. Its sale, it has been estimated, would have produced many times over the sum owed Fust by Gutenberg, and there exists no explanation as to why these tangible assets were not counted among Gutenberg's property at the trial.

After winning his suit, Fust gained control of the type for the Bible, and for Gutenberg's second masterpiece, a Psalter, and all of Gutenberg's other printing equipment. He continued to print, using Gutenberg's materials, with the assistance of Peter Schöffer, his son-in-law, who had been Gutenberg's most skilled employee and a witness against him in the 1455 trial. The first printed book in Europe to bear the name of its printer is a magnificent Psalter completed in Mainz on August 14, 1457, which lists Johann Fust and Peter Schöffer.

The Psalter

The Psalter is decorated with hundreds of polychrome initial letters and delicate scroll borders that were printed in a most ingenious technique based on multiple inking on a single metal block. Most experts are agreed that it would have been impossible for Fust and Schöffer alone to have invented and executed the intricate technical equipment necessary to execute this process between November 6, 1455, when Gutenberg lost control of his printing establishment, and August 14, 1457, when the Psalter appeared. It was Gutenberg's genius that was responsible for the Psalter decorations. In the 1960s it was suggested that he may also have had a hand in the creation of copper engraving, in which he may have recognized a method for producing pictorial matrices from which to cast reliefs that could be set with the type, initial letters, and calligraphic scrolls. It is at present no more than a hypothesis, but Gutenberg's absorption in both copper engraving and the Psalter decorations would certainly have increased Johann Fust's impatience and vindictiveness.

A number of smaller printings, many preserved only in

fragments, and printed in what appear to be experimental stages of Gutenberg's types, used to be attributed to him. They are now considered the work of other minor printers; among these is a 36-line Bible printed in Bamberg, a typographic resetting of the 42-line Bible. Attributed to Gutenberg himself is a *Türkenkalender*, a warning against the impending danger of Turkish invasion after the fall of Constantinople in 1453, printed December 1454 for 1455 use, some letters of indulgence, and some school grammars. The identity of the printer of a *Missale Speciale Constantiense* is still not established. One group of scholars believes it was printed by Gutenberg before the 42-line Bible, another group considers it to be a product of type acquired after Gutenberg's death in the 1470s.

After Gutenberg's financial ruin, a Mainz municipal official, Dr. Konrad Humery, lent him printing tools and equipment, but what these were, and what use Gutenberg made of them is not known. There is increasing doubt that the 1460 *Catholicon*, a weighty folio once attributed to him, could have been his work because of the book's size and the inferior quality of the types used. Toward the end of his life, when, according to one source, he was partially or totally blind, the elector Adolph von Nassau took pity on the destitute inventor and made him a member of his court, a tax-free sinecure that provided him with a yearly allowance of cloth, grain, and wine. He died most probably on February 3, 1468, and was buried in the Franciscan church in Mainz.

BIBLIOGRAPHY. ALOYS RUPPEL, *Johannes Gutenberg: Sein Leben und sein Werk*, 3rd ed. (1967), the definitive biography (a reprint of the 1947 2nd edition, not including some necessary updating); VICTOR SCHOLDERER, *Johann Gutenberg* (1963), a brief discussion of Gutenberg's invention, pleasant and easy to read and well illustrated; D.C. MCMURTRIE (ed.), *The Gutenberg Documents* (1941), a translation of the texts of the documents based with authority on the compilation by Karl Schorbach; HELMUT PRESSER, *Johannes Gutenberg in Zeugnissen und Bilddokumenten* (1967), a richly illustrated biography of Gutenberg, set into the framework of his time and the places where he lived and worked; FERDINAND GELDNER, *Die deutschen Inkunabeldrucker*, vol. 1 (1968), a detailed biography of German 15th-century printers with a full and richly documented discussion of Gutenberg and his invention; HELLMUT LEHMANN-HAUPT, *Gutenberg und der Meister der Spielkarten* (1962; enlarged Eng. trans., *Gutenberg and the Master of the Playing Cards*, 1966), an examination of the question whether Gutenberg may not have been involved with the origins or early developments of copper engraving, based on the recently discovered connections between Gutenberg and early copper engravings, illuminated manuscripts, and illuminated printed books produced around the middle of the 15th century; CARL WEHMER, *Mainzer Probedrucke* (1948), an important report on hitherto unknown proofs of early Mainz experimental printing discovered by the author in Cracow.

(H.E.L.-H.)

Guyana

Guyana is an independent republic and member of the Commonwealth of Nations located in the northeastern corner of South America. It is bordered by Venezuela to the west, Brazil to the southwest and south, Surinam to the east, and the Atlantic Ocean to the north. Its total area of 83,000 square miles (215,000 square kilometres) is largely uninhabited, and most of the country's 714,000 inhabitants occupy the narrow coastal strip. The national capital of Georgetown is located on the Atlantic coast and is one of the country's major ports. The name Guyana is derived from the term *guiana*, an Amerindian word meaning "land of waters," which applies to the entire region of South America between the Amazon, Orinoco, and Negro rivers.

Present-day Guyana continues to reflect its colonial past. Its economy is dominated by the sugar industry, which is still conducted by plantation agriculture and remains in British hands. The growing bauxite industry is also under British control, as is much of the country's commercial life. Guyana's social composition is also of colonial origin. The indigenous population is small and unintegrated, and the coastal peoples are largely descendants of slaves and indentured labourers who were imported to work the sugar plantations.

MAP INDEX

Political Subdivisions

East Berbice.....5·00n 57·58w
East Demerara....6·20n 58·00w
Essequibo.....7·00n 59·00w
Essequibo Islands.6·45n 58·35w
Mazaruni Potaro..6·00n 60·00w
North West.....7·30n 59·50w
Rupununi.....3·00n 58·30w
West Berbice.....6·20n 57·55w
West Demerara...5·25n 58·35w

Cities and towns

Bartica.....6·24n 58·37w
Bush Lot.....6·10n 57·18w
Buxton.....6·48n 58·02w
Charity.....7·24n 58·36w
Dadanawa.....2·50n 59·30w
Enmore.....6·45n 57·59w
Enterprise.....6·50n 58·25w
Everton.....6·12n 57·31w
Fort Wellington...6·30n 57·30w
Georgetown.....6·48n 58·10w
Holmia.....4·59n 59·40w
Hyde Park.....6·30n 58·16w
Isherton.....2·20n 59·25w
Issano.....5·49n 59·26w
Lethem.....3·20n 59·50w
Mabaruma.....8·10n 59·50w
Mackenzie.....6·00n 58·17w
Mahaicony.....6·36n 57·48w
Marlborough.....7·29n 58·38w
Matthews Ridge...7·30n 60·10w
Morawhanna.....8·17n 59·44w
New Amsterdam..6·15n 57·31w
Orinduik.....4·40n 60·03w
Parika.....6·51n 58·26w
Port Kaituma.....7·48n 59·52w
Potaro Landing...5·22n 59·09w
Queenstown.....7·12n 58·30w
Rockstone.....5·59n 58·32w
Rose Hall.....6·18n 57·23w
Rosignol.....6·16n 57·32w
Saint Ignatius Mission.....3·19n 59·47w
Skeldon.....5·57n 57·09w
Spring Garden....6·59n 58·30w
Suddie.....7·08n 58·29w
Tumatumari Fall..5·21n 58·59w
Vreed en Hoop....6·48n 58·11w
Wismar.....5·59n 58·18w

Physical features and points of interest

Acarai Mountains.2·00n 57·30w
Atlantic Ocean...8·00n 57·00w
Barama, *river*.....7·40n 59·15w
Barima, *river*.....8·33n 60·25w
Berbice, *river*.....6·20n 57·32w
Burro-Burro, *river*.....4·50n 58·50w
Corocoro Island...8·30n 59·50w
Courantyne, *river*..5·55n 57·05w
Cuyuni, *river*.....6·23n 58·41w
Demerara, *river*...6·50n 58·10w
Essequibo, *river*...6·50n 58·30w
Essequibo Islands.6·55n 58·30w
Illiwa, *river*.....3·55n 58·45w
Ireng, *river*.....3·33n 59·51w
Kaieteur Fall, *waterfall*.....5·10n 59·35w
Kaituma, *river*.....8·10n 59·43w
Kako, *river*.....5·45n 60·33w
Kamarang, *river*...5·53n 60·36w
Kamoa Mountains.....1·30n 59·00w
Kanuku Mountains.....3·12n 59·30w
Kassikaityu, *river*..1·50n 58·35w
Kuyuwini, *river*....2·15n 58·10w
Kwitaro, *river*.....3·20n 58·45w
Marina Fall, *waterfall*.....5·25n 59·30w
Mazaruni, *river*....6·25n 58·35w
New, *river*.....3·20n 57·37w
Oronogue, *river*...2·45n 57·25w
Pakaraima Mountains.....5·00n 61·00w
Pomeroon, *river*...7·35n 58·45w
Potaro, *river*.....5·25n 58·50w
Puruni, *river*.....6·01n 59·10w
Roraima, Mount, *mountain*.....5·12n 60·44w
Rupununi, *river*...4·03n 58·35w
Takutu, *river*.....3·33n 59·51w
Tiboku Falls, *waterfalls*.....5·44n 59·35w
Tiger Hill.....5·40n 58·23w
Waini, *river*.....8·20n 59·50w
Wonotobo Falls, *waterfalls*.....4·22n 57·57w

THE LANDSCAPE

The natural environment. *Relief features.* The narrow plain that stretches along the country's Atlantic coast is largely a man-made feature. Between three and four miles wide, the area lies below the high-tide level and has been reclaimed from the sea by canals and about 140 miles of dikes. Its inland border is also marked by canals that protect the plain from alluvial swamps.

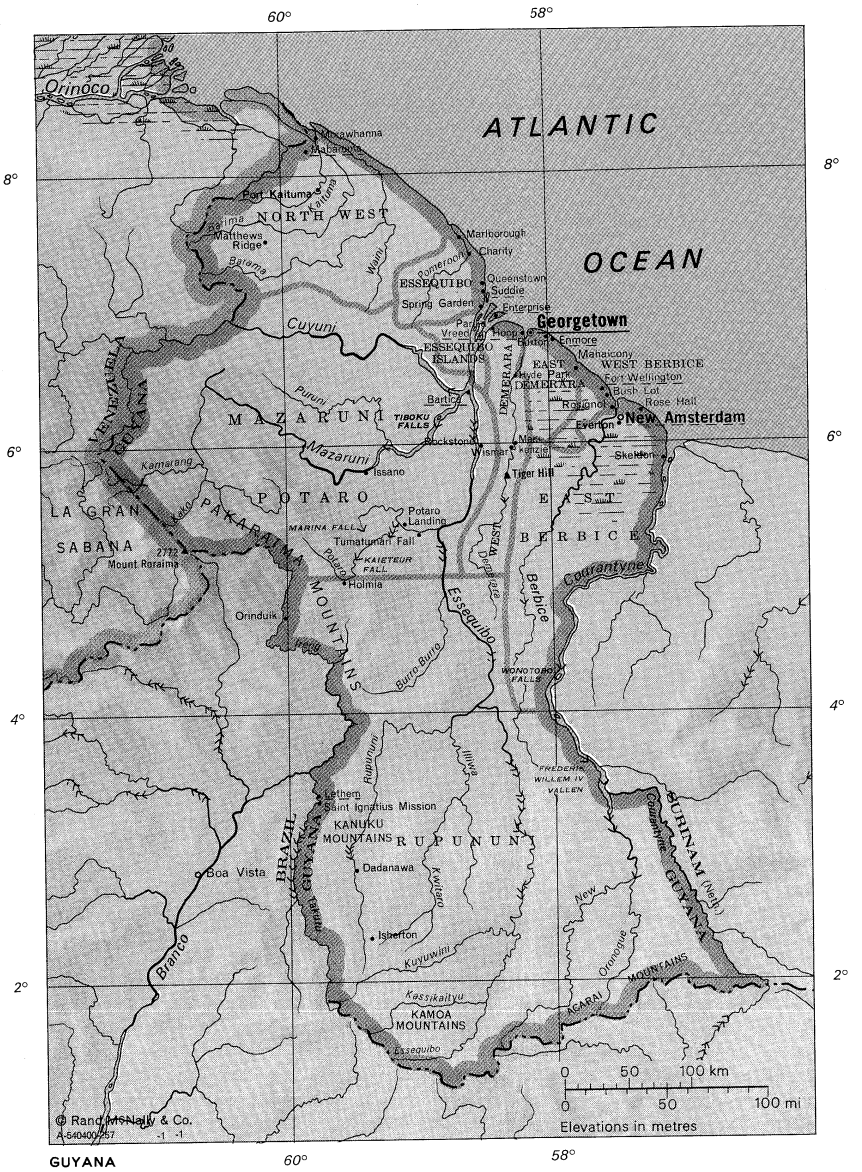
The coastal plain

About 40 miles from the coast is a region of undulating hills that rise inland from 50 feet on the coastal side of the region to 400 feet on the western side. The area is between 80 and 100 miles wide and is widest in the southeast. It is covered with sands, from which it takes its name as the white-sands (*zanderij*) region. A small savanna region in the east lies about 60 miles inland from the coast and is surrounded by the white-sands belt.

The sands partly overlie a low crystalline plateau that is generally less than 500 feet in elevation. It forms most of the country's centre and is penetrated by igneous intrusions that cause the numerous rapids of Guyana's rivers.

Beyond the crystalline plateau, the Kaieteurian Plateau lies generally below 1,640 feet above sea level. It is overlain with sandstones and shales that, in the south, form the extensive Rupununi savanna region. The plateau culminates in the Acarai Mountains (Serra Acarai), which rise to about 2,000 feet on the southern border, and is crowned on the western frontier by the Pakaraima Mountains, which rise to 9,094 feet (2,772 metres) in the Roraima massif. The Rupununi savanna is bisected by the Kanuku Mountains, which trend from west to east and rise to almost 3,000 feet above sea level.

Drainage and soils. Guyana's four main rivers—the Courantyne, Berbice, Demerara, and Essequibo—all flow from the south and empty into the Atlantic along the eastern section of the coast. Among the tributaries of the Essequibo, the Potaro, Mazaruni, and Cuyuni drain the northwest, and the Rupununi drains the southern savanna. The western coast is cut by shorter rivers, including the Pomeroon, Waini, Barima, and Kaituma.



GUYANA

The country's rivers are part of the watershed of the Amazon and Orinoco rivers (*qq.v.*), and the headwaters of the Rupununi in Brazil are often confused with those of the Amazon. Drainage is poor, because the average gradient is only one foot per mile, and there are swamps and flooding in the mountains and savannas. The rivers are not suitable for transportation because they are broken by interior falls and their mouths and estuaries are blocked by mud and by sandbars that may occur two to three miles out to sea.

Coastal soils

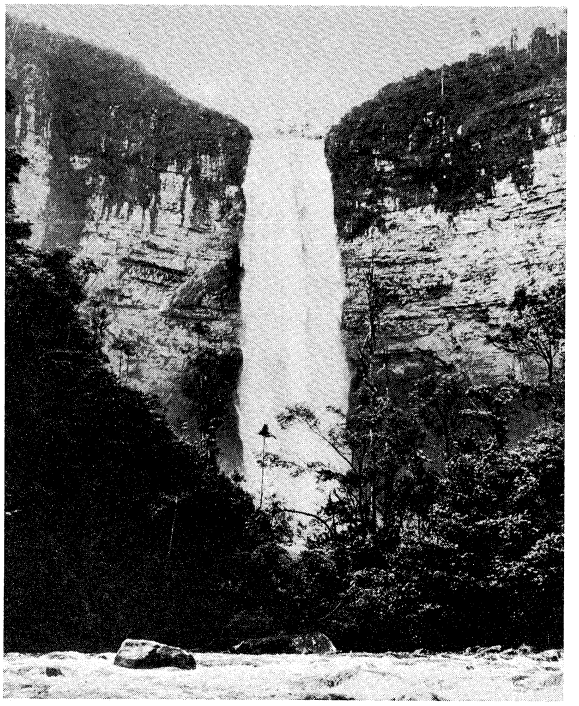
The coastal soils are fertile but acidic. The clays of the coastal plain are composed of alluvium from the Amazon that is deposited by the south equatorial ocean current and of alluvium from the country's rivers. They overlay white sands and clays and can support intensive agriculture but must be subjected to fallowing to restore fertility. *Pegasse* soil, a type of tropical peat, occurs behind the coastal clays and along the river estuaries, while silts line the banks of the lower rivers. Reef sands occur in bands in the coastal plain, especially near the Courantyne and Essequibo rivers. The rock soils of the interior are leached and infertile, and the white sands are almost pure quartz.

Climate. Guyana has the climatic characteristics of an equatorial lowland, with high temperatures, heavy rainfall with relatively small seasonal differences, high humidity, and high average cloud cover. Temperatures are remarkably uniform. At Georgetown the average monthly temperature varies from 79° to 82° F (26° to 28° C), with a daily range of about 15° F (8° C). The effects of constant heat and high humidity are mitigated near the coast by the northeast trade winds.

Rainfall derives mainly from the movement of the intertropical front, or doldrums. It is heavy everywhere on the plateau and the coast. The annual average at Georgetown is about 90 inches, and on the interior savannas it is about 70 inches. On the coast a long wet season from April to August and a short wet season from December to early February are sufficiently well marked on the average, but in the southern savannas the short wet season is not experienced. Total annual rainfall is variable and years of drought occur.

Vegetation and animal life. Many plants of the coast, such as the mangrove and various saltwater grasses, grow in shallow brackish water and help to protect or extend the land. The wet savanna behind the coast has coarse tufted grasses and a wide scattering of palms, notably the coconut, truli, and manicole. High rain forest, or selva, covers 73 percent of the land area and is of extraordinary variety and magnificence. Prominent trees include the greenheart and the wallaba on the sandy soils of the northern edge; the giant mora and crabwood on swampy sites; the balata and other latex producers; and many trees such as the siruaballi and hubaballi that yield handsome cabinet woods. The interior savanna is mostly open grassland, with much bare rock, many termite hills, and clumps of ita palm.

All forms of animal life are immensely varied and



King George VI, or Great, Falls, on the Kamarang River, near the Venezuelan border.
UPI Compix

abundant, though few, apart from birds and insects, are normally visible. The tapir is the country's largest land mammal, the ocelot the fiercest, and monkeys and deer the commonest. Among the more exotic species are the sloth; the great anteater; the capybara, or bush pig; and the armadillo. Bird life includes the vulture, kiskadee, blue sacki, hummingbird, kingfisher, and scarlet ibis of the coast and lower rivers; and the macaw, tinamou, bellbird, and cock-of-the-rock in the forest and savanna. The cayman (a reptile similar to the alligator) is the commonest of the larger freshwater creatures. The giant anaconda, or water boa, is the largest of many kinds of snakes and the bushmaster the most vicious. Lizards are numerous and include the iguana in the lower rivers. Sharks and stingrays are found offshore. The snapper and grouper are the most esteemed of the ocean fish commonly landed.

The landscape under human settlement. *Traditional regions.* The country is divided traditionally between the coast, which is inhabited by about 95 percent of the population, and the interior. The coastal population is heterogeneous, its inhabitants descended from Dutch settlers and the labourers brought in to work the sugar plantations. The interior, despite scattered ranching and mining settlements, is largely the province of the Amerindian.

Rural settlements

Settlement patterns. Guyanese society is predominantly rural, most of the people occupying small villages strung out along the coast. The highest concentrations of villages are found along the estuary of the Demerara River and near the mouths of the Berbice and Courantyne rivers. The villages range in size from several hundred to several thousand inhabitants and are separated from each other by drainage ditches. Buildings are usually constructed on stilts and are connected to the streets by foot bridges over the drainage canals that run parallel to the streets.

Georgetown, with a population of about 167,000, is the country's main port and its largest city. Located at the mouth of the Demerara River, it lies below sea level and is protected by dikes along both the river and the sea. The second largest town is Mackenzie, a mining centre with 20,000 inhabitants located 55 miles (88 kilometres) upstream from Georgetown. The market centre of New Amsterdam is located on the mouth of the Berbice River; it is the third largest town, with a population of 18,000. The city's general layout and central canals are reminiscent of The Netherlands. Most construction is of wood, painted light blue; houses have a rectangular boxlike appearance.

PEOPLE AND POPULATION

The Amerindians

Population groups. The indigenous peoples of Guyana are collectively known as Amerindians. Their numbers in 1970 were approximately 33,000, which represented more than a threefold increase since 1930 and a reversal of centuries of decline. They comprise less than 5 percent of the population. Indian groups include the Warraus, Arawaks, Caribs, Wapisianas (Wapishanas), Arecunas, the mixed "Spanish Arawaks" of the Moruka River, and many more in the forest areas. The Makusí (Macussí or Macushí) are the most prominent of the savanna peoples. The Amerindians are rarely seen in the populated coastal areas. Since 1953 they have had important rights of occupation in reservations totalling more than 6,000 square miles but are not confined to them. They are increasingly employed in the cattle and logging industries.

The other major elements in the population are all predominantly coastal dwellers. Descendants of African slaves form the oldest group of nonindigenous peoples; they abandoned the plantations after emancipation in 1834 to become independent peasantry or town dwellers. The Afro-Guyanese comprised 31 percent of the population in 1970. The East Indians came predominantly as indentured labour from India to replace the Africans in plantation work. They form the largest racial group in the country—51 percent of the total in 1970—and have been increasing more rapidly than any of the others. They are the mainstay of plantation agriculture, and many are independent farmers and landowners, have done well in trade, and are well represented among the professions.

The Chinese and Portuguese also entered originally as agricultural labourers but are now rarely found outside the towns. They are active in business and the professions, and their influence is disproportionate to their numbers; they have not been increasing, however, and together comprise only about 2 percent of the population. Europeans other than Portuguese are few, and most are short-term inhabitants. While every kind of racial mixture may be found, mulattos (persons of mixed white and black ancestry) are by far the most common. Most of them live in towns, and a high proportion are in clerical and professional work.

Over 33 percent of the population are Hindu. Approximately 40 percent are Christians, belonging to the Anglican, Roman Catholic, or Methodist churches. Another 9 percent practice Islām. Animistic religions are still practiced by some of the indigenous peoples. The official and principal language is English, but a creole patois is spoken throughout the country. Hindi and Urdu are also spoken.

Guyana, Area and Population

	area*		population	
	sq mi	sq km	1960 census	1970 census
Districts				
East Berbice	...	...	116,000	147,000
East Demerara	...	...	276,000	343,000
Essequibo	...	...	30,000	57,000†
Essequibo Islands	...	...	16,000	†
Mazaruni-Potaro	...	...	12,000	13,000
North West	...	...	13,000	16,000
Rupununi	...	...	10,000	14,000
West Berbice	...	...	27,000	37,000
West Demerara	...	...	62,000	87,000
Total Guyana	76,000 83,000	197,000 215,000	560,000‡	714,000

*No breakdown available. Of the two country total area figures, the first is the land area, the second the total area. †Essequibo includes Essequibo Islands. ‡Figures do not add to total given because of rounding.
Source: Official government figures.

Demography. In 1965 the population was estimated at 647,000, a figure almost two-thirds greater than the 375,000 in 1946. By 1970 it had grown to 714,000. This

growth is attributable to improved health conditions, resulting in a greatly reduced death rate. In 1970 there was an average of 6.4 registered deaths per 1,000 inhabitants and 35.7 registered live births per 1,000.

Migration patterns

Immigration is no longer important, and by the 1970s the number of foreign-born, long-term residents was insignificant. Emigration has fluctuated from as low as 1,500 persons per annum in the mid-1960s to 4,200 persons in 1970. Internal migration is also minimal, and there is no significant movement to the cities.

In 1967 more than 90 percent of the population lived on 4 percent of the land. The average population density in 1970 was nine persons per square mile. There were, however, more than 300 per square mile on the coastal plain and more than 1,000 per square mile on the east coast, where villages are most highly concentrated.

HISTORY

The Dutch period. The coast of Guyana was sighted by the Genoese explorer Christopher Columbus in 1498, but the area, long known as the Wild Coast, proved unattractive to the Spanish and Portuguese because of its lack of gold and difficulties of movement. Although Spain resisted, the Guiana coast was left to the Dutch, French, and English, whose attempts at colonization and exploration were almost wholly abortive until early in the 17th century. During the 17th and 18th centuries the Dutch were the leading colonists; most British activity was an offshoot of developments in the West Indies.

The early settlements were typically a few miles up the larger rivers. Dutch settlements existed on the Courantyne, Essequibo, and Cayenne rivers by 1616. The Dutch West India Company was founded in 1621; the slave trade was established; and official appointments in command of settlements were made. A successful British colony was planted on the Suriname River by Lord Willoughby of Parham in 1651, the colonists coming mainly from Barbados. Cayenne was finally established as French in 1674.

Early in the 18th century, colonists were attracted to the seacoasts by the fertility of the soil. The Berbice estuary was settled in 1732, and a fort was established on Fort Island in the Essequibo estuary. The greatest figure in 18th-century Guyanese history was Laurens Storm van's Gravesande, who was in Essequibo from 1738 to 1772. He began the orderly development of the Demerara region, kept the slave population in order by firm but humane action, established free immigration for all nationalities and remitted taxation for the immigrants' first ten years, and encouraged exploration.

Settlements changed hands with bewildering frequency during the wars (mostly between the British and the French) from 1780 to 1815. During a brief French occupation, Longchamps, later called Georgetown, was established at the mouth of the Demerara; the Dutch renamed it Stabroek and continued to develop it. The British took over in 1796 and remained in possession, except for short intervals, until 1814, when they purchased Demerara, Berbice, and Essequibo in the settlement that ended the Napoleonic Wars. Surinam went to the Dutch and Cayenne to the French. The 19th-century pattern was finally established when Berbice, Demerara, and Essequibo were united in 1831 as the colony of British Guiana.

The British period. The slave trade was abolished in 1807, when there were about 100,000 slaves in the three British colonies (Berbice, Demerara, and Essequibo). After emancipation in 1834 African labour left the plantations, and large areas of land went out of cultivation. Of the many attempts to organize immigration of suitable labour, only that for Asian Indians was at all successful. During the rest of the 19th century, the colony was almost overwhelmed with difficulties and was not helped by the indifference of Great Britain. Reliance on a single crop, sugarcane, was the main fault, and this was not corrected until well into the 20th century.

Settlement began on the Rupununi savanna in 1860 but proceeded very slowly. Gold was discovered in 1879, and a boom in the 1890s helped the colony over a difficult spell. The North Western District was organized in 1889 and was the cause of a sharp international flurry in 1895 when the United States supported Venezuela's claims to the territory. In the early 1960s Venezuela revived its territorial claims on British Guiana.



Masked dancers in the streets of Georgetown celebrating Guyana's independence from Great Britain on May 26, 1966.
Joseph Fabry, *Life* © 1966, Time Inc.

Governmental origins

The British inherited from the Dutch a singularly complicated constitutional structure. Except for a major change in 1891 that simplified it, the structure remained until 1928, when a typical crown-colony system was introduced. In 1953 a new constitution—with universal adult suffrage, a bicameral elected legislature, and a ministerial system—was introduced and, subject to several revisions, has remained in force.

From 1953 to 1966, the political history of the colony was stormy. The first elected government, formed by the People's Progressive Party (PPP) led by Cheddi Jagan, seemed so pro-Communist that the British government suspended the constitution in October 1953 and dispatched troops. The constitution was not restored until 1957. The PPP split along racial lines, Jagan leading a predominately East Indian party and Forbes Burnham leading a party of African descendants, the People's National Congress (PNC). In the elections of 1957 and 1961, the PPP was returned with working majorities and formed governments which pursued moderate policies. From 1961 to 1964 severe rioting and a long general strike led to the return of British troops to keep order, and the country was in grave economic difficulties.

To answer the PNC allegation that the existing electoral system unduly favoured the East Indian community, the British government introduced for the elections of December 1964 a new system of proportional representation. The result reflected racial divisions with considerable accuracy. The PNC formed a coalition government thereafter with the United Force Party, a group broadly representative of European and mixed racial interests. This government, led by Forbes Burnham, took the colony into independence under its new name, Guyana, on May 26, 1966. The PNC was returned to power by the general elections of December 1968 and on February 23, 1970, Guyana became a republic within the Commonwealth. Arthur Chung, a high court trial judge and the son of a Chinese farmer, was elected by the National Assembly as the first president of the republic of Guyana.

THE NATIONAL ECONOMY

Natural resources. The most important mineral resource is that of the extensive deposits of bauxite located between the Demerara and Berbice rivers. There are also significant deposits of manganese at Matthews Ridge in

the northwest, about 30 miles east of the Venezuelan frontier. Diamonds occur in the Mazaruni and other rivers of the Pakaraima Mountains. Gold is found in both alluvial and subsurface deposits. Other minerals include oil in the Rupununi savanna, copper, iron ore, molybdenite (a blue mineral that is the source of molybdenum, a metallic element used in strengthening steel), nickel, white sand (used in glass manufacture), kaolin (china clay), and graphite.

The main biological resource consists of the hardwoods of the tropical rain forest and especially the greenheart tree, which is resistant to termites, decay, and marine erosion. The shrimps off the coast and various inland fishes form the basis of the nation's fishing industry, and the grasses of the savanna regions are used for cattle grazing.

Hydroelectric potential in Guyana is considerable, especially at Tiger Hill on the Demerara River and Tiboku Falls on the Mazaruni. Development is hampered, however, by the remoteness of the falls and the large amounts of capital needed for generation and transmission facilities.

Sources of national income. *Agriculture, forestry, and fishing.* Agriculture is concentrated on the narrow coastal plain between the Essequibo and Courantyne rivers, where it supports about 90 percent of the population on 10 percent of the land. Because of the coast's relief and climate, certain agricultural features have evolved. Arable land is laid out in strips between the sea or a river and inland swamps. It is protected on all sides by dikes and canals that are used for both irrigation and drainage during the alternating dry and wet seasons. The land reclaimed from the sea is fertile but acidic; fertility must be returned to the soil by the process of periodic flood fallowing in which the fields are plowed, flooded for at least six months, drained, and planted.

Food crops include cassava, maize (corn), vegetables, and citrus fruit. Cash crops include sugarcane, rice, coffee, and cocoa. The production of sugarcane occurs almost totally on plantations owned by British companies. Bookers Sugar Estates Ltd. produces about 94 percent of the annual crop on 11 large estates. Two other estates are operated by the Demerara Company of Liverpool. Rice, however, is grown primarily on small private holdings by manual labour. Although agricultural production has generally increased since the 1960s, food must be imported to meet the demands of the expanding population. Further agricultural development is hampered by the high cost of mechanization and of preparation of new land.

Livestock production is concentrated mainly on the Rupununi savanna and on the coastal plain. Animals include beef cattle, dairy cattle, goats, sheep, pigs, and poultry.

Forestry activities are hampered by the lack of adequate transportation, the difficulty of cutting the extremely hard wood of Guyana's trees, and the shortage of facilities for the sawing, storing, and shipping of timber. Most of the timber produced for the domestic market and export is from the greenheart tree.

In the early 1970s fishing was not well developed, and fish were imported for food. The government hoped to improve activity through improved facilities and the promotion of fishponds. Shrimping is carried out primarily for export and is dominated by two companies from the United States and the United Kingdom.

Mining. Mining operations grew throughout the 1960s and produced 38 percent of all exports by value in 1970. Guyana is the fifth largest producer of bauxite after Jamaica, Australia, Surinam, and the Soviet Union. All alumina (aluminum oxide occurring in hydrated form in bauxite) and 85 percent of the bauxite mined is produced at Mackenzie by a Canadian company. The Mackenzie complex includes a processing plant for bauxite and alumina. The rest of the country's bauxite production is mined by a United States company at Kwakwani on the Berbice River; the company also operates a processing plant downriver at Everton.

Bauxite production

The mining of manganese by a United States company was suspended in 1968 because of depressed market conditions. The faltering gold industry is expected to revive following the discovery of two subsurface deposits in 1966 and the anticipated introduction of large-scale land and alluvial operations. Diamonds continue to be mined by hand and by suction dredges in the interior rivers.

Manufacturing. The vast industrial conglomerate of Bookers Sugar Estates Ltd. operates subsidiaries throughout the manufacturing sector of the economy. Guyana's industry centres upon the processing of agricultural products, and Bookers' companies operate nine of the country's 11 sugar plants; the other two plants are owned by the Demerara Company of Liverpool. Bookers also operates distilleries, a department and grocery store in Georgetown, and outlets for the sale of motor vehicles, agricultural machinery, hardware, electrical goods, office supplies, and pharmaceuticals. Its subsidiaries manufacture drugs and feed for livestock, process shrimp, produce concrete building blocks, retread tires, and engage in printing and the manufacture of boxes.

The Bookers industrial holdings

The country's 200 rice mills, like its rice fields, are generally small-scale individual units. Other domestic industries are oriented toward the replacement of consumer imports such as cigarettes and matches, edible oils, margarine, beverages, soap and detergents, and clothing.

Energy. The Guyana Electricity Corporation, the country's largest single supplier of electricity, has an installed capacity of 82,800 kilowatts. Private plants are also operated by the sugar estates and rice mills. Despite the country's large hydroelectric potential, all electricity in the early 1970s was produced by thermal generation and was available only on the coastal plain and at towns, such as Mackenzie, on the lower rivers.

Financial services. The Bank of Guyana, established in 1965, has the sole right of issue and acts as banker to the government and other banks. The country's four commercial banks are branches of British, Canadian, Indian, and United States banks. The Post Office Savings Bank mobilizes the deposits of small savers. Other financial services are provided by the New Building Society, which raises funds through the sale of shares at a fixed rate of interest; almost 40 insurance companies, most of which are foreign-owned; more than 700 cooperative societies, which serve as savings institutions and offer agricultural credit; and the Guyana Credit Corporation, which offers medium- and long-term loans for economic development.

Foreign trade. Guyana's major trading partners are the United States, the United Kingdom, Canada, and Trinidad and Tobago. About 36 percent of the country's exports are represented by bauxite and alumina, 27 percent by sugar, and 7 percent by rice. Shrimp, diamonds, molasses, rum, and timber are also sold abroad. Major imports include machinery, petroleum products, vehicles, textiles, footwear, and foods.

Management of the economy. The economy is primarily one of private enterprise, regulated by the law of supply and demand. Government participation is largely indirect, through its economic plans and the Guyana Development Corporation (GDC). Founded in 1963, the GDC is the only government agency for industrial development. The government's Private Investment Fund also promotes development, however; it offers medium- and long-term loans through the commercial banks. Guyana became a member of the Caribbean Free Trade Area (Carifta) together with Barbados and Antigua in 1965 in an attempt to expand the market for its domestic industrial products.

Most government revenues are derived from indirect taxes, which include customs, excise, and export duties. Direct taxes are largely derived from the corporate income tax, which is fixed at a flat rate of 45 percent, and a personal income tax. Revenues are dependent upon foreign trade, because customs duties are important (approximately 28 percent of the annual total revenue in 1970) and because most corporations are engaged in the exportation of sugar, rice, bauxite, and other commodities.

Sources of revenue

The country's most important labour organization is the

Trade Union Council, an association of the major unions. Its largest member is the Man-Power Citizens Association, an organization of sugar workers. Other organizations include the Guianese Agricultural Workers Union and unions of public service, mining, and industrial workers. The oldest employers' association is the Georgetown Chamber of Commerce and the largest is the Guyana Sugar Producers Association. Other organizations include the Consultative Association of Guyanese Industries, the Rice Millers and Landlords Association, and associations in the fields of mining, forestry, shipping, public works, and construction. Labour disputes are largely settled by informal collective bargaining and government arbitration.

Economic policies and problems

Government economic policy is based on the encouragement of foreign trade through Carifta, the GDC, and incentives to companies such as tax holidays and tax deductions. The government also plans to diversify local production in order to reduce imports and to provide employment and actively promote both public and private investment from abroad. These policies reflect the problems of an unfavourable balance of trade, the concentration of the production of goods subject to world price fluctuations, and growing unemployment. Guyana's economic prospects continue to depend, as in the past, on foreign aid, the slow development of the coastal plain, improved exploitation of the country's forests, and mineral exploration.

Transportation. There were about 1,200 miles (1,960 kilometres) of road in the early 1970s, of which about 450 miles were paved and the remainder were of burnt clay. A development scheme envisages the improvement of 400 miles of existing road and the building of 300 miles of new road; the Demerara River has been bridged at Mackenzie. There are also 400 miles of "vehicular trails" in the Rupununi savanna. The interior has few roads.

Coastal railways run from Georgetown to the Berbice and Essequibo rivers. A 30-mile (48-kilometre) line connects the manganese mines at Matthews Ridge with Port Kaituma on the Kaituma River.

Internal airways services of commercial and private aircraft use landing strips and the quieter stretches of rivers. Passengers and equipment are the chief freight. Timehri International Airport, 25 miles from Georgetown, is the country's main airport and is served by several international airlines.

Water transport constitutes by far the greatest mileage. It was used along the coast, on the canals of the sugar estates, by the ferry and steamer services of the lower rivers, and by the light craft of the interior.

Bauxite is loaded into oceangoing ships at Mackenzie and manganese ore at Port Kaituma, but, otherwise, the whole of the country's external trade passes through Georgetown, which maintains direct connections with the West Indies, Surinam, French Guiana, the United Kingdom, Canada, and the United States.

ADMINISTRATION, SOCIAL CONDITIONS, AND CULTURAL LIFE

The structure of government. *Constitutional framework.* In 1970 Guyana became a cooperative republic. The president is a ceremonial head of state who is elected to a six-year term by the country's unicameral legislative body, the National Assembly. Actual executive power resides in the prime minister, who is an elected member of the National Assembly and a member of the party with a majority of the assembly's seats. He serves upon the advice of the Cabinet, which is appointed by the president on the advice of the prime minister from members of the assembly or ministers who qualify to be elected. Ministers not of the assembly may not exceed four, and they may sit only as nonvoting members. The prime minister and the Cabinet are responsible to the assembly of 53 elected members, who serve four-year terms. The minority members of the assembly elect a leader of the opposition.

The executive structure also includes an ombudsman, who is charged with the investigation of administrative actions for malpractice or racial discrimination. The ombudsman is appointed upon the recommendation of the prime minister and the leader of the opposition to a four-year term.

Regional and local government. Apart from Georgetown and New Amsterdam, which are municipalities, the country is divided for local-government purposes into six coastal and three interior divisions. The coastal divisions are subdivided into villages and country districts, with councils responsible to a local government board. The interior divisions comprise seven-eighths of the area of the country but have a population of only 43,000. They are directly administered by the central government. A number of "Amerindian districts" have been demarcated for the sole occupation of Amerindians, and local councils have been formed in some of them.

The political process. The right to vote belongs to all Guyanese citizens over 21 years of age and any Commonwealth citizen who has lived in Guyana for more than one year. Voting is carried out by secret ballot under a system of proportional representation. Votes are cast for lists of candidates compiled by the political parties, and seats in the National Assembly are then allocated proportionately among the lists. Elections must be held at least every five years.

The two major political parties

The two dominant political parties are supported by about 85 percent of the electorate. They are the People's Progressive Party (PPP), which is primarily supported by the East Indian community, and the People's National Congress (PNC), which is supported by the Afro-Guyanese. The small United Force Party is composed of the Portuguese and the wealthy and has attracted the Amerindian vote. The parties reflect the country's ethnic divisions and, in their struggle for power, tend to exacerbate racial distrust.

Justice and the armed forces. Guyana has two legal traditions, the British common law and the Roman–Dutch code, the latter of which is now largely relegated to matters of land tenure. The constitution is the supreme law of the land. The court structure consists of magistrates courts, the High Court, and the Court of Appeal. The independent judiciary is headed by the chancellor, who presides over the Court of Appeal. The chancellor and the chief justice, who heads the High Court, are appointed by the prime minister, and all other judges and magistrates are appointed by the Judicial Service Commission.

The Guyana Defense Force (GDF) is composed of a regular and a reserve army and is charged with the maintenance of law and order. There is no navy or air force. There is no conscription, and the army is of minimal political importance. The GDF is directly controlled by the prime minister acting as minister of defense.

The social milieu. *Administrative services.* Education is free and compulsory from six to 14 years of age. Primary and secondary instruction are separate, although the lack of accommodations requires that some secondary classes are held in primary schools. The government accepts charge of most schools and gives grants-in-aid to denominational schools, many of which are located in the interior. Because of the lack of educational facilities, there are many private primary and secondary schools in the towns. The government's educational intentions are also hampered by the difficulties of attendance enforcement; only about one-third of the primary students continue into the secondary level. Yet, in spite of such problems, Guyana has the relatively high literacy rate of 85 percent. There is a teachers' training college in Georgetown. The Queen's College, founded in 1844, is government maintained. A technical institute has made progress since 1945, and the Carnegie School of Home Economics has had a long and distinguished history. In 1963 the University of Guyana was established in Georgetown. Well-organized educational broadcasting aids the schools.

Health and welfare services

The standard of health is generally good, and the control of malaria and other formerly endemic diseases, such as whooping cough, measles, tuberculosis, and leprosy, is effective in populous areas. Low dietary standards are being gradually raised by education and public-health education. Medical facilities include almost 150 medical

centres, including rural maternal and child-care clinics. There are about 24 general hospitals, 12 hospitals on the sugar estates, and specialized units for tuberculosis, psychiatric illness, and leprosy, as well as mobile health units. In 1969 there was one doctor for every 3,800 inhabitants.

Public health has long been well organized, and food hygiene, malaria control, and drainage receive high priority. Housing and reconstruction are promoted by a special ministry. Community development, or self-help, programs encourage communal projects and have been successful in getting people of different ethnic origins to work together. There are systems of social security, including old age pensions, and public assistance.

Housing presents a critical problem. Most of the houses in Guyana are of poor construction and are crowded; few have adequate sanitation or electricity services. Most are constructed of wood, roofed with corrugated iron or wooden shingles, and raised on posts or concrete blocks. Housing on the plantations includes tenements constructed of old barracks, and there are many tenements in Georgetown. The government has launched ambitious housing schemes, but few units have actually been constructed.

The 2,000-man Guyana Police Force (GPF) is an armed semi-military organization. It is charged with the maintenance of law and order, protection against fire, and civil defense.

Social conditions. In the sugar industry, price-rate wage systems prevail, and incentive schemes are particularly directed at improving the regularity of work. The sugar industry is by far the largest employer, but its labour force of 17,000 is being steadily diminished by mechanization. The bauxite industry is a relatively small employer, and rice milling is on a small scale except where connected with government irrigation schemes. The timber industry is the only other considerable employer.

Sugar and rice growing are seasonal occupations, so there has always been unemployment, underemployment, and casual labour. The rapid rise of population has aggravated the problem, which is being countered by efforts to make more drained land available and to introduce new local industries. Problematic inflation has given rise to labour unrest, and wages are not commensurate with the cost of living.

Social and economic divisions

The national social structure was inherited from the period of British colonial rule. The elite is composed of Europeans and those East Indians, Afro-Guyanese, and Portuguese of wealth, education, or the professions. The middle income group is not well defined, and social mobility tends to be related to occupation rather than to ethnic identity. The lower income group is composed of East Indian and Afro-Guyanese peasants, industrial workers, and artisans. The Amerindians remain apart from the country's main social stream. These social and economic groups are further complicated by the growing awareness of ethnic identity and the social stratification within each ethnic community.

Cultural life and institutions. Guyana is in the process of evolving a distinct national culture based on folk traditions and the contributions of the ethnic communities. Amerindian culture is important to music and dance; the landscape has long been central to painting and sculpture; and the literary arts reflect Guyana's history and its path to independence. The most impressive progress has been in the literary arts. Major writers include the original and sensitive Wilson Harris, the poet A.J. Seymour, and the novelist Edgar Mittelholzer.

Cultural institutions. Cultural institutions are concentrated in Georgetown. They include the Guyana Museum, which contains exhibits relating to industry, art, history, anthropology, and zoology, and the Guyana Zoo, with its collection of South American animals. Libraries include the Public Free Library, the Royal Agricultural and Commercial Society Library, and those of the British Council and the United States Information Service.

Press and broadcasting. Newspapers include three dailies, four Sunday papers, five weeklies, two fortnightlies, three monthlies, and three quarterlies. The *Guyana Graphic* of East Demerara is the major paper and is particularly influential among the Afro-Guyanese middle income group. The *Evening Post* is also published in East Demerara and is important among the trading community. The *Mirror* is the third daily. The weekly newspapers include *The New Nation*, an organ of the PNC; *The Sun*, an organ of the UFC; and the *Labor Advocate* of the Man-Power Citizens Association.

Radio services

There are two major radio broadcasting stations—Radio Demerara and the Guyana Broadcasting Service, which was purchased by the government in 1968. United Broadcasting also operates a broadcasting station. Programming includes school broadcasts, dramas, religious and ethnic programs, and sports. In 1968 television services were obtained from stations in Surinam, Trinidad, and Venezuela.

PROSPECTS FOR THE FUTURE

Guyana has had a relatively stable political life since achieving independence in 1966, and is seeking to establish its international position. The prospects of attracting foreign aid seem favourable, and finance for surveys, research, and prototype schemes is not difficult to acquire. The country's basic problems, however, remain much as they have been since the 19th century—diversification of the economy and development of natural resources. Agricultural development is necessarily slow, and it will probably remain true that Guyana's greatest asset is in the fertility of its coastal plain, of which only a small portion is now fully exploited. Successful forestry continues to be difficult, and future mining development will depend upon more intensive exploration with improved methods.

If political stability continues and a diminution in internal racial tensions can be achieved, the country could become an important settlement area for the overcrowded island populations of the surrounding Caribbean countries.

BIBLIOGRAPHY. SIR ROBERT H. SCHOMBURGK, "Reports to the Royal Geographical Society," *Geogrl. J.* (1836–45; substantially reprinted under the title *Travels in Guiana and on the Orinoco*, ed. by WALTER E. ROTH, 1931), represents a minor classic of 19th-century exploration. A botanical study of the same period is found in RICHARD SCHOMBURGK, *Travels in British Guiana, 1840–1844*, 3 vol. (1847–48; Eng. trans. by WALTER E. ROTH, 2 vol., 1922–23). Other historical works are C.A. HARRIS and J.A.J. DE VILLIERS (eds.), *Storm van's Gravesande . . .: The Rise of British Guiana* (Hakluyt Society, 2 vol., 1911); JAMES RODWAY, *Guiana: British, Dutch and French* (1912); SIR CECIL CLEMENTI, *The Chinese in British Guiana* (1915); DWARKA NATH, *A History of Indians in British Guiana* (1950); and RAWLE FARLEY, "The Rise of the Peasantry in British Guiana," University of the West Indies *Social and Economic Studies*, 2:87–103 (1954). Studies of the indigenous population include WILLIAM H. BRETT, *The Indian Tribes of Guiana* (1852); SIR EVERARD F. IM THURN, *Among the Indians of Guiana* (1883, reprinted 1967); and WALTER E. ROTH, *An Inquiry into the Animism and Folk-Lore of the Guiana Indians* (1915, reprinted 1970). RAYMOND T. SMITH, *British Guiana* (1962), is a valuable sociological work; and PETER NEWMAN, *British Guiana: Problems of Cohesion in an Immigrant Society* (1964), is also important. For natural resources, see VINCENT ROTH, *Handbook of the Natural Resources of British Guiana* (1946) and *Notes and Observations on Animal Life in British Guiana, 1907–1941* (1941); SMITH BRACEWELL, *Geology and Mineral Resources of British Guiana* (1947); *Report of the Commission of Enquiry into the Economic and Settlement Possibilities of British Guiana and British Honduras* (HMSO, Cmd 7533, 1948); and DENNYS B. FANSHAWE, *The Vegetation of British Guiana* (1952). MICHAEL SWAN, *British Guiana* (1957), presents the only fairly recent general account of the country; his *Marches of El Dorado* (1958), is a lively account of travel in the interior savannas. L.F.S. BURNHAM, *A Destiny to Mould* (1970), is a recent history written by Guyana's Prime Minister. For statistics, see the *Guyana Quarterly Statistical Digest*. A vast quantity of information may also be found in the *Reports of the British Guiana Geological Survey*, the *Annual Colonial Reports*, *Agricultural Journal of British Guiana*, and their counterparts renamed for "Guyana" since independence in 1966.

(Ed.)

Gymnastic and Weightlifting Sports

Gymnastic sports involve physical exercise, specifically systematic and usually rhythmic exercises (calisthenics) and performances on rings, bars, and other apparatus to promote strength, suppleness, agility, coordination, and body control. They also include the lifting of weights to increase strength and improve physical fitness. This article deals principally with the organized, competitive sports, gymnastics and weight lifting, which in modern times have evolved from these activities.

While gymnastics and weight lifting have common characteristics, especially in their use in physical conditioning, there are distinct differences between them. Weight lifting, as a competitive sport, has been practiced only by men, and it is generally associated with the quality of bodily strength, whereas both men and women engage in the sport of gymnastics, which is well-known for developing suppleness, grace, and coordination. Further, competitive weight lifting is an objective matter because the results are determined by a specific measurement—*i.e.*, the number of pounds that are lifted—while the execution of gymnastic exercises is entirely subjective, for it is practiced without the aid of any mechanical measurements, and results depend entirely upon human judgment and interpretation.

GYMNASTICS

Some form of gymnastics—particularly calisthenics, or floor exercises—is practiced in connection with most other sports. Systematic exercises involving movements of the arms, legs, torso, and other parts of the body are employed for loosening up muscles prior to engaging in competition and also for long range physical conditioning before the start of such competitive team sports as football, baseball, and soccer.

The competitive aspect of gymnastics differs from that of physical conditioning, for it not only is a form of physical conditioning in itself but is the end result of such exercise. It consists of an infinite variety of movements on the various apparatus and on the floor, requiring individual ingenuity in their composition, skill in their performance, and harmony in their combination.

The countries that have been strongest in competitive gymnastics in recent years are Japan, the Soviet Union, the United States, East and West Germany, Switzerland, and Czechoslovakia. They all have instituted sound programs to develop gymnasts of international calibre, starting at the high school level. Other countries have followed suit, and now gymnastics as a competitive sport is practiced throughout the world.

History. The sport of gymnastics evolved from the rudimentary exercises performed as training for strenuous combative sports that took place during the public games of ancient Greece. At that time, the general term gymnastics included activities that have since developed as separate sports—for example, track and field athletics, fencing, wrestling, and boxing. During the ancient Olympic Games, a primitive form of modern gymnastics was developed, and the sport came into its own as a part of those games.

Origins in ancient games

With the termination of the ancient games, all sports fell into a decline. In the Middle Ages jousts and various field sports were popular, but the systematic training of the body that the Greeks had associated with gymnastics fell into neglect (see also ATHLETIC GAMES AND CONTESTS).

The modern development of gymnastics started in the 19th century, when there was a revival of interest in all sports; and gymnastics early came to be recognized as a systematized form of physical exercise, having not only recreational but therapeutic value and offering a means of developing a high degree of discipline of both mind and body.

Gymnastic societies were founded first in Germany (Turnvereins) and in the Bohemia of the Austro-Hungarian Empire (Sokols); these were followed in France and Switzerland (where a system of gymnastics performed in unison by groups was developed and is still practiced). Similar societies gradually spread throughout western Europe. Gymnastics was one of the first sports to recognize that its recreational and therapeutic advantages were as valuable to women as to men, and the European gymnastic societies provided children's classes, starting youngsters as early as the age of five.

The original impetus in the United States was in the 1880s with the advent of the great tide of immigrants from the European countries. They brought with them the ideas of their gymnastic societies and founded the societies wherever they settled. The younger generations carried the influence of these gymnastic societies into the athletic programs of their schools and colleges, where the sport gained recognition competitively on a large scale.

Development as an organized sport. Internationally, gymnastics is governed by the Fédération Internationale de Gymnastique (FIG), which was founded in 1881 to formalize international competitions. FIG, which has a membership of more than 60 countries, each of which has its own national governing body, or federation, is administered by an executive committee (Comité Directeur) and by men's and women's technical committees that promulgate the rules and regulations governing competitions. The committees also prescribe international standards for the various apparatus to be used, the types of exercises to be performed, and the conduct of the competitions.

Modern competitive gymnastics has developed from two major systems. In Sweden a system of free exercises on the ground arose in the mid-19th century to develop perfect rhythm of movement. This system was introduced in England in 1879 and in the United States in about 1889; and for the next 20 years the relative merits of Swedish (rhythmic) gymnastics and the German gymnastic system (apparatus work of a formal nature, stressing muscular development) were debated intensively by both the educational authorities and adherents of the sport. A solution was found by the Fédération Internationale de Gymnastique in the 1920s, in blending the rhythmic, fluent movements of the Swedish system with the precision and developmental emphasis of the German system.

Major systems of gymnastics

Although the Swedish system is still popular in the country of its origin, the combined version has been adopted by national authorities elsewhere. Current trends, however, are to lean more and more to the aesthetic and fluent movements, rather than to movements of strength and the holding of static positions.

Competitions in gymnastics are conducted at all levels, from local school and gymnastic societies to national championships, international meets, world championships, and the Olympic Games. The Fédération Internationale de Gymnastique, which is recognized as the international authority by the International Olympic Committee and, as such, has jurisdiction over the gymnastic competitions of the Olympic Games, also conducts the world championships in gymnastics every fourth year (two years after each staging of Olympic Games).

Principles of competitive gymnastics. There is constant evolution of the sport. The innovation of movements—never seen before in gymnastic competitions—by gymnasts of international repute tends to popularize such movements in international competitions. Such innovations have a considerable influence on the competitive sport because they encourage originality and a greater expansion of the types of exercises that can be performed on the various apparatus and on the floor. For instance, a movement on the horizontal bar by Josef Stalder of Switzerland never seen previously in international competition has since been known as the "Stalder Shoot"; a vault on the vaulting horse by Haruhiro Yamashita of Japan in the 1964 Olympic Games in Tokyo has ever since been known as the "Yamashita"; and as another example, a movement on the parallel bars in the 1966 Fédération Internationale de Gymnastique World Championships in Dortmund, West Germany, by Sergey Diomidov of the Soviet Union, gave rise to its being included in the nomenclature of movements on the parallel bars as a "Diomidov."

Innovative aspects of the sport

As a competitive sport, gymnastics is akin to diving and figure skating, for it is a demonstration sport, and the effectiveness of the competitor is assessed solely by the judgment of officials who have a knowledge of the technical rules and regulations governing the competition. In other words, these demonstration sports are competitions in which there is no physical contact with an opponent to guide an official, such as a touch of the foil that occurs in fencing, a knockdown in boxing, an advantage in wrestling—and where there is no mechanical measurement to evaluate the effort of the competitor, such as timing a race, measuring the height or distance of a jump or throw, or weighing the number of pounds lifted.

Men's events and apparatus. The men's competition in international gymnastics consists of seven events: the horizontal bar, parallel bars, side or pommelled horse, long or vaulting horse, stationary rings, floor exercises, or calisthenics, and the all-around event consisting of the combined scores in the first six events. (International rules require that a gymnast perform in all events similar to the decathlon event in track and field.)

Each competitor performs two exercises in each event. One is a prescribed exercise that all competitors must perform, and it is composed by the governing body of the sport (for international competitions, the FIG). Its performance is judged solely on its execution—*i.e.*, the form of the gymnast, the fluency of his performance, the correctness of the execution, and the beauty of combining the component parts of the exercise.

Prescribed and optional exercises

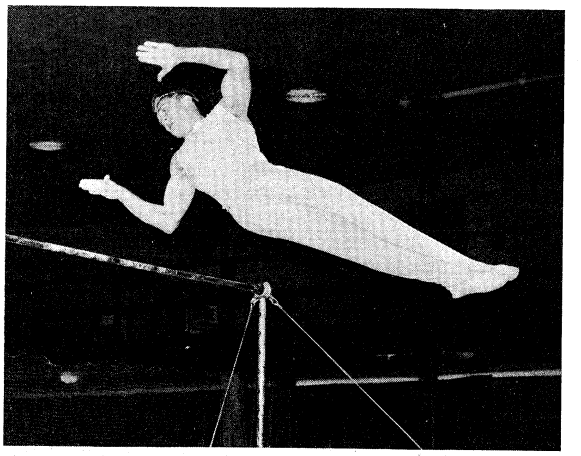
In addition, an optional exercise, composed by the gymnast himself, must also be performed. The element of difficulty of the component movements of the exercise enters into the evaluation of the optional exercise. Besides difficulty, other elements evaluated are originality, beauty of combining the various movements, and the fluency and perfection of execution. There is a minimum of 11 parts, or skills, required for a complete optional exercise.

The requirements of the events and the apparatus used for the men's competition are described below.

Horizontal bar. The horizontal bar is a polished steel bar, 28 millimetres (about 1.1 inches) in diameter, 2.35 to 2.50 metres (7.7 to 8.2 feet) long, and 2.40 to 2.50 metres (7.9 to 8.2 feet) high from the floor, supported at the ends by steel poles that are held upright by guy wires. Only movements of swinging and vaulting, without pause, are permitted. Such exercises as upstarts, or kips, back and front (*i.e.*, moving with an impetus from a hanging position in which the hips are bent and the stretched legs and feet are stretched over the head, to a support position with the head above the feet and the body either in front or back of the bar); uprises (in which the gymnast swings himself from a hang to a support position above the bar); giant circles forward and backward (rotating around the bar from a handstand position with the arms fully extended), with changes from one grip to another; vaulting over the bar, releasing the grip and regrasping the bar; movements requiring turns and changes of body position and the releasing and regrasping of the bar; and finishes, or dismounts, with straddles over the bar, or forward and backward somersaults from the bar are performed on this apparatus.

Parallel bars. The parallel bars consist of two bars made of wood, oval in form and 42 to 48 centimetres (17 to 19 inches) apart, 3.50 metres (11.5 feet) long, and 1.60 to 1.70 metres (5.2 to 5.6 feet) high. The requirements for this apparatus are movements combining swings, vaults, strength, and balances, although swings and vaults must predominate. There may not be more than three balances (such as handstands) in an exercise (a sequence of at least 11 different movements or tricks), and there must be at least one movement of strength (such as a handstand that is pressed into by strength, rather than swung up into). Movements below the bars and the release and regrasping of the bars are also required.

Side, or pommelled, horse. This apparatus is a leather-covered form 1.60 metres (5.2 feet) long, 35 to 37 centimetres (14 to 15 inches) wide, and raised 1.08 metres (3.5 feet) from the floor (measured to its top) by a support in its centre. Curved wooden pommels, 12 centimetres (five inches) high, are inserted in the top of the horse 40 to 45 centimetres (16 to 18 inches) apart in the centre, or "saddle" of the horse. The gymnast supports himself on the pommels and performs movements of the trunk and legs without stops, such as single or double leg circles and crosses of the legs (scissors) with turns and changes of the grasp to the forward (neck) part of the horse, the centre (saddle), and the rear (croup). It is necessary to vary the movements and also to perform them both to the left (clockwise) and to the right (counterclockwise).

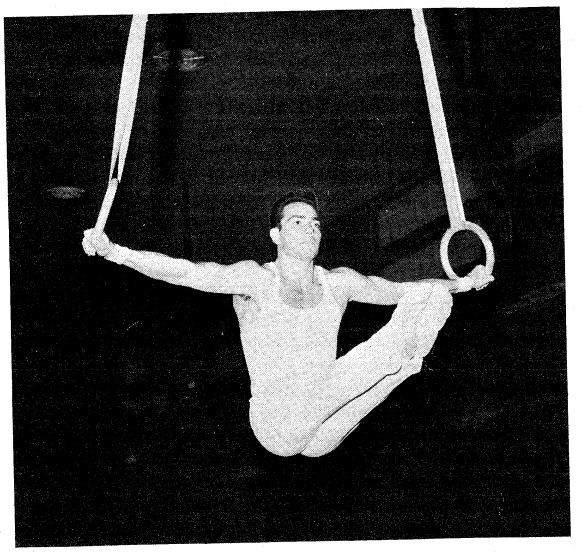


Horizontal bar.
A full turn (pirouette) at the end of a backward swing (releasing grasp), performed by Bruno Klaus (U.S.).
John Nicolas

Long or vaulting horse. This is the same device that is used for the side horse, except that the pommels are withdrawn and the horse is placed lengthwise, so that the part nearest to the gymnast's stand is the croup, and the neck is farthest away from him. The height of the horse is raised to 1.35 metres (4.4 feet). An elastic board, or springboard, is placed in front of the croup end of the horse.

The gymnast runs toward the horse, rebounds off the elastic board and, supporting his hands on the horse, vaults over it. A variety of vaults is performed, such as vaulting over with straddled legs, with the legs together and bent into a squatting position, or with the legs straight and the hips bent, Handsprings, cartwheels, and other vaults are performed over the horse. The hands may be placed on either the croup or the neck end of the horse, but the space for the support of the hands is limited and encroaching beyond this space incurs a penalty on the scoring of the vault. Each vault is evaluated by a table as to its difficulty.

Stationary rings. The stationary rings are made of wood, 28 millimetres (1.1 inches) in thickness and 18 centimetres (seven inches) in diameter (inside). They are suspended by straps at a height of 5.50 metres (18 feet) from the floor, with the rings themselves hanging 2.40 to 2.50 metres (7.9 to 8.2 feet) above the floor. The exercises must be performed with the rings in a stationary position (without a swinging or pendulum movement of the rings), and they must combine swinging movements of the body, strength and holding of positions. There must be at least two handstands in each exercise, one of which must be pressed into by strength with the body held above the rings. In the other, the athlete swings up into the handstand from a hanging position below the rings. At least one other example of strength must be exhibited, such as a cross (holding the body vertical with the arms fully stretched sideways), or a lever (hanging with straight arms with the body stretched horizontally, parallel to the floor, either in front of the arms or in a dislocated hang; *i.e.*, hanging in an inverted position with the head nearest the floor and then lowering the body backward, stretching it in a horizontal position). Combined with such required move-



Rings.
A cross hang, with legs in "L" position, performed by Abie Grossfeld (U.S.).
John Nicolas

ments may be either forward or backward upstarts, forward or backward uprises, straight or inverted hangs, and circles of the body, either forward or backward.

Floor exercises, or calisthenics. No apparatus is used in the floor exercises, or calisthenic event. The exercise is done on the floor, which may be covered by canvas or felt, in an area 12 by 12 metres (39 by 39 feet), and it must be at least 50 seconds and not more than 70 seconds in duration. The type of exercise required is a series of combined movements that demonstrate flexibility (tumbling movements such as handsprings, cartwheels, somersaults in the air), jumps, strength, holding of poses, balances, and other tricks. The exercise must be performed with rhythm and harmony, and the gymnast must move in different directions and utilize a major portion of the allotted area.

The exercise usually starts and finishes with a series of tumbling movements, such as a cartwheel with a half a turn (a roundoff) and continues with handsprings and somersaults backward or forward in the air. In between the start and finish, balances are held on one leg (scales), handstands are pressed into by strength, and exercises resembling ballet and tumbling movements and jumps are interposed.

Competition for women

Women's events and apparatus. Generally the basic requirements for the women's competition are the same as those for the men as regards the prescribed and optional exercises and their evaluation. The women's events are performed on the balance beam, the uneven parallel bars, and the vaulting horse; women's floor exercises, like the men's, are performed without apparatus.

Balance beam. The balance beam is a wooden beam five metres (16 feet) long and ten centimetres (four inches) wide, which is raised 120 centimetres (about four feet) from the floor. The performer begins the exercise by mounting the beam either by a vault or a jump and executes movements that must include steps, running, jumps, turns, sitting and lying positions, and some held, or posed, positions. The duration of the exercise is from one minute 20 seconds to one minute 45 seconds.

Uneven parallel bars. These are of the same dimensions and construction as the men's parallel bars, except that the top bar is 2.30 metres (7.5 feet) above the floor whereas the lower bar is only 1.50 metres (4.9 feet) high. This apparatus is the latest to be developed, having been used first in the 1936 Olympic Games; it permits a great variety of movements, although hanging and swinging exercises predominate (such as hanging from the higher bar and swinging forward to the lower bar), interspersed with support and balance movements. The performer uses either of the bars or combines movements using both bars.

Vaulting horse. This is the same as the men's long horse, except that it is only 1.10 metres (3.6 feet) high and is placed sideways instead of lengthwise. Women also use the elastic board and perform vaults similar to those done by men, except that they are done over the width of the horse, rather than its length.

Floor exercises or calisthenics. The women's floor-exercise event is similar to the men's but it is performed to music and lasts from one to 1.5 minutes.

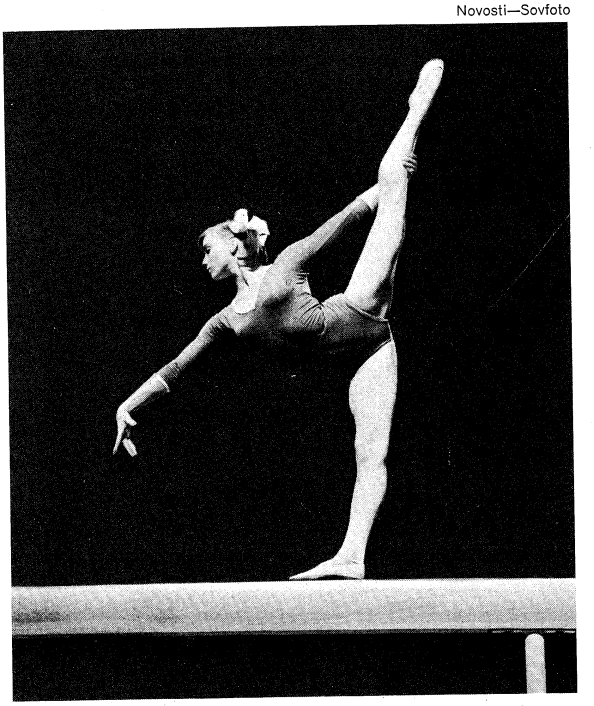
Scoring and judging. Although the scoring and judging of gymnastic competitions is solely subjective, there are detailed and comprehensive rules and regulations promulgated by the FIG that attempt to eliminate the factor of human judgment and provide guidelines for as objective an evaluation as is possible without the aid of mechanical measurements.

Evaluation and scoring of individual and all-around events

Four judges supervised by a superior judge evaluate and score each exercise on a one-tenth basis; *i.e.*, 8.8, 9.3, and so on, with 10.0 given for a perfect exercise. For the prescribed exercise, the entire ten points are confined to its execution. In the evaluation of the optional exercise, 3.4 points are allotted for the difficulty of the exercise, 2.6 points for the element of combination of the various movements, and 4.0 points for the execution of the exercise as a whole. All four judges evaluate and score each exercise independently. Then, in order to eliminate gross errors of judgment, the scorers delete the highest and the lowest marks of the four judges, and the two middle marks are averaged for the score of the exercise. For example, if the four judges' marks are 9.6, 9.2, 9.1, and 8.7, the 9.6 and 8.7 are deleted, and the 9.2 and 9.1 are averaged, giving the final score of 9.15. The average score of the prescribed exercise is added to that of the optional exercise to determine the final score of the competitor in the event and his standing as compared with the other competitors. To obtain the all-around score, the final scores of the six events (four in the case of women) are added, and the total comprises the score of the gymnast in the all-around competition.

The championships of the individual events are determined by a separate competition known as the "finals" wherein the gymnasts who have achieved the six highest scores in each event perform an additional optional exercise in that event, and the score of this final exercise is added to the gymnast's previous score to determine the standings of the six best competitors in each event.

Novosti—Sovfoto



Balance beam.
A front scale, with a vertical split of the legs, performed by Larisa Petrik (U.S.S.R.).

For Olympic champions, see ATHLETIC GAMES AND CONTESTS. For records of gymnastic world championships, see also SPORTING RECORD in the *Ready Reference and Index*.

Other forms of gymnastics. *Trampoline.* In trampoline tumbling, or rebound tumbling, the performer bounces upward from a resilient platform or bed and does acrobatics while in flight. It is performed on a table-height metal frame three by five metres (ten by 17 feet), with an elastic nylon webbing bed attached by a steel-spring suspension system. Competition consists of one compulsory and one voluntary routine, with judging similar to that in conventional competitive gymnastics. The sport has been especially popular in the United States. Annual world trampoline championships for men and women were instituted in 1964. For a list of champions, see also SPORTING RECORD in the *Ready Reference and Index*.

Moderne Gymnastics. A concept of women's gymnastics termed Moderne Gymnastics is an expression of very free movements. The floor exercises, done to music, permit more fluidity, dance, and ballet movements. The other events on this program are performed with hand apparatus, such as a *corde* (similar to a skipping rope), a hoop (rolling hoop), a ball, and a length of cloth ribbon. Another feature of this program is a team drill in which the entire team (six women) performs as a unit and is judged according to its execution, uniformity, difficulty of composition of the drill, harmony, and similar factors. Although the Moderne Gymnastics competition is recognized internationally by the Fédération Internationale de Gymnastique, and world championships in this program have been held biennially since 1963, it is not included in the Olympic program.

Gymnastic festivals and demonstrations. This form of gymnastics is popular mainly in European countries, where it is practiced by groups ranging from a few gymnasts to mass demonstrations of many thousands. The Sokols, gymnastic societies in Czechoslovakia, have promoted gymnastic festivals in which as many as 10,000 gymnasts participated simultaneously in a mass calisthenic drill. The Turnvereins of Germany and the gymnastic societies of Switzerland and the Scandinavian countries also participate in mass demonstrations, not only in calisthenic drills but also on apparatus.

Gymnastic societies and gymnaestrada

The Fédération Internationale de Gymnastique recognizes the value of and encourages such noncompetitive aspects of gymnastics on national and international levels. For this purpose the FIG established in 1953 the international Gymnaestrada, sponsored quadrennially the year after each Olympiad, at which all member federations of the FIG are invited to demonstrate a noncompetitive gymnastic activity, preferably native to their country.

In addition to the internationally recognized programs of competitive events, local and national programs may include other activities such as tumbling, rope climbing, and exercises on the swinging rings. (J.F.H.)

WEIGHT LIFTING

Weight lifting is a competitive sport controlled by the International Weightlifting Federation, which is responsible for the technical rules governing the sport, the organization of annual world championships, continental championships, international tournaments and matches, recognition of world and Olympic records, and appointment of international referees. Weight lifting is included in the Olympic Games and all regular regional games.

History. Throughout his history, man has been interested in strength and trials of strength. For many prehistoric tribes the traditional test of manhood was to lift a special rock. Such stones, with centuries-old inscriptions giving the name of the athlete who first lifted them, can still be found in Greece. Manhood stones, or *clach cuid fir*, have been found in some ancient Scottish castles, indicating that this activity persisted through the Middle Ages. In Munich, a plaque above a stone in the courtyard of the Apothekerhof, weighing 364 pounds (165 kilograms), records that in the year 1490 Duke Christopher of Bavaria lifted and threw it. Stone-throwing competitions survive in West Germany and Switzerland, and rock lifting is still a popular rural activity in the Basque area of Spain.

Modern weight lifting began in the late 18th and 19th centuries, fostered by the feats of professional strongmen such as Eugene Sandow and Arthur Saxon of Germany, George Hackenschmidt of Russia, Louis Apollon of France, and others who performed in circuses and music halls. Their feats were copied by enthusiastic amateurs, mostly in western Europe. The first open world championships were held at the Café Monico, London, in 1891, and won by the Englishman Lawrence Levy. Two weight lifting competitions were included in the first revival of the Olympic Games, held in Athens in 1896. The winner of the one-hand lift competition was Launceston Elliot of Great Britain with 156½ pounds (71 kilograms); and the winner of the two-hand lift competition was Viggo Jensen of Denmark, with 244½ pounds (111 kilograms).

Beginnings of modern weight lifting

The one- and two-hand events were held again in 1904; then, in 1920, at the suggestion of the International Olympic Committee, the International Weightlifting Federation was formed to control the sport and, in particular, to formulate technical rules so that competitions in the Olympic Games could be held under universally accepted conditions. Since then, weight lifting has been included in the Olympic Games. The international federation originally included 14 member countries, but in the mid-1970s there were approximately 100 countries affiliated from all parts of the world. France and Germany were the leading nations until the 1930s, when Egypt became the leader. After World War II the United States was supreme until 1953. Since then, weight lifters of the Soviet Union have been world-team champions and have held most of the world records.

Principles of competitive weight lifting. *Equipment.* The weight used in modern weight lifting is a barbell, a steel bar, or rod, to which metal (cast-iron) disk weights are fastened at each end on a revolving sleeve. The range of weights is 25 kilograms (55 pounds), 15 kilograms (33 pounds), 10 kilograms (22 pounds), five kilograms (11 pounds), 2½ kilograms (5½ pounds), 1¼ kilograms (2¾ pounds). The lifts are performed on a wooden platform four metres (13 feet 1¼ inches) square. If a lifter steps off this platform during the course of a lift, that lift will not be passed.

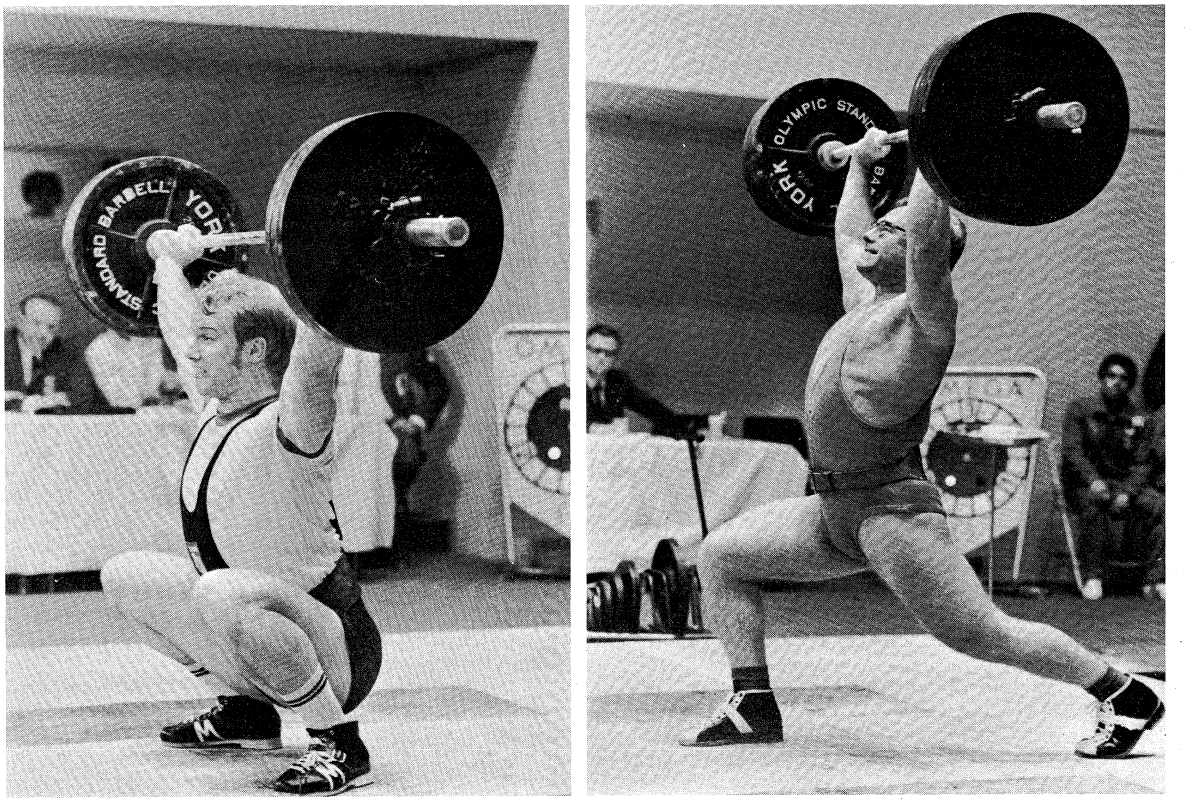
Weight categories. Originally there were five bodyweight categories for the contestants but others were added in 1947, 1951, and 1968, and there are now nine: flyweight 52 kilograms (114½ pounds); bantamweight 56 kilograms (123¼ pounds); featherweight 60 kilograms (132¼ pounds); lightweight 67.5 kilograms (148¾ pounds); middleweight 75 kilograms (165¼ pounds); light heavyweight 82.5 kilograms (181¾ pounds); middle heavyweight 90 kilograms (198¼ pounds); heavyweight 110 kilograms (242½ pounds); super heavyweight over 110 kilograms (over 242½ pounds).

The Olympic lifts. Originally, in International Weightlifting Federation competition five different lifts were used. These were reduced to three in 1928 and finally to two in 1972. The one eliminated in 1972 was the press (or clean and press), which is a two-part lift. The lifter first hefts the barbell to his shoulders in a single, clean movement. He may move his feet to do this. When he is standing erect, the chief referee gives him a signal to complete the lift, which he does by pressing the barbell in a steady, continuous movement to arm's length overhead, without any assistance from the legs.

The press, snatch, and clean and jerk

Only two lifts may now be called Olympic lifts, the snatch and the clean and jerk. In the snatch, the barbell is lifted from the floor to arm's length overhead in a single, continuous, explosive movement, with the lifter being permitted to move his feet or to squat under the barbell as he lifts it, after which he must recover to an erect position again.

The other remaining lift, the clean and jerk, is a two-part lift. After lifting the barbell to his shoulders, the lifter jerks it overhead to arm's length in his own time and



Weightlifting.
(Left) The snatch (squat style), performed by B.O. Johanson (Sweden). (Right) The jerk, performed by Gyozo Veres (Hungary).
Oscar State

without any restrictions on leg action. In both lifts, the lifter must finish with his feet in line, body erect, arms and legs extended, the barbell under control overhead, and he must wait for the referee's signal to lower the barbell back to the floor.

In the clean and the snatch, two main techniques are employed, either the split or the squat. In the split style the lifter lunges under the barbell, thrusting one foot forward and the other foot simultaneously backward to land in a fore-and-aft straddle position. This is also the invariable technique for lifting the barbell in the jerk. In the squat technique, the lifter bends both knees and drops vertically below the barbell into a low, squatting position with the trunk erect. The squat technique is more often used because it is mechanically superior to the split.

Judging of competition. The competition begins at the lowest weight specified by the lifter, and the weight of the barbell is then increased, as required, in multiples of 2½ kilograms (5½ pounds). Each lifter takes his turn as the barbell reaches the weight that he requires, but he is allowed only three attempts on each of the three lifts. The increase in weight from the first to the second attempt must be at least five kilograms (11 pounds) and from the second to the third attempt at least 2½ kilograms (5½ pounds). If a competitor fails a lift and makes another attempt, he may either take the same weight again or increase it, as he wishes. His best results on each of the three lifts are added to give a total. The lifter with the highest total is the winner. In case of a tie in totals, the lifter with the lighter body weight is the winner, and the rest are placed according to their totals. All lifters in official championships are weighed 1¼ hours before the competition begins. Two referees and a chief referee adjudicate the lifts and signal their decisions by means of a system of lights: a white light indicates a "Good Lift" and a red light "No Lift." A majority of two is equally decisive as a unanimous verdict.

For Olympic champions see ATHLETIC GAMES AND CONTESTS. For world weight lifting records see also SPORTING RECORD in the *Ready Reference and Index.*

Weight training. Weight training is distinguished from weight lifting in that it is not a competitive sport but a system of physical training using weights in the form of barbells and dumbbells and other specialized equipment such as pulley weights and leg machines. It is used extensively by track and field athletes, swimmers, football and soccer players, tennis players, and other sportsmen for whom basic strength is an important factor in their overall training program.

Weight lifting for strength and physical fitness

Other weight trainers use weights in a moderate program to obtain general physical fitness. There are also large numbers who train with weights to develop their physique in order to enter body-building contests. In the mid-1970s the medical profession was making increased use of weight training for physical rehabilitation after illness and injury. (O.S.)

BIBLIOGRAPHY. RICHARD M. ARONSON (ed.), *The Art and Science of Judging Men's Gymnastics* (1970), articles by recognized authorities on the sport, giving their interpretations of the rules and regulations contained in the FIG Code of Points, and their experiences in officiating.

Official handbooks: FEDERATION INTERNATIONALE DE GYMNASTIQUE, *Men's Code of Points* (1968, with supplements); *Women's Code of Points* (1970, with supplements); *Moderne Gymnastic Code of Points* (1970, with supplements), detailed technical rules and regulations for judging exercises and competitions; *Measurements, Dimensions and Forms of Competitive Apparatus* (1965, with supplements); *List of Results of FIG World Championships, Olympic Games Gymnastic Competitions, and European Championships* (1948–).

Periodicals: FIG Quarterly Bulletin, reports, results of international matches, and articles regarding new gymnastic techniques and changes in rules and regulations; *Olympische Turnkunst,* contains comments and opinions of international authorities regarding competitions and personalities in the sport; *AAU Gymnastic News Letter, Modern Gymnast,* and *Mademoiselle Gymnast,* contain current articles, results of meets, and editorials pertaining to the sport.

Weightlifting: FEDERATION HALTEROPHILE INTERNATIONALE, *FHI Constitution and Rules,* rev. ed. (1972), rules for competition; G.W. KIRKLEY, *Modern Weightlifting,* 2nd ed. (1961), a brief history of weightlifting, description of lifts, rules, and training schedules; *Weightlifting and Weight Training* (1964),

a description of lifts, exercises, and training schedules; and with J. GOODBODY, *The Manual of Weight Training* (1967); J. MURRAY and P.V. KARPOVICH, *Weight Training in Athletics* (1956), training schedules for various sports; A. MURRAY, *The Theory and Practice of Olympic Lifting* (1950), detailed analysis of competition lifts; J. MURRAY, *Weightlifting and Progressive Resistance Exercises* (1955), brief descriptions of competition and training lifts and schedules; R. HOFFMAN, *Weightlifting* (1939), a history of American weightlifting and descriptions of several different competition lifts; O. STATE, "Weightlifting," in the *Oxford Companion to Sports and Games* (1972), a history of world weightlifting and a detailed description of competition procedure; D.P. WEBSTER, *Two Hands Snatch* (1964) and *The Development of the Two Hands Clean and Jerk* (1967), analytical descriptions of the lifts; *The Complete Physique Book* (1966), pictorial descriptions of bodybuilding exercises.

(J.F.H./O.S.)

Gymnosperm

Gymnosperms constitute one of the two major groups of seed-producing plants—the other group being the flowering plants, or angiosperms. Although they number only about 750 species, gymnosperms rank among the most familiar, important, and biologically interesting of all plants; they include such well-known species as pines, spruces, cedars, junipers, araucarias, and kauris. They are the major constituents of the vegetation in many parts of the world. Many species of gymnosperms are significant as timber trees, as ornamental plants, as food plants, as sources of essential oils and other products, and as medicinal plants. Gymnosperm forests of prehistoric times formed a large portion of the coal available today. Some gymnosperms are considered weeds, and a few are poisonous. Gymnosperms are of great biological interest because of their diversity of form and structure, their present and past distributions, and their well-preserved and rather complete fossil record.

The major groups of living gymnosperms are the coniferophytes, including conifers, taxads, and the ginkgo; the cycadophytes, including cycads; and the gnetophytes, including joint firs, gnetums, and *Welwitschia* (Figure 1).

ECONOMIC IMPORTANCE

Gymnosperms are among the most valuable timber trees. Outstanding in the Northern Hemisphere are pines, Douglas firs, cedars, and the redwood; in the Southern Hemisphere are kauris, podocarps, dacrydiums, and araucarias. Notable as sources of pulpwood for paper manufacture are hemlocks, spruces, firs, and pines. Gymnosperm wood is "softwood," structurally different from the wood ("hardwood") of angiospermous trees; despite that designation, a few gymnosperms produce wood that is harder and heavier than that of some angiosperms.

Ornamental gymnosperms

Gymnosperms rank high in value and popularity as ornamental plants. Pines, spruces, firs, Douglas firs, cedars, junipers, larches, arborvitaes, hemlocks, yews, podocarps, *Cunninghamia*, *Chamaecyparis*, araucarias, ginkgo, and cycads are well represented in cultivation. Some of these, particularly spruces, junipers, and *Chamaecyparis*, are notable for their many horticultural forms. Of those gymnosperms grown in greenhouses, cycads, araucarias, and kauris are perhaps the most common. Most gymnosperms grow to be too large for use as house plants, but juvenile forms of some, such as the araucarias, are popular. Some gymnosperms—*e.g.*, pines and junipers—are used in the Japanese art of tree dwarfing (bonsai).

Gymnosperms are of only local significance as food plants. Pine "nuts," the seeds of the pinyon, or nut, pines, are a minor article of commerce. Seeds of araucarias, ginkgo, gnetums, and *Torreya* are eaten locally by man. Of all the gymnosperms, however, the cycads have been most used as a food source, especially *Cycas*, *Encephalartos*, *Macrozamia*, and *Zamia*. They yield a starch, either from the seeds or from the stems, important particularly in areas of periodic famine or where the food supply is naturally limited. In the preparation of cycad starch, the seed kernels and the stem core are treated in various ways—often with water—to separate the starch from the fibrous matter and to remove a poison that is present in certain forms. Extraction of starch called Florida arrowroot from *Zamia* stems, a pioneer industry in Florida, continued until the mid-1920s. The only gymnosperm of importance as a flavouring is the common juniper, whose "berries" (*i.e.*, fleshy cones) contain an essential oil widely used to impart a distinctive flavour to gin and other alcoholic beverages.

Some gymnosperms are of prime benefit as sources of food and cover for wildlife. Pines, for example, furnish nutritious seeds for many birds and edible foliage and twigs for browsing mammals. Other gymnosperms of similar significance are spruces, hemlocks, firs, junipers, and joint firs. The seeds of some cycads are eaten by animals; *Encephalartos* seeds are taken in such quantity by baboons in southern Africa that the cycad is threatened with local extinction.

On the other hand, the leaves of pines, cypresses, junipers, yews, and cycads are dangerous to browsing livestock. Poisoning of stock by ingestion of cycads is especially common in parts of Australia, where cycad eradication on ranges has been attempted. Cycad seeds are poisonous to man. The use of cycad-seed and cycad-stem starch for food is, therefore, dangerous unless the poisonous substances are removed or made harmless by special treatment. In areas where cycads are used as food plants, they have been implicated in the high incidence of certain neurological disorders. The attractive seeds of yews, with their scarlet fleshy covering and hard toxic kernel, have caused death in man.

Oils and tannins

Certain gymnosperms, among them spruces, pines, junipers, firs, hemlocks, and arborvitaes, are the source of essential (volatile) oils much used industrially as scents in soaps, air fresheners, disinfectants, pharmaceuticals, cosmetics, and perfumes. These oils, obtained by distillation, are used primarily for their pleasing aromas but also, in certain cases, for their odour-fixative or antiseptic properties. Tannins, basic to the manufacture of leather from animal skins, are derived from many kinds of plants including, among the gymnosperms, hemlocks and spruces, which have a tannin-rich bark.

Natural resins, once much used commercially (especially for the manufacture of varnishes), are now of greatly reduced importance. The best known gymnospermous resin is kauri copal, from *Agathis australis* of New Zealand, obtained largely by digging pieces of resin from the ground in which they have become buried. Baltic amber, a "fossil" resin, was produced by a species of pine now extinct. Among the most important oleoresins (mixtures of resin and essential oil) are the turpentines, derived from pines. Distillation of the oleoresin yields oil of turpentine (the "turpentine" of commerce) and rosin, which have many industrial uses. The turpentine, or "naval stores," industry is well developed in the southeastern United States and in parts of Europe and Asia. Gymnosperms other than pines that yield turpentines include the balsam fir (source of Canada balsam), spruces, and the European larch.

Although many gymnosperms have local medicinal uses, the only ones of commercial importance are certain of the joint firs, the natural source of the alkaloid drug ephedrine, used to treat certain respiratory afflictions.

Extinct gymnosperms—the seed ferns, the cordaitaleans, and the early conifers—of Carboniferous times (280,000,000 to 345,000,000 years ago) contributed greatly to the formation of coal. The lush vegetation built up thick layers of peatlike material that, after deeper burial and compression over millions of years, became coal.

Although most gymnosperms are not considered to be weeds—plants growing where man does not want them—some may be so called under certain conditions. The value of some western United States rangelands has been lowered by invasion by pinyon pines and junipers, which have replaced more desirable forage vegetation. Pines and junipers have a marked tendency to invade abandoned fields, making the re-use of these lands for crops difficult. The aggressiveness of pines, in fact, may enable them to encroach upon vegetation that, for various reasons, is considered more valuable than pine-dominated vegetation.

Junipers are alternate hosts for the fungus that causes



Figure 1: *Diversity of form among gymnosperms.*
(Top left) Giant sequoia (*Sequoiadendron giganteum*), the largest gymnosperm. (Top centre) Ginkgo, or maldenhair tree (*Ginkgo biloba*), native to northern China and Japan, alone constitutes the order Ginkgoales. (Top right) Norfolk Island pine (*Araucaria excelsa*), noted for its unusual growth habit. (Centre) *Encephalartos*, a genus of cycads that produces the largest known cone. (Centre right) *Ephedra*, sole genus of the family Ephedraceae. (Bottom left) Japanese yew (*Taxus cuspidata*), commonly used in landscaping. (Bottom right) Tumboa (*Welwitschia*), a South African gymnosperm noted for its bizarre form, habit, and development.
(Top left) Authenticated News International; (top centre) Eric Hosking—National Audubon Society; (top right) Lyn—Annan Photo Features; (centre) H.R. Allen—EB Inc.; (centre right) Jack Dermid; (bottom left) J.C. Allen and Son; (bottom right) Gordon L. Maclean

cedar-apple rust disease of apple trees and other members of the rose family. White pines (*Pinus*) of North America are seriously affected by the fungus that causes white-pine blister rust; gooseberries and related *Ribes* species are the alternate hosts.

Among the plants whose roots harbour nitrogen-fixing micro-organisms (*i.e.*, micro-organisms that can render atmospheric nitrogen usable by green plants) are the cycads. Although the significance of cycads in the nitrogen cycle is not known, in areas where these plants are abundant they have certainly played a role in the maintenance of the nitrogen supply in the soil.

The Christmas-tree industry is dependent upon various kinds of gymnospermous trees, including Douglas firs,

firs, pines, junipers, and spruces, harvested from natural stands or from plantations.

Conservation uses

In erosion control, gymnosperms—especially conifers—are valuable in both natural and planted vegetation. Pines, araucarias, podocarps, cryptomerias, and larches are much used for reforestation purposes. The role of forests, including those of gymnosperms, in the protection of watersheds is of the greatest importance. Junipers, spruces, Douglas firs, and *Callitris* are much planted in shelterbelts and windbreaks.

NATURAL HISTORY

Distribution and abundance. Gymnosperms are widely distributed, occurring on all continents except Antarctica. They are characteristic especially of temperate latitudes; most of those in the tropics occur in mountains rather than in lowlands. Eurasia has more gymnosperm species than any other continent—about 340. North America has about 150. Africa, South America, and Australia have about 50 species each. Approximately two-thirds of the species of gymnosperms are native in the Northern Hemisphere and about one-third in the Southern. Of the 67 genera, 35 are restricted to the Northern Hemisphere and 20 to the Southern. Three genera of cycads (*Cycas*, *Macrozamia*, and *Zamia*), the joint firs, the gnetums, and seven genera of conifers (*Agathis*, *Dacrydium*, *Juniperus*, *Papuacedrus*, *Phyllocladus*, *Pinus*, and *Podocarpus*) have ranges that cross the Equator. Thirty-nine genera of gymnosperms are restricted to the Old World, 11 to the New World; 17 occur in both.

The genus of gymnosperms that occupies the greatest area is *Juniperus*, the junipers, the range of which extends over much of the Northern Hemisphere from the Arctic to tropical mountains and into the Southern Hemisphere in East Africa. Other genera of wide distribution in North America and Eurasia are *Abies* (firs), *Larix* (larches), *Picea* (spruces), and *Pinus* (pines). The southern genus occupying the greatest area is *Podocarpus*, found from Mexico to southern South America and in Africa, Madagascar, southeastern Asia, Indonesia, Australia, New Zealand, and many South Pacific islands.

In contrast to these widely distributed genera are many of limited distribution—*e.g.*, *Sequoia*, confined to California and Oregon; *Sequoiadendron*, to California; *Microcycas*, to Cuba; *Saxegothaea*, to Chile and Patagonia; *Stangeria*, to southern Africa; *Bowenia*, to northeastern Australia; *Microcachrys*, to Tasmania; *Neocallitropsis*, to New Caledonia; and *Taiwania*, to Taiwan.

Gymnospermous trees reach more northerly latitudes than do angiospermous trees. On the fringes of the Arctic (*i.e.*, along the continental "tree line") in North America and Eurasia are only two kinds of trees, spruces and larches, both gymnospermous. Gymnosperms grow at almost the highest altitudes reached by vascular plants. A joint fir has been found at 5,350 metres (about 17,500 feet) in Kashmir.

Gymnosperms, especially the conifers, are dominant plants in vegetation in many parts of the world but mainly in the Northern Hemisphere. The coniferous forest of greatest extent is the so-called boreal forest in northern North America and northern Eurasia south of the Arctic. South of the boreal forest—in mountains of the north temperate zone—are also extensive forests of conifers, as in the Rocky Mountains, the Alps, and the Himalayas. Conifers are the dominant forest trees in many parts of the southeastern United States. Genera well represented in these Northern Hemisphere coniferous forests include *Picea*, *Pinus*, *Abies*, *Larix*, *Tsuga*, *Juniperus*, and *Cedrus*.

Gymnosperms of the Southern Hemisphere

In the Southern Hemisphere, forests dominated by conifers are of much less extent than in the Northern. Mention must be made of the kauri forests of New Zealand (*Agathis*) and the Parana "pine" forests of southern Brazil (*Araucaria*). Many of the Southern Hemisphere conifers grow in mixed forests with various hardwood trees. They may be major elements of such forests, as in certain Australian rain forests (*Araucaria*, *Agathis*, *Podo-*

From *Diversity of Life* by J. Gottlieb; © 1968 by Litton Educational Publishing, Inc. Reprinted by permission of Van Nostrand Reinhold Company

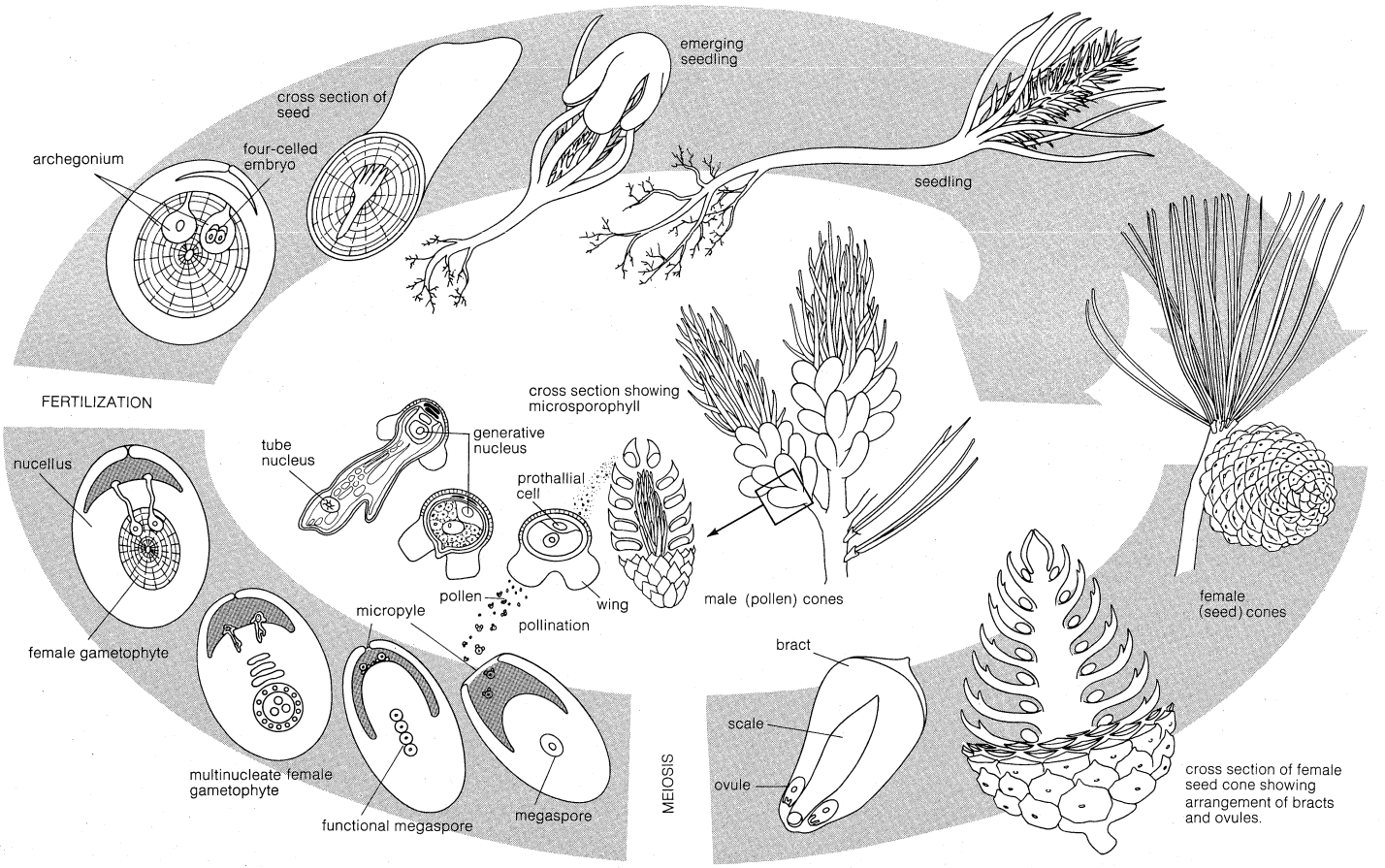


Figure 2: Life cycle of a representative gymnosperm (pine).

carpus) and in the Knysna Forest of South Africa (*Podocarpus*).

Gymnosperms other than conifers that may form a conspicuous feature of the landscape are cycads, as in parts of Australia and southern Africa, and the joint firs, which may be conspicuous in grasslands, as in parts of the southwestern United States.

It is of interest that gymnosperms are almost never dominant in vegetation as a result of superiority in numbers of species present but only in the size and conspicuousness of the plants; for example, in the boreal forest, which is dominated by relatively few species of pines, firs, spruces, and larches, the angiosperms (and frequently the thallophytes as well) are represented by many more species than are the gymnosperms.

Certain kinds of gymnosperms were more widely distributed in geologic times. Ginkgoes, for instance, once occurred in both the Eastern and the Western Hemispheres. The sole surviving species, *Ginkgo biloba*, or maidenhair tree of China, may well have been saved from extinction by being taken into cultivation by man. It is ironic that this survivor from ages past should be one of the most pollution-resistant trees for use in cities. Cycads, once distributed from the Arctic to the Antarctic, are now restricted to a few regions in the tropics and subtropics. As a group they seem to be on the way to extinction. The dawn redwood, *Metasequoia*, once grew in North America and Eurasia. It was abundant in North America from the Upper Cretaceous to the Miocene (from 26,000,000 to about 70,000,000 years ago). Indeed, the genus was first known from fossil remains. Not until the mid-1940s was the dawn redwood discovered to be still in existence, in a remote region of China, the surviving remnant of a once-widespread type.

Generalized life cycle. The life cycle of gymnosperms is typified by the pine (Figure 2). All gymnosperms exhibit an alternation of generations between an asexual phase—the sporophyte—and a sexual phase—the gametophyte. The sporophyte is the tree, with its roots, trunk, branches, leaves, and cones; it produces spores, which develop into gametophytes. The gametophytes—male and female—are microscopic and produce gametes, the egg and sperm. The fertilized egg then develops into a sporophyte.

Cone development

Pollen and seed strobili (cones) are produced on the same tree. In the pollen sacs (microsporangia) of the pollen cones, many cells, called microsporocytes, undergo meiotic cell division, which halves the chromosome number from diploid (24 in *Pinus*) to haploid (12). From each microsporocyte, four haploid microspores are produced. Each microspore ultimately develops into a pollen grain. A pollen grain, at time of shedding, is four-celled: it contains two degenerating prothallial cells, a generative cell, and a tube cell. Pollen grains are wind-borne and may be so abundant as to colour the soil surface yellow or to form a yellow scum on pools and ponds in the vicinity of pine trees. They are carried to the seed cones, where they sift down between the scales and come into contact with the ovules (immature seeds).

Each scale of the seed cones bears on its upper surface near the axis two ovules. Each ovule consists of a central portion, the megasporangium (nucellus), wrapped in a covering, the integument. At the tip of the ovule is an opening, the micropyle, facing the axis of the cone. A single megasporocyte develops within the megasporangium, enlarges, and undergoes meiotic cell division. All but one of the resulting haploid cells disintegrate. The remaining cell, the megaspore, then grows into the female gametophyte within the confines of the megasporangium. At maturity, this gametophyte is a small multicellular (probably 2,000 or more cells) structure that produces usually two to six archegonia at the end of the gametophyte nearest the micropyle. Each archegonium produces one egg.

Pollen grains that have been carried to the ovule adhere to the pollination drop, a drop of fluid exuding from the micropyle. The grains float through this drop or are drawn by its evaporation into a chamber between the integument and the megasporangium. There the grains germinate, each producing a pollen tube that grows through the megasporangium to an archegonium, where it discharges two non-motile sperm. The pollen tube with its sperm constitutes the male gametophyte.

Following fertilization of the egg, the embryo begins to grow. The ovule enlarges to become the seed; the seed coat develops from the integument. At the time of seed dispersal, the embryo consists of an axis bearing several seed leaves, or cotyledons (average of eight in *Pinus*). The lower end of the axis is the radicle, which gives rise to the primary root; the upper end is the plumule, which gives rise to the upper part of the stem and the first leaves of the seedling. The embryo is surrounded by nutritive tissue produced by the female gametophyte. Although usually more than one archegonium—each with its egg—is produced by the gametophyte and usually more than one egg is fertilized, normally only one embryo develops in the seed.

Pollination

The conifers rely on wind for pollination, but at least some cycads, some joint firs, and *Welwitschia* may be pollinated by insects. Agents of dispersal of gymnospermous seeds are several, especially wind and animals. Seeds that are dispersed by wind are mostly provided with wings, as in pines, larches, and kauris. Animal-dispersed seeds, sought by animals (especially birds and mammals) as a food source, include seeds of cycads, gnetums, yews, and podocarps, all of which have a fleshy edible covering. The seeds of bald cypresses are carried by water. Those of junipers are spread by animals, particularly birds, that feed on the berrylike cones.

FORM AND FUNCTION

Diversity in size and structure. Perhaps 60 percent of the gymnosperm species are trees; most of the remainder are shrubs. Some gnetums are vines with twining stems. Many cycads look like palms; others resemble ferns. Joint firs are densely branched shrubs; like many other plants of desert regions, they bear very small leaves. The most bizarre of gymnosperms—and of all plants—is *Welwitschia mirabilis* of southwestern Africa. The top of its turnip-shaped stem barely emerges from the soil, seldom more than 45 centimetres (about 1½ feet); frequently it is almost buried. From the exposed portion of the stem, which may reach 60 centimetres (two feet) in diameter, grow two leathery leaves, one on either side, that persist for the life of the plant, becoming as much as 300 centimetres (about 10 feet) long. As they age, the leaves rest on the ground, eventually splitting lengthwise into ribbons. Mature *Welwitschia* plants have been likened to giant octopuses sunning themselves on desert sands.

Probably the smallest gymnosperm is the Cuban cycad *Zamia pygmaea*, which grows only five to eight centimetres (about two to three inches) tall. The conifer *Dacrydium fonkii* of Chile is a shrub less than 30 centimetres (12 inches) tall. The world's bulkiest gymnosperms—and among the tallest and oldest plants—are the sequoias of California, the trunks of which may be nearly 12½ metres (almost 40 feet) in diameter at ground level. One, said to be the "largest living organism," is the General Sherman Sequoia, which contains more wood than any other single tree. The tallest gymnosperms are the redwoods of California, several of which have exceeded 112 metres (365 feet) high. The oldest gymnosperm—and the oldest trees—are bristlecone pines growing in California's White Mountains.

Most gymnosperms are evergreen plants, retaining their leaves for more than one year. Exceptions are the ginkgo and several conifers, including the larches, dawn redwood, and bald cypresses, which lose their leaves after a single growing season. Most gymnosperms have simple leaves; only the cycads have compound leaves—*i.e.*, leaves divided into leaflets.

Leaf diversity

In form, the leaves are highly varied. Those of cycads resemble the leaves of palms or ferns. The leaves of ginkgo are fan-shaped in outline and, unique among living gymnosperms, are dichotomously veined (*i.e.*, the veins fork equally and repeatedly). The leaves of gnetums, with their veins forming a network, are strikingly similar to those of many angiosperms. The leaves of

conifers show a wide range of form difficult to classify. Three main forms may be recognized: needlelike, scalelike, and broad. Needlelike leaves are long and narrow; they may be round, 3- or 4-angled, or flattened in cross section. Representative needle-leaved conifers are pines, spruces, firs, larches, hemlocks, cedars, and Douglas firs. Scalelike leaves are relatively small, flat, and more or less triangular in outline (at least at the tip); they often overlap like fish scales. Representative scale-leaved conifers are junipers, arborvitaes, sequoia, cypresses, and *Chamaecyparis*. The broad leaves of certain conifers are similar in outline to the leaves of some angiosperms; the flat blades may be up to 30 centimetres (12 inches) long and five centimetres (two inches) wide, with the margins curving out from the leaf base and meeting at the leaf tip. Such leaves are typically thick, even leathery, and may have one vein (the midrib) or several veins running more or less parallel to each other. Broad-leaved conifers include the kauris, araucarias, and podocarps.

The arrangement of the leaves may be spiral ("alternate"), opposite, or whorled (like the spokes of a wheel). In many cases crowding of the leaves or their small size obscures the arrangement. The leaves of some conifers are dimorphic; *i.e.*, of two distinct forms. In pines, for example, the primary leaves are small and papery, hardly leaflike; they constitute the overlapping scales easily seen in the terminal buds. In many pines the bases of these primary leaves remain on the branches behind the needles, or secondary leaves.

The joint firs and *Phyllocladus* are unique among gymnosperms in that their main photosynthetic organs are not leaves but stems. The leaves of the joint firs are rudimentary, whitish to brownish, opposite or whorled scales found at the joints of the green stems. In *Phyllocladus* the leaves are the minute scales found along the edges of the strikingly leaflike, flattened, green branchlets (cladodes).

Living gymnosperms produce pollen and seeds in separate conelike structures, or strobili. These two kinds of strobili may be borne on different plants, as in cycads, yews, ginkgo, joint firs, and *Welwitschia*, or on the same plant, as in most conifers.

The pollen (or "male") cones range in size from those of some junipers, about two millimetres (less than ⅛ inch) long, to those of some cycads, which may reach 80 centimetres (30 inches) in length and 20 centimetres (eight inches) in diameter. Each pollen cone consists of a number of scales (microsporophylls) spirally arranged on a central axis. Each scale bears on its under surface (*i.e.*, the surface toward the base of the cone) two or more microsporangia, or pollen sacs. The hanging pollen strobili of the ginkgo bear little resemblance to cones but look like the catkins of oaks.

Seed cone size

The seed (or "female") cones, especially those of conifers, are familiar objects. The largest cones of any plant, living or extinct, are the seed cones of certain cycads. Those of the Australian *Macrozamia denisonii* may be nearly 100 centimetres (40 inches) long and weigh 38 kilograms (about 84 pounds). The cones of the African *Encephalartos caffer* may weigh slightly more. Among the conifers, the largest cones are the seed cones of the sugar pine (*P. lambertiana*) of far western United States, which reach 50 centimetres (20 inches) in length. The smallest seed cones, those of certain junipers, are only four to five millimetres (3/16 inch) in diameter. The scales of conifer seed cones may be papery or woody and distinct, as in spruces and pines, or they may be fleshy and fused together, the cone being berrylike, as in junipers.

In some gymnosperms (including some conifers) the seeds are borne in fruitlike structures. Those of the ginkgo hang, either singly or in pairs, on long stalks, whereas those of yews, podocarps, *Torreya*, and *Cephalotaxus* occur singly among the leaves. In their commonly fleshy outer covering and hard, stonelike interior, these fruits bear a structural resemblance to plums.

The strobili of many living gymnosperms are simple in that the central axis bears only one kind of structure—sporophylls. Such is the case in the pollen cones of conifers, the pollen strobilus of the ginkgo, and both kinds of cones of cycads. In contrast, the seed cones of conifers such as pines are compound. The central axis bears two kinds of structures, (1) bracts and (2) scales, that morphologically are not sporophylls but are greatly modified lateral shoots. Each scale arises from the axil (upper angle) of a bract and bears one to many seeds on its upper surface (toward the apex of the cone). The pollen strobili and the seed strobili of joint firs, gnetums, and *Welwitschia* are also compound.

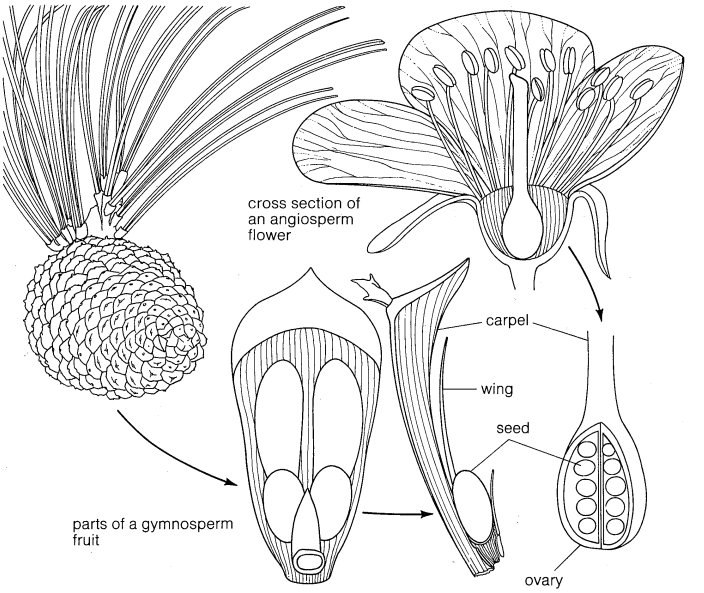


Figure 3: Comparison of gymnosperm cone and angiosperm flower; the seed and the carpel in gymnosperms and angiosperms.

From G. Simpson et al., *Life, An Introduction to Biology;* © 1965 by Harcourt Brace Jovanovich, Inc., reproduced with permission.

Distinguishing characteristics. The major distinguishing feature of the gymnosperms is their so-called naked seeds, which are borne, more or less exposed to the air, on the surface of cone scales or on stalks among the leaves (Figure 3). In contrast, the seeds of angiosperms are borne within fruit tissue that develops from the ovary and sometimes other parts of a flower. Pollen grains are carried to the ovules (immature seeds) and germinate there, extensions called pollen tubes penetrating the ovular tissues. The pollen grains of angiosperms, on the other hand, do not come into contact with the ovules but are carried to the receptive portion (stigma) of a flower, where they germinate; the pollen tubes grow through considerable tissue before they reach the ovary and finally penetrate ovular tissues.

Distinction between gymnosperms and angiosperms

Additional features that distinguish gymnosperms from angiosperms are: (1) the absence of vessels from the wood (except in Gneticae); (2) the lack of specialized cells called companion cells in the phloem; (3) the many-celled or many-nucleated (500 or more) female gametophyte; (4) the presence of archegonia in the female gametophytes (except in *Gnetum* and *Welwitschia*); and (5) the lack of double fertilization (unique to angiosperms). All gymnosperms are woody plants; many angiosperms are not.

Many plant morphologists would restrict the term flower to the reproductive structure of angiosperms, which typically consists of such parts as sepals, petals, stamens, and pistils. On the basis of this restriction, there are no flowers in gymnosperms.

EVOLUTION AND PALEONTOLOGY

The first seed plants (except for two genera of carboniferous lycopsids) were gymnosperms from the Late Devonian Period (about 345,000,000 years ago). It was not until the Lower Cretaceous—some 240,000,000 years later—that the first indisputable angiosperm fossils were laid down. The gymnosperms reached their greatest development during the Mesozoic, when they were the dominant plants in the world. During the Cretaceous,

however, they began to dwindle, and by the end of that period (65,000,000 years ago) their dominant position had been taken over by the angiosperms.

Many morphologists recognize three major groups among the gymnosperms. They are treated as the subdivisions Gneticae, Cycadicae, and Pinicae in this article.

The smallest of the subdivisions is the Gneticae, the gnetophytes, with three genera strikingly different from each other and from other gymnosperms: *Ephedra* (40 species), *Gnetum* (30 species), and *Welwitschia* (one species). The characteristics that unify the Gneticae are (1) the presence of vessels in the wood; (2) the compound strobili, both pollen and seed; (3) the presence of a "perianth," superficially corolla- or calyx-like, associated with the pollen-producing and seed-producing structures; and (4) the presence of two integuments, the inner one produced into a long, tubular micropyle. The Gneticae possess certain features that make them appear remarkably "angiosperm-like." Formerly, the group was even regarded by some morphologists as ancestral to the flowering plants. The Gneticae are isolated and have almost no fossil record. Essentially nothing is known about the origin and evolution of the group.

The remaining two subdivisions of the gymnosperms, the Cycadicae (cycadophytes) and the Pinicae (coniferophytes), are distinguished from each other largely on features of leaves, wood, and seeds. They represent lines that have been in existence since the Devonian (395,000,000 years ago) and that have left a rich fossil record. Whether the lines had a common origin is speculative. Both the Cycadicae and the Pinicae contain orders that have become extinct.

The cycadophytes are characterized generally by often unbranched trunks; by relatively soft and loose wood; by large, mostly compound leaves; and by radially symmetrical seeds. Three of the four orders are extinct. The oldest of these, Cycadofilicales, or seed ferns, ranged from the Devonian to the Jurassic (395,000,000 to 190,000,000 years ago), being most abundant in the Carboniferous. Typical seed ferns, appearing much like present-day tree ferns, had relatively slender stems and large, fernlike leaves that often were much divided. Their seeds were borne either on foliage leaves or on specially modified leaves (megasporophylls). Although seed ferns have many fernlike features, the origin of these plants from true ferns has not been demonstrated. Seed ferns and true ferns may well have arisen from a common ancestral group, one not exactly in either category. It is unsettled whether the seed ferns gave rise to the other orders of Cycadicae.

The second extinct order of the Cycadicae, the Caytoniales, was discovered in 1926. It lasted from the Triassic Period to the Lower Cretaceous Period (136,000,000 years ago). The plants had palmately compound leaves with three to six leaflets. Their ovules were clustered in pouches (cupules) that were paired along an axis. Similarities between the cupules, with their contained seeds, and the angiosperm carpel led to much speculation that the Caytoniales were actually primitive angiosperms, a view now discarded. The order is now regarded as a seed fern offshoot that left no descendants.

Cycadeoids and cycads

The third extinct order of the Cycadicae, the Bennettitales (cycadeoids), ranged from the Triassic to the end of the Lower Cretaceous. The Mesozoic Era was once often called the "Age of Cycads," but it has been demonstrated that much of the cycad-like foliage of that time belonged in reality to cycadeoids, cycad-like plants with either branched or unbranched stems and compound (or, rarely, simple) leaves. They were remarkable for the structure of their strobili, which contained either seed-producing or pollen-producing structures, or both, subtended by numerous leaflike bracts. When both kinds of structures were present, the pollen organs were arranged along the axis, and the ovules were clustered at the end of the axis. The occurrence of both kinds of structures on the same axis originated with the cycadeoids. The ovules were frequently long-stalked; among them were leafy tabs called interseminal scales. All in all the strobilus was superficially quite flowerlike. The hypothesis that these bisexual strobili were ancestral to the bisexual angiosperm flower, however, is now rejected. The origin of cycadeoids is problematical. Relationship to the seed ferns has been suggested but is not yet supported by conclusive evidence. The cycadeoids have no modern descendants.

The one surviving order of Cycadicae is the Cycadales (cycads), with nine genera and about 100 species. Palmlike or fernlike plants with pinnately compound leaves, they typically bear both their pollen and their seeds in cones on different individual plants. *Cycas* is exceptional in bearing the seeds on the margins of leaflike structures (megasporophylls) not grouped into cones. The oldest cycads are from the early Triassic. The present-day species are a remnant of a group that has been in existence for at least 200,000,000 years. The sperm of cycads are motile and are the largest known in vascular plants; as much as 300 microns long, they are barely visible to the unaided eye. The cycads possibly evolved from the seed ferns, but their origin is uncertain.

The final subdivision of the gymnosperms is the Pinicae (coniferophytes), which are characterized by branching stems, relatively dense wood, relatively small simple leaves, and bilaterally symmetrical seeds. One of the four orders, Cordaitales, is extinct. Ranging from the Devonian to the Permian (395,000,000 to 280,000,000 years ago), the cordaitaleans reached their greatest development in the Upper Carboniferous (about 300,000,000 years ago), when they formed great forests of tall, stately trees with a crown of branches near the summit of the trunk. The leaves were strap-shaped and had dichotomous (forking) veins. The pollen-bearing and seed-bearing structures, though separate, were constructed similarly. A slender axis bore two rows of bracts. In the axil of each bract was a budlike organ consisting of scalelike appendages spirally arranged on an axis. Those appendages that were terminal in each bud bore either pollen sacs or ovules. Little is known concerning the origin of the Cordaitales. They and the seed ferns may have arisen from the same stock. The Cordaitales are probably ancestral to the conifers and the ginkgo.

The ginkgo

The Ginkgoales, the second order of Pinicae, were once almost cosmopolitan in distribution. The sole living representative, *Ginkgo biloba*, is now widely planted as a street tree. It can well be called a "living fossil": leaves identical with those of modern ginkgoes have been found in Triassic deposits (200,000,000 years old). The oldest members of this order are from the Lower Permian. The Ginkgoales include branching trees with strap-shaped (fossil members only) or fan-shaped leaves that may be deeply divided. The veins of the leaves are dichotomous. The pollen strobili are catkin-like; the ovules are borne in groups of two to 10 (paired in the modern species) on long, branched, or almost unbranched axes. The sperm are motile. The origin of Ginkgoales is uncertain, but it may lie with some cordaitalean stock.

The third order of Pinicae, the Taxales (taxads), includes five living genera with about 20 species. Although frequently placed with the conifers, the Taxales have been a distinct group as far back as their fossil record goes—about to the Triassic. The best known living taxads are the yews. Members of the order have small, needlelike leaves. The pollen is produced in small cones. The ovules are solitary and terminal on short shoots. Each seed is arillate; *i.e.*, it arises from a juicy, cuplike structure. The origin of Taxales is unknown.

The final order of gymnosperms, the Coniferales (conifers), is the largest and most widespread at the present time. It contains almost 50 genera and about 570 species in six families. The fossil record of conifers extends back well into the Carboniferous. The Petrified Forest of Arizona represents the remains of a coniferous forest of Mesozoic times, with trees that may have resembled present-day araucarias. Conifers are trees or shrubs, with typically small, simple leaves. Both kinds of reproductive structures are borne in cones; the pollen cones are simple, and the seed cones are compound. The early conifers bear a close similarity to the Cordaitales and may have arisen from them.

CLASSIFICATION

Distinguishing taxonomic features. Traditionally the gymnosperms have been classified together in one large group (with rank of division, subdivision, or subclass, depending on the classification scheme used) equal in rank to the angiosperms. Such classification places great weight on one attribute of these plants—the naked seeds. In contrast, some classification schemes divide the gymnosperms into two to four groups, each equal in rank to the angiosperms. Such classification de-emphasizes the importance of the naked seeds but places great weight on the many basic differences among the various major kinds of gymnosperms.

Annotated classification. The classification adopted in this article—after Armen Takhtajan, a Soviet botanist—is no more "correct" than several others. The definitive classification of so large and diverse a group as the gymnosperms will perhaps never be achieved—mainly because each scheme presented depends upon individual and varying interpretations of data.

About 50 of the genera of gymnosperms have fewer than a dozen species; of these, 25 are monotypic (*i.e.*, having but one species). The two largest genera are *Podocarpus*, with about 80 species, and *Pinus*, with about 95. Groups marked with a dagger (†) are extinct and known only from fossils.

DIVISION PINOPHYTA (GYMNOSPERMAE) (gymnosperms)
Woody plants; low shrubs to tall trees. Ovules naked, not enclosed in an ovary; female gametophyte with 500 or more cells (or nuclei) and (except in *Gnetum* and *Welwitschia*) producing definite archegonia; wood without vessels (except in Gneticae); phloem without companion cells; double fertilization lacking.

Subdivision Cycadicae
Palmlike to fernlike plants. Leaves mostly compound. Living species with motile sperms; wood without vessels, relatively soft and loose; microstrobili, when present, simple; ovules with one integument; seeds radially symmetrical.

†Order Cycadofilicales

Extinct fernlike plants from the Devonian to the Jurassic. Leaves mostly relatively large, pinnately compound to decompound. Ovules borne separately along the margins or on the surface of pinnately compound megasporophylls, the latter commonly like the foliage leaves, not in strobili; microsporophylls pinnately compound, not in strobili.

†Order Caytoniales

Extinct plants from the Triassic to the Cretaceous. Leaves of 3–6 palmately arranged leaflets with netted venation; ovules borne in clusters within pouches on pinnately compound megasporophylls, these not in strobili; microsporophylls pinnately compound, not in strobili.

†Order Bennettitales

Extinct cycad-like plants from the Triassic to the Cretaceous. Leaves entire or pinnately compound; stomates of characteristic structure (syndetocheilic); ovules borne singly and terminally on simple megasporophylls, these in strobili, the cluster of megasporophylls sometimes subtended by microsporophylls; microsporophylls pinnately compound.

Order Cycadales

Palmlike or fernlike plants from the Triassic to the present. Dioecious; leaves pinnately or bipinnately compound; stomates of characteristic structure (haplocheilic); ovules borne on usually simple megasporophylls, these in strobili (except in *Cycas*); microsporophylls simple, in strobili. One modern family, Cycadaceae, with 9 genera.

Subdivision Pinicae
Plants with simple leaves, not like those of palms or ferns; living species without motile sperm (except in *Ginkgo*); wood without vessels, relatively dense; microstrobili simple; ovules with 1 integument; seeds bilaterally symmetrical.

Order Ginkgoales

Trees from the Permian to the present. Leaves strap-shaped or fan-shaped in general outline, with dichotomous venation; ovules 2–10, terminal on branched or almost unbranched axes; microstrobili catkin-like; sperm motile. One living species, *Ginkgo biloba*, the ginkgo or maidenhair tree, native to China, widely cultivated; many fossil representatives.

†Order Cordaitales

Extinct trees from the Devonian to the Permian. Leaves strap-shaped; strobili of two kinds, both simple, each kind bearing sterile appendages below and fertile appendages above.

Order Coniferales

Trees or shrubs from the Carboniferous to the present. Leaves needlelike, scalelike, or "broad"; microstrobili simple, conelike; megastrobili compound, often conelike. Six modern families: Pinaceae, with 10 genera; Taxodiaceae, 10; Cupressaceae, 19; Podocarpaceae, 7; Cephalotaxaceae, 1; and Araucariaceae, 2. Cosmopolitan.

Order Taxales

Trees or shrubs from the Triassic to the present. Leaves needlelike; microstrobili conelike; ovules solitary and terminal on short shoots, arillate. One modern family, Taxaceae, with 5 genera, in North America, Eurasia, Indonesia, and the Philippines.

Subdivision Gneticae
Plants with simple leaves, not like those of palms or ferns. Living species without motile sperm; wood with vessels; microstrobili compound; ovules with 2 integuments.

Order Ephedrales

Densely branched shrubs with jointed stems and minute scalelike leaves; strobili conelike. One family, Ephedraceae, with a single genus, *Ephedra*, of North and South America and Eurasia. *Ephedra* pollen from the Eocene, *Ephedra*-like pollen from the Permian; otherwise no fossils.

Order Welwitschiales

Plants with a large, turniplike, mostly subterranean stem and two thick, leathery, elongated leaves that persist for the life of the plant, becoming split lengthwise into ribbons; strobili conelike. One family, Welwitschiaceae, with a single species, *Welwitschia mirabilis*, of Angola and southwest Africa. *Welwitschia*-like pollen from the Permian; otherwise no fossils.

Order Gnetales

Woody vines or, less frequently, shrubs or trees. Leaves opposite, with netted venation, in aspect angiosperm-like; strobili not conelike. One family, Gnetaceae, with a single genus, *Gnetum*, of tropical South America, Africa, southeastern Eurasia, and Indonesia. No fossil record.

Not included in the above outline is the Pentoxylales, an order of Jurassic gymnosperms whose position is uncertain.

Critical appraisal. As an indication of the taxonomic instability of this rather diverse group of plants, taxonomic categories above the family level have long been and continue to be subjected to change. The yews (Taxales) in particular are problematically included by some authorities as a part of the order Coniferales. On the basis of the solitary position of the ovule and of the long, distinct fossil record, however, it seems best to give them separate and equal rank as an order.

Although most families of gymnosperms are well established taxonomically, a few present certain problems which have not yet been resolved. The family Cupressaceae, because of its diverse structure of cones (fleshy and dry) and its variety of foliage (scale and awl leaves), is difficult to sort out at the generic level. The Cupressaceae is frequently included as a subfamily in the pine family (Pinaceae), as is the Taxodiaceae. The family Podocarpaceae is equally diverse and presents a complex collection of plants which are not easily separable into genera. Although close to the conifers, the family Cephalotaxaceae eludes exact placement.

BIBLIOGRAPHY. C.J. CHAMBERLAIN, *Gymnosperms: Structure and Evolution* (1935), a classic work, with an extensive bibliography of early writings; J.M. COULTER and C.J. CHAMBERLAIN, *Morphology of Gymnosperms* (1910), a somewhat dated textbook but still of value; K.R. SPORNE, *The Morphology of Gymnosperms: The Structure and Evolution of Primitive Seed-Plants* (1965), a general book on gymnosperms, with a good summary of evolution and fossil species, and a helpful bibliography; W. DALLIMORE and A. BRUCE JACKSON, *A Handbook of Coniferae and Ginkgoaceae*, 4th ed. rev. by S.G. HARRISON (1966), a well-illustrated descriptive work; THEODORE DELEVORYAS, *Morphology and Evolution of Fossil Plants* (1962), a paleobotany textbook with much description of extinct gymnosperms; RUDOLF FLORIN, "The Systematics of the Gymnosperms," in *A Century of Progress in the Natural Sciences, 1853–1953* (1955), a summary statement by a renowned investigator of gymnosperms; GERD KRUSSMANN, *Handbuch der Nadelgehölze* (1971–72), a profusely illustrated work on the conifers and the taxads.

(J.W.Th.)

Gyroscope

A gyroscope consists of a rapidly spinning wheel set in a framework that permits it to tilt freely in any direction; that is, to rotate about any axis. The momentum of such a wheel causes it to retain its attitude when the framework is tilted; from this characteristic derive a number of extremely valuable applications. A properly mounted gyroscope on board a ship retains its original orientation while the ship rolls, and so can give an accurate measurement of the extent of the roll. Quite similarly the spin axis of a free gyroscope can be orientated to point to true north and will remain pointed in that direction regardless of changes in direction of the vehicle (except for the effects of the rotation of the Earth and the convergence of the meridians when steaming in an easterly or westerly direction; these will be discussed below).

Gyroscopes are used in such instruments as compasses and automatic pilots on board ships and aircraft, in the steering mechanisms of torpedoes, in antiroll equipment on large ships, and in inertial guidance systems.

HISTORY

A 19th-century French scientist, J.-B.-L. Foucault, is responsible for giving the name gyroscope to a wheel, or rotor, mounted in gimbal rings; that is, a set of rings that permit it to turn freely in any direction. Foucault's original experiment, conducted in Paris in 1852, was carried out with a pendulum rather than a rotor. He constructed a pendulum consisting of a very long wire suspending a heavy weight, then drew a line on the floor directly under the pendulum's arc. As spectators watched, the pendulum appeared to move slowly, until its arc was discernibly at variance with the line on the floor. What was actually happening was that the pendulum was remaining faithful to its original arc, with respect to space, but the Earth, and the floor of the exhibition hall, were moving under it due to the Earth's rotation. The phenomenon, startling in 1852, is still a feature of many science museums.

Foucault's experiment

Foucault later carried out a similar experiment using a rotor mounted with three supporting frames (Figure 1, left) and demonstrated that the spinning wheel maintained its original orientation in space regardless of the Earth's rotation. The word gyroscope he derived from the Greek *gyros*, "rotation," and *skopein*, "to view"; a method of "viewing" or demonstrating the rotation of the Earth.

The ability of a gyroscope to maintain its orientation as shown by Foucault suggested its use as a direction indicator, but it was not until 1910 that the first workable gyrocompass was installed, on board a German warship. In the year 1911 Elmer A. Sperry, in the United States, marketed a gyrocompass, and a little later one was produced in Britain. Few fundamental modifications have been made to these early types, and the basic design can still be recognized despite changes in outer appearance. Although it was designed to indicate direction on either land or sea, the gyrocompass has found its greatest application in marine navigation. The rotors of modern gyros are driven by a self-contained electric motor.

In 1909 Sperry built the first automatic pilot using the direction-keeping properties of a gyroscope to keep an aircraft on its course. The first automatic pilot for ships was produced by the Anschütz Company in Kiel, Germany, and installed in a Danish passenger ship in 1916.

A three-frame gyroscope (Figure 1, left) was used, also in 1916, in the design of the first artificial horizon for aircraft. This instrument indicates roll (side to side) and pitch (fore and aft) attitude to the pilot and is especially useful in the absence of a visible horizon.

One of the earliest uses of a gyroscope in a missile was to steer a torpedo on a straight course. The spin axis was aligned parallel to the longitudinal axis of the torpedo. Any angular error in the course was detected by the gyroscope and used to initiate rudder deflection to correct the error.

Early use in torpedoes

Conventional three-frame gyroscopes are used in ballistic missiles for automatic steering together with two-frame gyroscopes (Figure 1, right) to correct turn and pitch motion. Beginning in the 1930s German engineers were pioneers in this field and their knowledge was used in the design of guidance systems for the V-1 flying bomb, a pilotless aircraft, and the V-2 rocket, an early ballistic missile.

In 1915 the Sperry Company, employing a two-frame gyroscope, devised a gyrostabilizer to reduce the rolling of ships, thus minimizing damage to cargo, reducing stresses in the hull structure, and adding to the comfort of passengers. It was installed in about 40 ships, the most notable installation being a massive three–rotor gyrostabilizer in the 41,000-ton Italian liner "Conte di Savoia" in 1933. The roll-reducing action of this type of gyrostabilizer was quite effective and was independent of the speed of the ship. It had a number of disadvantages including its weight (over 600 tons), cost, and space requirements (each rotor was 13 feet in diameter and rotated 900 revolutions per minute), and it was not installed on later ships, in part because of the introduction by Japanese shipyards of an underwater fin-type ship stabilizer in 1925. At first manually controlled, the fins were later given automatic controls, using small gyroscopes.

The ability of a gyroscope to define a direction in space with a great degree of accuracy, used in conjunction with sophisticated control mechanisms, led to the development during World War II of stabilized gunsights, bombsights, and platforms to carry guns and radar antennas aboard ships.

Missiles and vehicles that depend on accurate navigation or on guidance or both are equipped with inertial navigation systems. Such a system requires a small platform, stabilized by gyroscopes to an extraordinary degree of precision. It was not until the 1950s that this type of platform was perfected, following work in the design of air-supported bearings and flotation gyroscopes (see below *Inertial navigation systems*).

From W. Burger and A. Corbet, *Ship Stabilizers, Their Design and Operation in Correcting the Rolling of Ships* (1966); Pergamon Press Limited

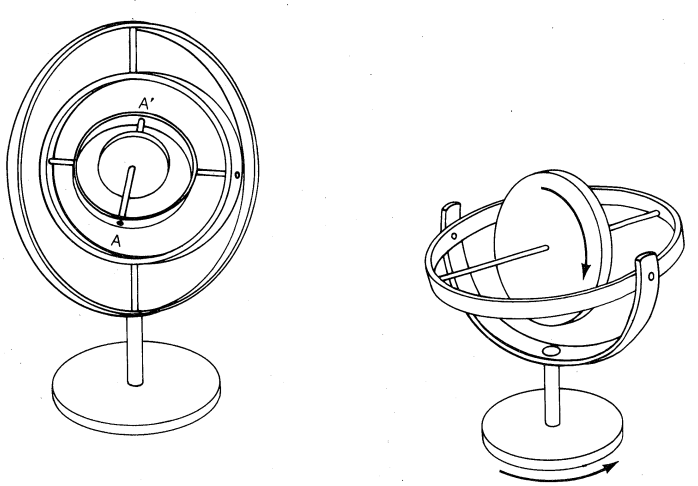


Figure 1: (Left) Three-frame gyroscope and (right) two-frame gyroscope.

BASIC OPERATING PRINCIPLES

Gyroscopic inertia and precession. If the base of a three-frame gyroscope model (Figure 1, left) is held in the hand with the rotor spinning and turned about any of the three axes, the rotor axle will continue to point in the original direction in space. This elemental property is known as gyroscopic inertia or rigidity in space. If the speed of the wheel decreases, the gyroscopic inertia gradually disappears; the rotor axle begins to wobble and ultimately takes up any convenient position. Rotors with a high speed and a concentration of mass toward the rim of the wheel display the strongest gyroscopic inertia. It is apparent, therefore, that gyroscopic inertia depends on the angular velocity and the moment of inertia of the rotor, or on its angular momentum. The rotor wheel is subject to the laws of rotational motion and inertia in that a freely rotating body will maintain a fixed direction in space, and the rotor tends to preserve its angular mo-

mentum, or spinning action, unless acted on by some external force.

The consequence of gyroscopic inertia is that to the observer on Earth the spin axis of a gyroscope makes an apparent movement over a period of time, although this apparent motion merely reflects the revolution of the Earth about its axis. There is one exception to this, that when the spin axis points toward the polar star, there is no movement of the spin axis with respect to the observer's surroundings, as the axis is parallel to the Earth's axis and points toward the celestial poles. This apparent movement is shown in Figure 2, in which, at position 1, the spin axis is parallel to the horizontal plane and the end of the spin axis, marked A, points due north. As the direction of the Earth's rotation is counterclockwise when seen from above the North Pole, the relative direction of the A-end will change through northeast, east, southeast, south (position 5), etc., and this clockwise movement will continue until, at the end of one period of rotation of the Earth (23 hours 56 minutes) the rotor and spin axis revert to their original position with respect to the observer on the Earth's surface. While this is taking place, the A-end is apparently tilting upward between positions 1 and 5 and tilting downward between positions 5 and 1. The change in azimuth (direction) of the spin axis is often referred to as "drifting"; sometimes "tilting" and "drifting" are collectively called "apparent wander."

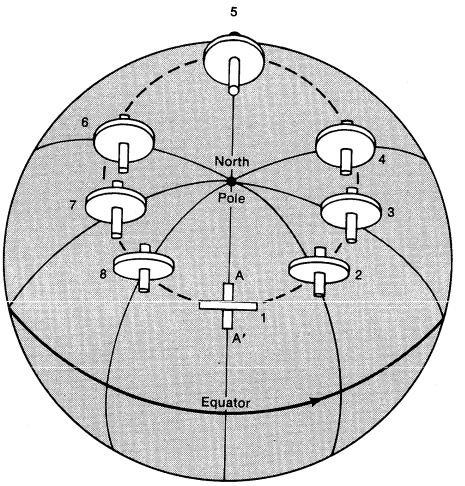


Figure 2: Apparent movement of the spin axis of a gyroscope (see text).
From W. Burger and A. Corbet, *Ship Stabilizers, Their Design and Operation in Correcting the Rolling of Ships* (1966); Pergamon Press Limited

Figure 2 is drawn for a gyroscope location in a latitude north of the Equator. A similar figure drawn for a latitude south of the Equator would reveal a counterclockwise drifting. In fact, the spin axis, whatever the place, would follow the movement across the heavens of a reference star at which it is pointing.

Precession. If, while the rotor of a model of a three-frame gyroscope (Figure 1) is spinning, a slight vertical downward or upward pressure is applied to the horizontal gimbal ring at A or A′, the rotor axle will move at right angles in a horizontal plane. But no movement will take place in the vertical plane; in fact, a definite resistance to motion is felt in this direction. Similarly if a sideways pressure is applied at the same point, the rotor axle will tilt upward or downward. This second property is called "precession." A precession or angular velocity in the horizontal plane is caused by the application of a couple—*i.e.*, parallel forces equal and opposite, in the vertical plane perpendicular to that of the rotor wheel. Precession is the tendency of the rotor's axis to move at right angles to any perpendicular force applied to it.

Determining the direction of precession

A convenient way to remember the direction in which precession takes place is to regard the force as though it acted at a single point on the rim of the wheel. This point will not move in response to the force; instead, a point 90° beyond in the direction of the wheel's rotation will move away.

To explain this strange anomaly—that cause and effect seem to operate in planes at right angles to one another—consider a particle m at the upper and lower rim of the rotor; *i.e.*, at those places where the vertical couple exerted on the axle has maximum effect (Figure 3). It will

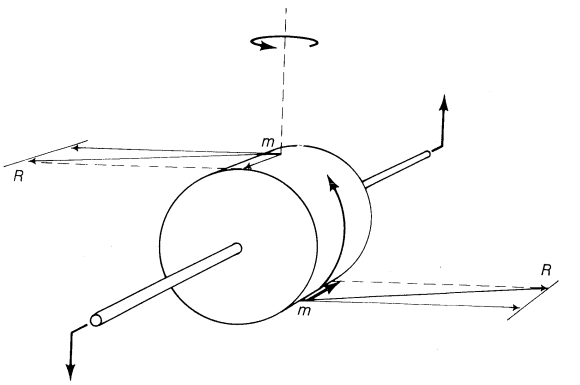


Figure 3: Precessional effect (see text).

then be apparent that the particle m is subject to two velocities, one relatively very large owing to the speed of the rotor, the other small as a result of the applied couple. The directions of the velocities are mutually at right angles, and the particle m will try to move in the direction of the resultant R of the two velocities; in other words, the rotor will acquire an angular velocity in the horizontal plane.

Sensing angular displacement. *Three-frame gyroscopes.* The unrestrained or "free" three-frame gyroscope has little practical use, because its spin axis is subject to tilting and drifting owing to the rotation of the Earth. In the controlled state, however, it is widely used. The term control of a gyroscope implies that the spin axis, by small continuous or intermittent applications of torque, is made to precess so that it points at—or, rather, oscillates around—a reference mark fixed in relation to coordinates on the Earth rather than in relation to space.

Controlled gyroscopes fall into three categories: north-seeking, directional, and gyrovertical.

The north-seeking gyro is used for marine gyrocompasses. In the settling (or normal) position the spin axis is kept horizontal and in the plane of a meridian.

The directional gyroscope is used in aircraft and is sometimes called a self-levelling free gyroscope corrected for drift. With its spin axis horizontal it has directional properties but does not automatically seek the meridian plane.

The gyrovertical has its spin axis vertical and is used to detect and measure angles of roll and pitch.

Those three types of three-frame gyroscopes are known as displacement gyroscopes because they can measure angular displacements between the framework in which they are mounted and a fixed reference direction—the rotor axis.

Two-frame gyroscopes. The following simulated experiments conducted on the two-frame gyro (Figure 1, right) illustrate the basis of important applications.

If, with the rotor spinning and the spin axis in the horizontal plane, the base is rotated uniformly in the horizontal plane, a definite resistance owing to gyroscopic inertia will be felt. At the same time, the spin axis will begin to precess in the vertical plane and will continue to do so until the axis is vertical and all gyroscopic inertia disappears.

If the experiment is then repeated as before, except that while the base is being turned in the horizontal plane, the precessional movement of the spin axis is stopped by the application of force to the end of the axle where it terminates in the gimbal ring, the resistance to the turning motion of the hand due to gyroscopic inertia will cease to exist. In effect, the precession process will have been reversed. A vertical downward force applied to the end of the rotor axle, then, introduces a torque (twisting force) that makes the base precess at the same rate and in

the same direction as the turning movement of the hand. The quicker the base is turned, the greater the force that must be exerted on the axle to stop the precession. Two important conclusions may be drawn from this experiment: (1) There is resistance to the turning motion of the base if, and only if, the spin axis precesses. (2) The force needed to stop the precession is directly proportional to the rate of turning of the base. This force can be exerted by a spring arrangement in which the gyroscope (Figure 4A) measures the rate of change of azimuth and is used in aircraft and ships as a "rate-of-turn indicator."

From W. Burger and A. Corbet, *Ship Stabilizers, Their Design and Operation in Correcting the Rolling of Ships* (1966); Pergamon Press Limited

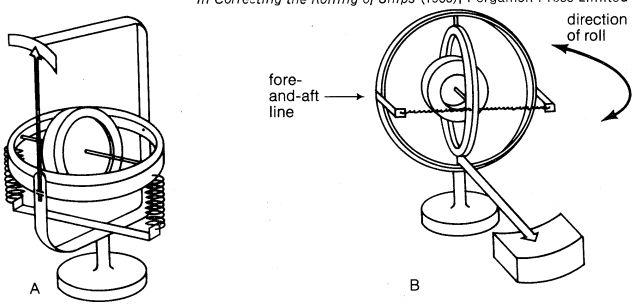


Figure 4: *Rate gyroscopes.*
(A) Velocity gyro for measuring rate of turn. (B) Velocity gyro for measuring rate of roll.

Rate-of-turn indicator

The greater the rate of turn of the vehicle—and base—the larger will be the force required to stop the precession in the vertical plane. Additional force results in more spring extension and greater angular displacement between the gimbal ring and the horizontal plane. The deflection of the instrument's pointer, when it is steady, therefore, gives an indication directly proportional to the rate of change of azimuth.

The gyroscope illustrated in Figure 4B is used on ships to measure the rate of roll; *i.e.*, the angular roll velocity. The spin axis is positioned at right angles to the fore-and-aft line and the rate of roll is measured about this line.

These gyroscopes are called velocity or rate gyroscopes as distinct from "displacement" gyroscopes. The sensitive or input axis of a rate gyroscope is at right angles to its spin axis, while with a displacement gyroscope the spin axis is directly equivalent to the sensitive or input axis. A north rate gyroscope combined with a north displacement gyroscope, therefore, have their spin axes at right angles to each other.

GYROSCOPIC DEVICES

Stabilizer for ships. The main components of a ship stabilizer are a set of fins and the gyroscopes. The fins protrude from the ship's hull and are so operated that the forward motion of the ship produces a tilt in one direction on one fin and in the opposite direction on the other fin. When properly controlled, therefore, these fins oppose the rolling motion. The gyroscopes sense the vertical angular displacement and the roll velocity and provide the proper control for the fins.

A block diagram of a stabilizer is shown in Figure 5. The installation is designed in accordance with the principle that the movement of the fins is a total function of the roll angle, the roll velocity, the roll acceleration, the amount of helm (rudder), and the natural list of the vessel. The "natural-list" unit operates if the vessel develops a permanent list. It generates a signal that makes it possible for the stabilizers to act about a mean position corresponding to the ship's permanent list, thus avoiding power waste in attempting to correct the list. Provision is made, however, to switch out this signal if desired; when this is done the stabilizer will operate to correct the list.

The roll-angle sensor is a vertical displacement gyroscope; the roll-velocity sensor, a rate gyroscope; and the accelerometer is an electrical device that measures the roll acceleration by determining the rate of change of movement of the roll velocity gyroscope. All the signals are combined and amplified to actuate a hydraulic relay unit, the output arm of which controls the action of a hydraulic pump that in turn controls the movement of the hydraulic fin-tilting mechanism.

Inertial navigation systems. Neither position nor velocity can be sensed directly by an inertial system. An acceleration (change of velocity), however, can be detected by an accelerometer and this can be used to determine the position of a ship, aircraft, or space vehicle. Basically this navigational system comprises three components: the platform, the gyroscopic frame, and the computer.

The platform. The accelerometers, mounted with their input axes mutually at right angles, are carried on a platform. Two accelerometers measure acceleration in the horizontal plane—the requirement for surface navigation. For space navigation an additional accelerometer measures acceleration in the vertical plane. Each of the acceleration signals can be converted into distance travelled by determining, first, the total change in velocity which, added to the known initial velocity, gives the vehicle velocity; and, second, the total change in position that, added to the known initial position, yields the present vehicle position. The platform on which the accelerometers are grouped must be stabilized; generally the stabilization is about the local vertical with the input axis of one accelerometer pointing north, one axis pointing east, and the third, when required, pointing vertically.

The gyroscopic frame. This component is responsible for the stabilization of the platform. Three rate gyroscopes are fitted in the frame with their three input axes mutually perpendicular. Two of the gyroscopes provide the horizontal alignment of the platform—an essential requirement to eliminate the influence of accelerations due to gravity—while the third is responsible for the north–south alignment.

The three components of the system

Pitch, roll, and yaw are detected by the three gyroscope input axes. The gimbal deflection of each of the gyroscopes is converted into a signal voltage that, when amplified, drives a servomotor which, via a gear train, rotates the frame back to its original position. Stabilization of the platform is achieved by coupling it to the gyroscopic frame.

Tilting and drifting due to the Earth's rotational movement (Figure 2) are also detected by the gyroscopic frame. If the platform is to be kept horizontal and north-south stabilized, torque signals must be applied to the roll and pitch servomotors to offset the precession caused by tilting, as well as to the azimuth servomotor to eliminate the precession caused by drifting.

The rate gyroscopes are not spring-restrained. Instead, "flotation gyroscopes" in which the precession is opposed by the viscous drag of a liquid are employed. The opposing torque is therefore proportional to the precession rate, instead of the precession displacement, as in a spring-restrained gyroscope.

The computer. This unit performs the necessary calculations. Specifically, it applies certain corrections to the

From W. Burger and A. Corbet, *Ship Stabilizers, Their Design and Operation in Correcting the Rolling of Ships* (1966); Pergamon Press Limited

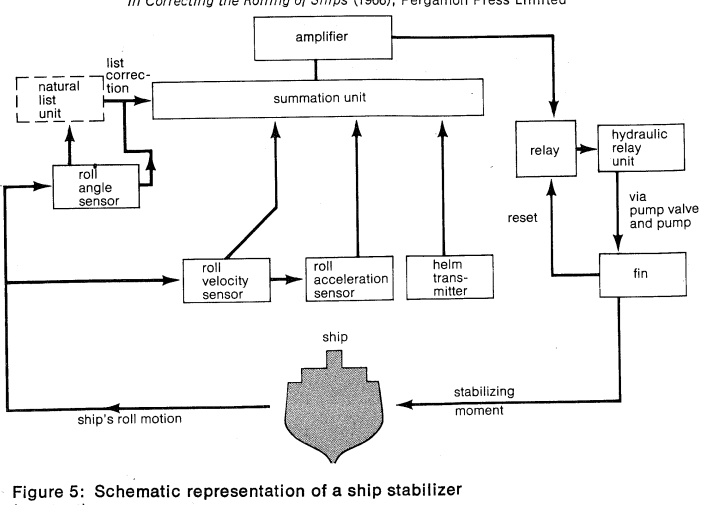


Figure 5: Schematic representation of a ship stabilizer (see text).

acceleration; integrates acceleration to velocity and velocity to distance; computes latitude and longitude; and converts geocentric latitudes into geographic latitudes. If the inertial system is used for inertial guidance in space navigation, then the computer also compares the vehicle's position with the destination or target position to provide steering commands and compares the vehicle's velocity vector (both direction and magnitude) with the programmed velocity vector to provide rocket steering and engine cutoff commands.

Stabilized platforms and gunsights. The inertial type of platform is extremely small and must be stabilized to an extraordinary degree of precision, but the method of stabilization used for gun platforms is essentially the same. The gyroscopes that detect platform displacement are not as accurate and expensive as the flotation type. The servomechanisms must be capable of exerting considerable torque and are often electrohydraulic in character.

The gyroscopic gunsight revolutionized aerial gunnery. The sight fitted on the gun contains a rate gyroscope capable of measuring angular velocities in two planes at right angles to each other. The gyroscope sight can be thought of as the three-frame gyroscope shown in Figure 1, left, constrained by horizontal and vertical springs to the inner and outer gimbal, respectively. Instead of a mechanical spring arrangement, variable-strength magnetic fields are used to constrain the rotor axle in azimuth and elevation. The field coils for producing the horizontal component of this magnetic field are coupled to the range finder. The current through the vertical field coils is adjusted so that the field depends on the drop of the projectiles due to gravity. The sensitivity of the gyroscope in the horizontal plane is a function of the sighting range; in the vertical it is a function of the gravity drop.

Use of magnetic fields

In operating this gunsight, often called a "predictor" sight, the gunner holds the image of a central dot over the target while the gun is automatically aimed by the gyroscope at the place where the target will be at the expiration of the time of flight of its projectile.

Aircraft instruments. The three primary gyroscopic instruments fitted to the flight panel are a rate-of-turn indicator, a directional gyroscope, and an artificial horizon. The rate-of-turn indicator was discussed above. Such gyroscopes may be driven by electric motors or by air jets.

Directional gyroscope. This instrument forms a standard reference for the pilot and navigator. It is a three-frame gyroscope with its spin axis in the horizontal plane. As soon as tilt develops, a switch is closed between the gyroscope housing and the vertical gimbal ring and a motor introduces a torque in the horizontal plane that causes the gyroscope to precess back toward the horizontal. The reduction of azimuth drift to its lowest practical limit is achieved by an adjustable "latitude" balance nut attached to one side of the inner gimbal ring. Random precessional errors (real wander) are always present because of wear of bearings, variation of temperature resulting in contractions and expansions, and mechanical imperfections. Owing to the convergence of the meridians at the poles, an apparent wander (drift) takes place when an aircraft changes its longitude, except when it is flying along the Equator.

Artificial horizon. The artificial horizon displays the rolling and pitching motion of the aircraft. It consists basically of a three-frame gyroscope with its spin axis vertical and automatic correcting devices to counteract the apparent motion of the spin axis around the celestial pole (the North Pole in celestial coordinates) and any other random precessions.

Other applications. The gyroscope principle has been utilized in many other applications, such as the gyrocompass and gyropilot, and in nonrotating gyroscope devices.

Gyrocompass. A compensated magnetic compass, free from external accelerations, indicates magnetic north, which varies from true north from place to place on the Earth's surface. A gyrocompass, however, when properly adjusted, can be made to indicate true north.

The marine gyrocompass is a three-frame gyroscope with its spin axis horizontal. To achieve the north-seeking and actual location (or meridian-settling) properties of a gyroscope, use is made of the tilting effect of the spin axis when it is not pointing true north. As soon as tilt develops, a pendulum type device introduces torques that precess the spin axis toward the meridian, causing it to describe a spiral with an ever-decreasing radius. When settled (stabilized) the spin axis is maintained in the meridian plane by a precession equal but opposite to the drift at the particular latitude.

Marine gyrocompasses

When there is no tilting effect the marine gyrocompass will lose its directional properties and become useless. This is the case at the poles and also when a vehicle moves due west with a speed equal to the surface speed of the Earth. Because the latter condition can easily exist in an aircraft in the middle and upper latitudes, it cannot be used for air navigation.

Aircraft gyrocompass. Aircraft gyrocompasses are based on automatically monitored directional gyroscopes in which the monitoring device senses the direction of the meridian and ensures that the gyroscope axis is maintained in this direction. The monitoring device consists of a magnetic sensing unit, called the flux valve, and allowance is made for variation in the direction of the Earth's magnetic field.

Gyropilot. The gyropilot—or automatic pilot, as it is usually called—consists basically of three devices, each of which detects disturbances of the aircraft in one plane and corrects for these disturbances by moving the appropriate control: the rudder control for azimuth and sudden change in heading (yaw) disturbances, aileron control for roll disturbance, and elevator control for pitch disturbance.

In the past some gyropilot designs have relied entirely on displacement sensors, but with this system accurate control is not possible. In the later designs rate detection forms the principal reference and displacement detection plays a secondary role. In such designs yaw disturbance is detected by a rate gyroscope and the change in heading is detected, to a lesser degree, by the associated gyrocompass. The two signals are added electronically and cause corrective rudder control to be applied to the rudder servomotor. The roll disturbance is detected by a roll rate gyroscope and to a lesser degree by a roll angle pendulum, which senses displacement. The aileron servo applies corrective action. Pitch disturbance is detected by a pitch rate gyroscope and to a lesser degree by a pitch pendulum. The elevator servo applies corrective action.

Ships' gyropilots primarily use displacement signals from the gyrocompass. Rate signals can be taken from a small generator geared to the displacement shaft of the gyrocompass.

Miscellaneous applications. Extensive use is made of two-frame gyroscopes to measure the rate of turn of a vehicle. Rate gyroscopes are also mounted on theodolites used for orientation of field artillery and other weapons and for surveying. Theodolites, mounted on gyrocompasses, are used in underground mine exploration, since magnetic compasses would be disturbed by metal deposits.

Use with theodolites and field artillery

Two-frame gyroscopes have been used successfully in the stabilization of ships, monorail cars, and two-wheel cars.

Vertical three-frame gyroscopes with pen recorder attachments often are used to analyze rolling and pitching movements of ships and rocking motions of trains.

Gyroscopes without rotors. The word gyroscope means "viewing the rotation," which implies that a rotating wheel is not a necessary part of a gyroscope. All instruments that are able to measure angular velocity relative to axes fixed in inertial space can be called gyroscopes. One example of such a gyroscope is the ring laser. In this instrument a laser beam (see LASER AND MASER) is split by means of mirrors into two beams that are sent in opposite directions around a closed circuit in the horizontal plane. The velocities of the two beams measured relative to the Earth will differ, since one beam travels in the direction of the Earth's rotation (drifting), while the other beam travels against the direction of this component. This leads to a displacement of the two signals with

respect to each other (phase shift), which can be measured when the two beams are recombined. The phase shift is proportional to the rate of turn. Thus, the ring laser measures the drifting due to the Earth's rotational motion and can therefore be classed as a rate gyroscope.

EFFECTS OF PRECESSION

The Earth as a gyroscope. The Earth itself belongs to the class of three-frame gyroscopes and owing to gyroscopic inertia it has maintained its axis in a (nearly) fixed direction. This characteristic, together with the facts that the Earth moves around the Sun in one year and that its axis is inclined to the plane of the ecliptic (the orbit around the Sun) at an angle of about 66½°, explains the changes of season.

The Earth can be thought of as a sphere that bulges toward the Equator. As a result of this equatorial bulge—coupled with the fact that the Earth's axis is not perpendicular to the plane of the ecliptic and the plane of the Moon's orbit—the forces exerted on the Earth by the Moon (and to a lesser degree by the Sun) in general do not pass through the centre of gravity of the Earth. This causes the Earth's axis to precess along a small ellipse with a period of 18⅔ years while at the same time the centre of the ellipse precesses along a circle with spherical radius of 23½° with a period of 26,000 years.

The former motion is called "nutation." The latter motion causes a precession of the equinoxes (the points on the ecliptic where day and night are of equal length), with a mean displacement of 50.26 angular seconds annually.

Satellites. Artificial satellites can be considered part of a spinning wheel and, provided they are not polar satellites (*i.e.*, those that have their orbital planes at right angles to the Equator), the equatorial bulge exerts a pull on the satellite. Since the direction of the pull falls outside the orbital plane, this results in a precessional movement of the satellite's orbit. This precession can be quite considerable; for example, a satellite at a height of 800 kilometres (500 miles) with an orbital inclination of 40° with the Equator and travelling eastward, precesses five degrees toward the west each day.

Moving vehicles. Precessional effects also are experienced in fast-moving motor cars and locomotives. Road and rail engineers must take these factors into account when working out banking angles for curves. Similar problems occur when large rotors, such as steam or gas-turbine rotors, are carried in moving vehicles. The pitching movement of such vehicles can generate precessional effects, resulting in a tendency to yaw.

Projectiles. Gun barrels are rifled to give the projectile a high rate of spin about its fore-and-aft axis. This imparts gyroscopic inertia to the projectile, preventing it from toppling over. Since wind resistance can produce precession, ballistic calculations must take this into account.

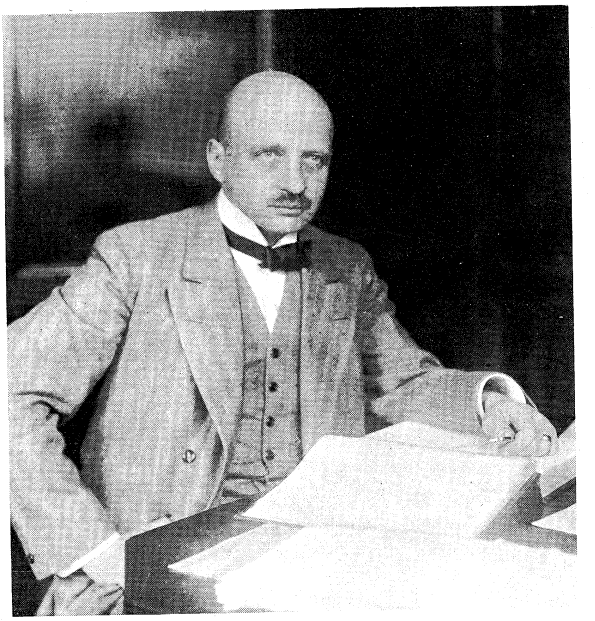
BIBLIOGRAPHY. W. BURGER and A.G. CORBET, *Marine Gyro-Compasses and Automatic Pilots*, 2 vol. (1963–64), of special interest to the mariner, but the first four chapters of vol. 1 give a comprehensive account of the basic principles of gyroscopes; *Ship Stabilizers* (1966), ch. 2 offers an interesting and simple approach to the principles of gyroscopes, ch. 3–4 deal with anti-roll devices in general and are easily understandable for the lay reader, while the remainder of the book throws light on the technical and engineering principles embodied in ship stabilizers. *Gyros* in the *Institution of Mechanical Engineers Proceedings*, vol. 179, pt. 3E (1964–65), includes a full account of the history and development of all types of gyroscopes and all instruments related to gyroscopes; most of it can be understood by the reader who does not have a technical background; R.H. PARVIN, *Inertial Navigation* (1962), a well-written book but requiring a fundamental knowledge of mathematics; K.I.T. RICHARDSON, *The Gyroscope Applied* (1954), one of the best books ever written about gyroscopes for the layman.

(W.B./A.G.C.)

Haber, Fritz

The German physical chemist Fritz Haber was distinguished not only for his researches, which won him the Nobel Prize, but also for his services to industry and to his country. Haber and the research institute he directed contributed to a wide range of advances in physical chemistry. His most outstanding scientific achievement was his synthesis of ammonia virtually from air and water, which solved the urgent problem of meeting the world demand for nitrogen fertilizer.

EB Inc.



Haber.

Early life

He was born on December 9, 1868, in Breslau, Silesia (now Wrocław, Poland), the son of a prosperous chemical merchant. After the usual classical education of the *Gymnasium*, and student years in Berlin, Heidelberg, and Zürich, he entered his father's business, but his impatient spirit soon led to a break.

Deciding on an academic career, he first took up organic chemical research at the University of Jena, but its orthodox methods gave him little satisfaction. In later years he delighted in telling how chance brought him at the age of 25 to a junior post at the Technische Hochschule of Karlsruhe, where he immediately threw himself with tremendous zest into the teaching of physical chemistry (a subject in which he was essentially self-taught) and into research. His intensive early researches in electrochemistry and thermodynamics soon gained him the position of professor of physical chemistry (1898); his reputation was much enhanced by his timely book *The Theoretical Basis of Technical Electrochemistry* (1898) and especially by *The Thermodynamics of Technical Gas Reactions* (1905), a pioneering work that had considerable influence on teaching and research.

In the first decade of the 20th century the rapidly increasing demand for nitrogen fertilizer greatly exceeded the supply, which still came mainly from Chilean nitrate. The problem of utilizing atmospheric nitrogen for this purpose had become of worldwide concern. Haber developed a method for synthesizing ammonia from nitrogen and hydrogen. By 1909 he had established conditions for the large-scale synthesis of ammonia. The process was handed over to Carl Bosch of the Badische Anilin und Soda-Fabrik AG for industrial development, leading to the Haber–Bosch ammonia process. Haber was awarded the Nobel Prize for Chemistry in 1918.

Director of the Kaiser Wilhelm Institute

In 1911, at the age of 42, he was appointed director of the Kaiser Wilhelm Institute for Physical Chemistry in Berlin, a new research establishment that was to become even more famous than the school he had built up in Karlsruhe. With the outbreak of World War I in 1914, he immediately placed himself and his laboratory at the service of the government, his first concern being to organize the supply of essential war matériel. After the development of trench warfare he was made head of the chemical-warfare service, and his institute became a

major military establishment. He played a leading part in the development of poison gas as a weapon.

The war years were for Haber a period of intense effort motivated by his strong patriotism. He felt the outcome as a personal tragedy. When Germany was required to pay enormous reparations, Haber sought to find a way of extracting gold from seawater. Long efforts ended in 1926 with the conclusion that the gold content of seawater was far less than had been thought. The failure was a bitter disappointment to him, though it had positive results of much scientific value.

After the war Haber's institute became the world's leading centre of research in physical chemistry, with a large and distinguished international staff. All his life he had been an advocate of close relations between science and industry, and he now became a recognized authority in this field; he was active in promoting the national organization of research and in fostering friendly relations with foreign scientists. He was much attracted to Japan and in 1930 established the Japan Institute, with headquarters in Berlin and Tokyo, to promote mutual understanding and cultural interests. He enjoyed the high title of privy councillor (*Geheimer Regierungsrat*) and was an honorary fellow of leading chemical societies.

The breakup of Haber's institute began in 1933, when, with the rise of the Hitler regime and its anti-Semitic policy, this great German chemist became the "Jew Haber." He resigned in 1933 and accepted an invitation to work at Cambridge. After four months he left to spend the winter in Italy, but he suffered a heart attack on the way in Basel and died there on January 29, 1934.

Haber's first wife was Clara Immerwahr, with whom he had one son. In later life he married Charlotta Nathan, with whom he had a son and a daughter.

A striking feature of Haber's life and work was his versatility. He and the institutions he directed contributed in a fundamental way to nearly all the important branches of physical chemistry. In fact, his scientific life reflected the main developments of physical chemistry over a period of 40 years. Furthermore, Haber was strongly imbued with the conviction that the basic purpose of science was the betterment of mankind.

BIBLIOGRAPHY. A biography of Haber, as well as a survey of his scientific work, is J.E. COATES, "Haber Memorial Lecture," *J. Chem. Soc.*, pp. 1642–72 (1939). M. GORAN, *The Story of Fritz Haber* (1967), contains an exhaustive bibliography of writings by and relating to Haber. Of the many works by German authors about Haber, probably the best is ch. 10 of RICHARD WILLSTATTER, *Aus meinem Leben* (1949; Eng. trans., *From My Life*, 1965). Of Haber's own writings, apart from his scientific books and articles, two relate to his life and ideas: *Fünf Vorträge aus den Jahren 1920–1923* (1924); and *Aus Leben und Beruf; Aufsätze, Reden, Vorträge* (1927).

(J.E.Co.)

Habsburg, House of

Also known as the House of Austria, the Habsburgs were one of the greatest of the sovereign dynasties of Europe.

ORIGINS

The name Habsburg is derived from the castle of Habsburg, or Habichtsburg ("Hawk's Castle"), built in 1020 by Werner, bishop of Strasbourg, and his brother-in-law, Count Radbot, in the Aargau overlooking the Aar River, in what is now Switzerland. Radbot's grandfather, Guntram the Rich, the earliest traceable ancestor of the house, may perhaps be identified with a Count Guntram who rebelled against the German king Otto I in 950. Radbot's son Werner I (died 1096) bore the title count of Habsburg and was the grandfather of Albert III (died *c.* 1200), who was count of Zürich and landgrave of Upper Alsace. Rudolf II of Habsburg (died 1232) acquired Laufenburg and the "Waldstätte" (Schwyz, Uri, Unterwalden, and Lucerne), but on his death his sons Albert IV and Rudolf III partitioned the inheritance. Rudolf III's descendants, however, forming the House of Habsburg–Laufenburg, sold Laufenburg and other districts to Albert IV's descendants before dying out in 1408.

AUSTRIA AND THE RISE OF THE HABSBURGS IN GERMANY

Albert IV's son Rudolf IV of Habsburg was elected German king as Rudolf I in 1273. It was he who, in 1282, bestowed Austria and Styria on his two sons Albert (the future German king Albert I) and Rudolf (reckoned as Rudolf II of Austria). From this date the age-long identification of the Habsburgs with Austria begins (see AUSTRIA, HISTORY OF). The family's custom, however, was to vest the government of its hereditary domains not in individuals but in all male members of the family in common, and though Rudolf II renounced his share in 1283, difficulties arose again when King Albert I died (1308). After a system of condominium had been tried, Rudolf IV of Austria in 1364 made a compact with his younger brothers that acknowledged the principle of equal rights but secured de facto supremacy for the head of the house. Even so, after his death the brothers Albert III and Leopold III of Austria agreed on a partition (Treaty of Neuberg, 1379): Albert took Austria, Leopold took Styria, Carinthia, and Tirol.

King Albert I's son Rudolf III of Austria had been king of Bohemia from 1306 to 1307, and his brother Frederick I had been German king as Frederick III (in rivalry or conjointly with Louis IV the Bavarian) from 1314 to 1330. Albert V of Austria was in 1438 elected king of Hungary, German king (as Albert II), and king of Bohemia; his only surviving son, Ladislas Posthumus, was also king of Hungary from 1440 and of Bohemia from 1453. With Ladislas the male descendants of Albert III of Austria died out in 1457. Meanwhile the Styrian line descended from Leopold III had been subdivided into Inner Austrian and Tirolean branches.

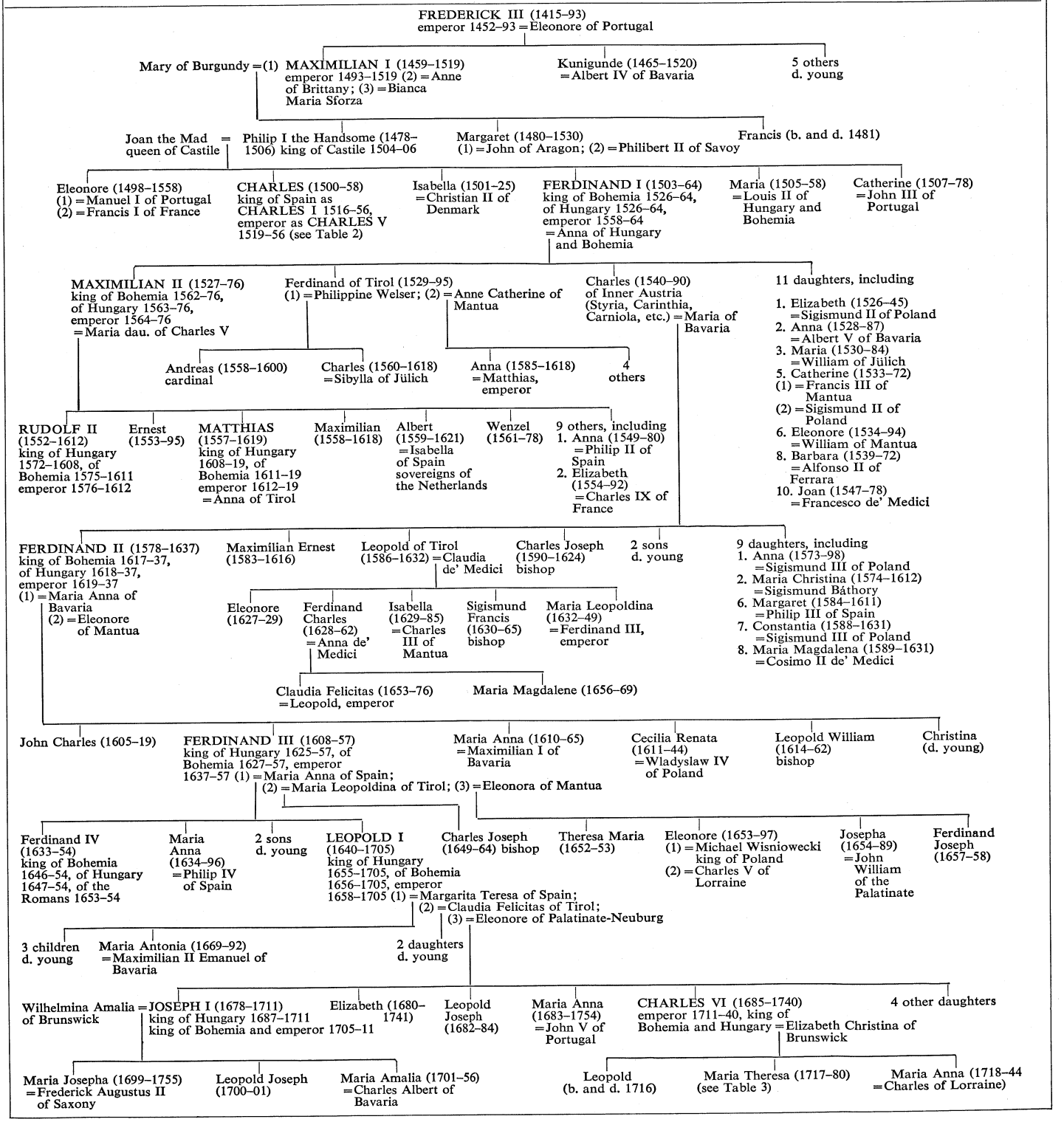
Frederick V, senior representative of the Inner Austrian line, was elected German king in 1440 and crowned Holy Roman emperor, as Frederick III, in 1452—the last such emperor to be crowned in Rome. A Habsburg having thus attained the Western world's most exalted secular dignity, a word may be said about the dynasty's major titles. The imperial title at this time was, for practical purposes, hardly more than a glorification of the title of German king; and the German kingship was, like the Bohemian and the Hungarian, elective. If Habsburg was to succeed Habsburg as emperor continuously from Frederick's death in 1493 to Charles VI's accession in 1711, the principal reason was that the hereditary lands of the Habsburgs, whether within the empire or outside it, formed an aggregate large enough and rich enough to enable the dynasty to impose its candidate on the other German electors (the Habsburgs themselves had an electoral vote only in so far as they were kings of Bohemia).

The emperor Frederick III

For the greater part of Frederick's reign it was scarcely foreseeable that his descendants would monopolize the imperial succession so long as they did. The Bohemian and Hungarian kingdoms were lost to the Habsburgs for nearly 70 years from the death of Ladislas Posthumus in 1457; the Swiss territories, lost in reality from 1315 onward (see SWITZERLAND, HISTORY OF), were finally renounced in 1474; and Frederick's control over the Austrian inheritance itself was long precarious, not only because of aggression from Hungary but also because of dissension between him and his Habsburg kinsmen. Yet Frederick, one of whose earliest acts in his capacity as emperor had been to ratify, in 1453, the Habsburgs' use of the unique title of "archduke of Austria" (first arrogated for them by Rudolf IV in 1358–59), may have had some prescient aspiration toward worldwide empire for the House of Austria: the motto *A.E.I.O.U.*, which he occasionally used, is generally interpreted as meaning *Austriae est imperare orbi universo* ("Austria is destined to rule the world"), or *Alles Erdreich ist Österreich untertan* ("The whole world is subject to Austria"). He lived long enough, in any case, to see his son Maximilian make the most momentous marriage in European history (see below); and three years before his death he also saw the Austrian hereditary lands reunited when Sigismund of Tirol abdicated in Maximilian's favour (1490).

Before explaining what the Habsburgs owed dynastically

Table 1: The Imperial Succession of the Habsburgs, 1493–1740



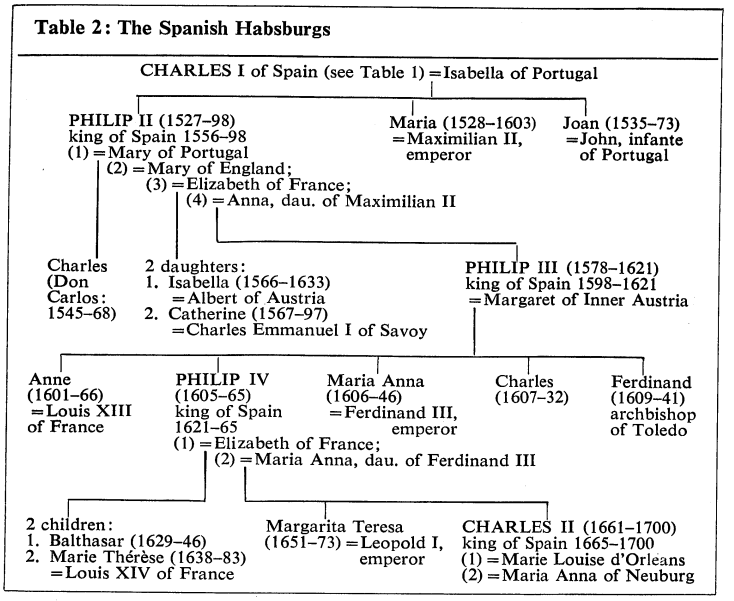
to Maximilian, mention can be made of a physical peculiarity characteristic of the House of Habsburg from the emperor Frederick III onward: his jaw and his lower lip were prominent, a feature supposed to have been inherited by him from his mother, the Mazovian princess Cymbarka. Later intermarriage reproduced the "Habsburg lip" more and more markedly, especially among the last Habsburg kings of Spain.

THE WORLD POWER OF THE HABSBURGS

Even before Frederick III's time the House of Habsburg had won much of its standing in Germany and in central Europe through marriages to heiresses. Frederick's son Maximilian carried this matrimonial policy to heights of unequalled brilliance. First he himself in 1477 married the heiress of Burgundy, Charles the Bold's daughter Mary, with the result that the House of Habsburg, in the person of their son Philip, inherited the greater part of Charles the Bold's widespread dominions: not the duchy of Burgundy itself, which the French seized, but Artois, the Netherlands, Luxembourg, and the County of Burgundy or Franche Comté. Secondly, though he failed after Mary's death in 1482 to secure Brittany also by a similar coup (France frustrated his proxy marriage to the

Maximilian: the Burgundian and Spanish marriages

Table 2: The Spanish Habsburgs



Breton heiress Anne), he procured Philip's marriage, in 1496, to Joan, prospective heiress of Castile and Aragon: thus securing for his family not only Spain, with Naples–Sicily and Sardinia, but also the immense dominions the Spaniards were about to conquer in America. Maximilian's matrimonial achievements were the occasion of the famous hexameter *Bella gerant alii, tu felix Austria nube* ("Let others wage wars: you fortunate Austria, marry").

Since Philip I of Castile died prematurely, his son was already ruler of the Burgundian heritage and of Spain when, in 1519, he succeeded Maximilian as ruler of the Habsburgs' Austrian territories. In the same year he was elected Holy Roman emperor as Charles V.

Charles V

The threat of force as well as an enormous expenditure in bribes was necessary to secure Charles's election. Besides the fact that many of the German princes were reluctant to saddle themselves with so mighty a sovereign, there was the opposition of France, which saw itself already half-encircled, from the northeast clockwise to the southwest, by Charles's possessions. Dating from Maximilian's Burgundian marriage, antagonism between the French kings and the Habsburgs was to persist, to the progressive detriment of the latter, until the middle of the 18th century, and until the second half of the 17th the other European powers would mostly sympathize with France. The Habsburgs in the 16th century were too formidable not to provoke envy and anxiety.

Charles V's responsibilities at the time of his becoming emperor were moreover too great for one man to assume, as he himself could acknowledge: they had to be divided. By the Treaty of Brussels (1522) he assigned the Habsburg-Austrian hereditary lands to his brother, the future emperor Ferdinand I. In 1521 Ferdinand had married Anna, daughter of Louis II of Hungary and Bohemia; and Louis II's untimely death in 1526, after his defeat by the Turks in the Battle of Mohács, prompted Ferdinand to stand as candidate for his succession, to which, despite rivals, he was elected.

The Habsburgs reached the zenith of their power before the end of the 16th century: the duchy of Milan, annexed by Charles V in 1535, was assigned by him to his son, the future Philip II of Spain, in 1540; Philip II conquered Portugal in 1580; and the Spanish dominions in America were ever expanding. There were, however, three faults in the power structure—two of them historical accidents, the third an effect of the Habsburg dynasty's own measures for self-preservation.

External and internal weaknesses in the 16th century

In the first place, the ascendancy of Charles V coincided with the outbreak of the Protestant Reformation in Germany, which was to spread turmoil for decades over Europe from the Netherlands to Hungary. As Charles, from his Spanish upbringing, was imbued with ideas of Catholic uniformity and as his successors, with the exception of the enigmatic Maximilian II, sought also to realize those ideas, religious resistance to the Habsburgs' authority came to aggravate or to camouflage political resistance. At the same time, the papacy, overawed though it was by the Spanish military presence in Italy, did not always subscribe to the Habsburg's special policy for Catholicism.

Secondly, Ferdinand's accession to Hungary meant that the Habsburgs had to bear the brunt of the Ottoman Turkish drive from the Balkans into central Europe, just as Habsburg Spain had to confront Turkish incursions into the western Mediterranean. The great victory of Lepanto (1571), won by a Habsburg bastard, did not end these troubles, which were exploited, against the dynasty, by Hungarian dissidents and, more covertly, by France.

The third flaw in the Habsburg edifice was latent in the 16th century. Mindful of what they had won by marriages, the Habsburgs sought to preclude rival dynasties from turning the tables on them by the same means: to keep their heritage in their own hands, they began to intermarry more and more frequently among themselves. The result, in a few generations, was a fatal inbreeding that brought the male line of Charles V to extinction.

By a series of abdications toward the end of his life Charles V transferred his Burgundian, Spanish, and Italian possessions to his son Philip II and his functions as emperor to his brother Ferdinand, who succeeded him formally as such after his death (1558). This division of the dynasty between imperial and Spanish lines was definitive: Ferdinand's male descendants were Holy Roman emperors until 1740 (for the emperors from Frederick III to Charles VI see Table 1), Philip's were kings of Spain until 1700 (see Table 2). The imperial line was inevitably concerned to maintain its position in Bohemia and to assert itself against the Turks in divided Hungary, because the loss of the two kingdoms would have meant the reduction of its possessions to what the Habsburgs had had hereditarily before Frederick III's time (the Austrian duchies and scattered holdings in Swabia and in Alsace)—a reduction that in turn would have compromised its chances of continuing to be elected to the German kingship. Philip II of Spain remained territorially the greatest sovereign in the Western world until his death in 1598; but the Revolt of the Netherlands (see LOW COUNTRIES, HISTORY OF), which he proved unable to subdue, was an irritation that his English and French enemies did their worst to inflame.

The imperial and the Spanish lines

Cooperation between imperial and Spanish Habsburgs in the 17th century failed to maintain the hegemony that the dynasty had enjoyed in the 16th. For the imperial

line, religious troubles in Germany and in central Europe went on even when the domestic conflict between the insane emperor Rudolf II and his brothers was over (1612); and the Bohemian insurrection of 1618 gave rise to that chain of wars involving the Austrian Habsburgs that, because it was prolonged until 1648, is known conventionally as the Thirty Years' War (*q.v.*). For the Spanish Habsburgs, their truce of 1609 with the Dutch ended in 1621, whereupon the renewed conflict in the Netherlands became merged with the struggles of their Austrian cousins. The Peace of Westphalia (1648) finally abolished Habsburg sovereignty over the northern Netherlands, severely restricted the emperor's authority over the other German princes, and transferred the Habsburg lands in Alsace to France; however the ordinance of 1627, whereby the Bohemian crown had been converted into a hereditary one for the Habsburgs, was permitted to stand.

The Thirty Years' War

France from the late 1620s had made the most of the Thirty Years' War to distress the Habsburgs of both lines, and peace with the imperial line did not prevent France from continuing its war against the Spanish until 1659, when by the Peace of the Pyrenees it obtained Gravelines, most of Artois, and part of Hainaut, together with some places south of Luxembourg.

The challenge of 17th-century France

The next 30 years saw the end of the Habsburg dynasty's claim to European hegemony in any real sense. The aggressions of Louis XIV of France, from 1667 onward, took territory after territory from the Spanish Habsburgs—large parts of Flanders, the rest of Artois, and other areas in the Netherlands, as well as the whole Franche Comté and, in 1684, the stronghold of Luxembourg—and demonstrated at the same time that the imperial Habsburgs, preoccupied as they were with the Turkish assault from Hungary, could not effectively defend the German frontier west of the Rhine. After being saved from the crisis of the Turkish siege of Vienna in 1683, the imperial Habsburgs did indeed obtain one dynastically significant success—the conversion, in 1687, of the Hungarian crown into an hereditary one for themselves—but by this time it was plain to Europe that the most formidable dynasty was no longer the Habsburg but the Bourbon. In the War of the Grand Alliance (1689–97) the rising powers that 100 years earlier had been Habsburg Spain's principal enemies and feeble France's most fluent encouragers, the Dutch and English, led those supporting the Habsburgs against Louis XIV.

Apart from the Bourbon ascendancy, there was a further reason for other powers to watch with jealous solicitude over the fate of Spain. The physical debility of Charles II of Spain was such that no male heir could be expected to be born to him, and the question of his succession was one of great concern to the European powers. Up to 1699 it was understood that his crowns would pass to the electoral prince of Bavaria, Joseph Ferdinand, son of his niece Maria Antonia, daughter of the emperor Leopold I; and this arrangement was generally acceptable because, by transferring the Spanish inheritance to the Bavarian House of Wittelsbach, it would not necessarily upset the balance of power between the imperial Habsburgs and Bourbon France. In 1699, however, when Joseph Ferdinand died, the moribund Charles II's next natural heirs were the descendants (1) of his half-sister, who had married Louis XIV of France, and (2) of his father's two sisters, of whom one had been Louis XIV's mother and the other the emperor Leopold I's. Critical tension developed: on the one hand neither the imperial Habsburgs nor their British and Dutch friends could consent to their Bourbon enemy's acquiring the whole Spanish inheritance; on the other neither Bourbon France nor its British and Dutch enemies wanted to see an imperial Habsburg reunite in one pair of hands most of what the emperor Charles V had had in 1519. Charles II in the meantime regarded any partition of his inheritance as a humiliation to Spain: dying in 1700, he named as his sole heir a Bourbon prince, Philip of Anjou, the second of Louis XIV's grandsons. The War of the Spanish Succession ensued.

The Spanish succession

THE HABSBURG SUCCESSION IN THE 18TH CENTURY

To allay British and Dutch misgivings, Leopold I and his elder son, the future emperor Joseph I, in 1703 renounced their own claims to Spain in favour of Joseph's brother Charles, so that he might found a second line of Spanish Habsburgs distinct from the imperial; but when Joseph I died, leaving only daughters, in 1711, and was succeeded by his brother as emperor (Charles VI) and as ruler of the Austrian, Bohemian, and Hungarian lands, the British and the Dutch lost interest in making him king of Spain and together began serious negotiations with France. Their Treaties of Utrecht (1713), which recognized the Bourbon accession to Spain and to Spanish America, virtually forced the hand of the reluctant Charles, who made peace with France by the Treaty of Rastatt in 1714: out of the whole inheritance of the Spanish Habsburgs, he had finally to content himself with the southern Netherlands and with the former Spanish possessions on the mainland of Italy, together with Mantua (annexed by him in 1708) and Sardinia. Sardinia, however, was exchanged by him in 1717 for Sicily, which the peacemakers of Utrecht had assigned to the House of Savoy. With characteristic obstinacy, Charles remained technically at war with Bourbon Spain until 1720, when an armistice was declared (formal recognition of the Bourbon accession came only in 1725).

Meanwhile the extinction of the Spanish Habsburgs' male line and the death of his brother Joseph left Charles, in 1711, as the last male Habsburg. He had therefore to consider what should happen after his death. No woman could rule the Holy Roman Empire, and furthermore the Habsburg succession in some of the hereditary lands was assured only to the male line. In order, therefore, to secure the indivisibility of his Habsburg inheritance he issued his famous Pragmatic Sanction of April 19, 1713, prescribing that, in the event of his dying sonless, the whole inheritance should pass (1) to a daughter of his, according to the rule of primogeniture, and thence to her descendants; next (2) if he himself left no daughter, to his late brother's daughters, under the same conditions; and finally (3) if his nieces' line was extinct, to the heirs of his paternal aunts. The attempt to win general recognition for his Pragmatic Sanction was Charles VI's main concern from 1716 onward (his baby son died in that year). By 1738, at the end of the War of the Polish Succession (in which he lost both Naples and Sicily to a Spanish Bourbon but got Parma and Piacenza for the Habsburgs in compensation), he seemed to have won his point: Saxony, Bavaria (grudgingly and with an express reservation), Spain, Russia, Prussia, Hanover-England, and finally France (with a reservation about third-party rights) had all, in one way or another, acknowledged the Pragmatic Sanction. His hopes were illusory: less than two months after his death, in 1740, his daughter Maria Theresa had to face a Prussian invasion of Silesia, which unleashed the War of the Austrian Succession. Bavaria then promptly challenged the Habsburg position in Germany; and France's support of Bavaria encouraged Saxony to follow suit and Spain to try to oust the Habsburgs from Lombardy. Great Britain came, late enough, to support Maria Theresa rather out of hostility toward France than out of loyalty to the Pragmatic Sanction.

The Pragmatic Sanction and the Austrian succession

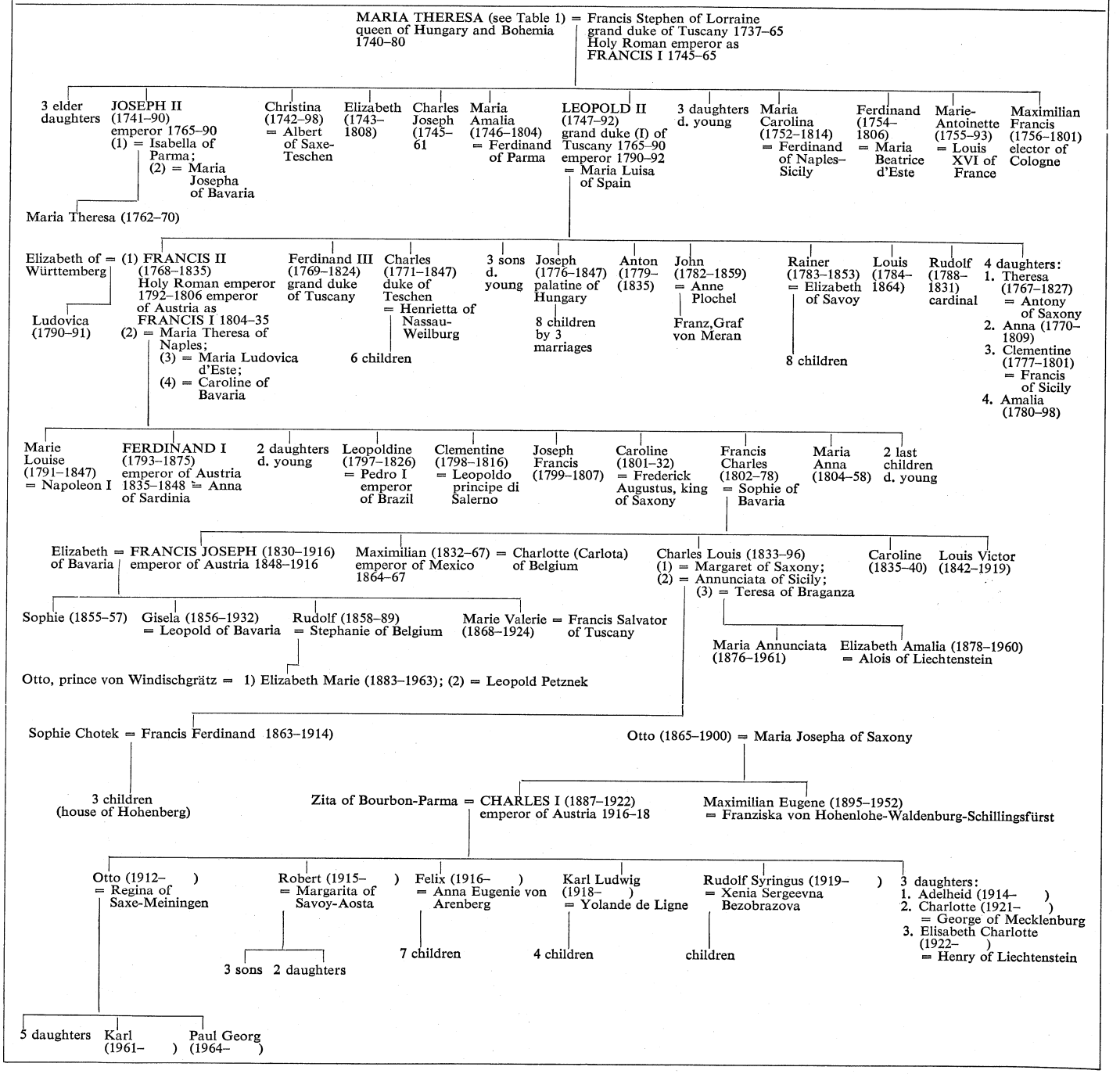
The challenge of Prussia

HABSBURG–LORRAINE

The War of the Austrian Succession cost Maria Theresa most of Silesia, part of Lombardy, and the duchies of Parma and Piacenza (Treaty of Aix-la-Chapelle, 1748) but left her in possession of the rest of her father's hereditary lands. Moreover, her husband, Francis Stephen of Lorraine, who in 1737 had become hereditary grand duke of Tuscany, was finally recognized as Holy Roman emperor, with the title of Francis I. He and his descendants, of the House of Habsburg–Lorraine (see Table 3 below), are the dynastic continuators of the original Habsburgs.

The peace of 1748 did not last long. Prussia was not

Table 3: The House of Habsburg–Lorraine in Austria



satiated by the seizure of Silesia from the Habsburgs, and they in turn were even more determined to recover Silesia than anxious to ensure the protection of their outlying possessions in the Netherlands against the continuing danger of French attack. The so-called Diplomatic Revolution, which preceded the Seven Years' War (*q.v.*) of 1756–63, was the product, basically, of these situations: finding that their former British friends were more interested in conciliating Prussia than in abetting Austro-Russian plans for destroying it, the Habsburgs played their part in the "reversal of alliances" by achieving a reconciliation with France, hitherto their longest-standing enemy. Though the new alliance and the consequent war brought no territorial profit to the Habsburg dynasty, an Austro-French entente was subsequently maintained until 1792: the marriage of the archduchess Marie-Antoinette to the future Louis XVI of France (1770) was intended to confirm it.

To secure their imperial status in Germany against Prussian enterprises, the Habsburgs exerted themselves to consolidate and to expand their central European bloc of territory. For this purpose Tuscany and the Netherlands were practically irrelevant. Tuscany in fact was kept separate from the ancient Habsburg inheritance: when the emperor Francis I died (1765), his eldest son, the emperor Joseph II, became coregent with his mother of the Austrian dominions, but Joseph's brother Leopold became grand duke of Tuscany; and similarly when Leopold succeeded to Joseph's titles (1790), his own second son succeeded to Tuscany as Ferdinand III. Thereafter the Tuscan branch of the Habsburgs remained distinct from the senior or imperial line.

The northeastward expansion of Habsburg central Europe, which came about in Joseph II's, time was a result not so much of Joseph's initiative as of external events: the First Partition of Poland (1772), which gave

him Galicia and Lodomeria, was a Russo-Prussian arrangement disgusting to his conscientious mother, who remembered Silesia; and his subsequent acquisition of Bukovina (1775), geopolitically logical though it was as bridging a gap between his Transylvanian and his new Galician lands, was a side effect of the Russo-Turkish Treaty of Küçük Kaynarca (1774).

Joseph II was considerably more interested in westward expansion, over Bavaria, which would have both strengthened his western frontier strategically and enhanced his status among the German princes politically. Prussia's forceful opposition. however, reduced his gains in the War of the Bavarian Succession to the Innviertel (1779) and frustrated his plan for ceding the Netherlands to the House of Wittelsbach in exchange for Bavaria five years later (1784).

The French Revolutionary and Napoleonic Wars (*q.v.*) brought a kaleidoscopic series of changes. Three were clearly significant for the future of the House of Habsburg: (1) the formal dissolution of the Holy Roman Empire in 1806, in anticipation of which Leopold II's successor Francis II had in 1804 begun to style himself "hereditary emperor of Austria," a title that, as Francis I, he could retain come what might; (2) the definitive renunciation of the southern Netherlands by the Habsburgs in 1797; and (3) the awakening of the spirit of nationalism in the modern sense.

The Austrian Empire

On Napoleon's downfall the Congress of Vienna (1814–15) inaugurated the Restoration, from which the battered House of Habsburg naturally benefitted. Francis I of Austria recovered Lombardy (lost in 1797), Venetia and Dalmatia (both of them acquired in 1797 but lost in 1809), and Tirol (lost also in 1809); Ferdinand III of Tuscany recovered his grand duchy; another Habsburg was recognized as sovereign duke of Modena, because his father, a brother of the Holy Roman emperors Joseph II and Leopold II, had in 1771 married the heiress of the House of Este; and Napoleon's Habsburg consort, Marie Louise, received the duchies of Parma and Piacenza for her lifetime (after which they were to revert to the Bourbons). The territory of Salzburg, which the Habsburgs had acquired in 1803 but lost to Bavaria in 1809, was finally restored to Austria in 1816. Though the Congress of Vienna did not restore Austrian rule over "Western Galicia" (the Habsburgs' share under the Third Partition of Poland in 1795, lost likewise in 1809), a small part of that area, namely the territory of Cracow, was annexed by Austria in 1846.

Nationalism against the dynasty

The history of the House of Habsburg for the century following the Congress of Vienna is inseparable from that of the Austrian Empire, a bastion of monarchical conservatism that the forces of nationalism—German, Italian, Hungarian, Slav, and Romanian—gradually eroded. The first territorial losses came in 1859, when Austria had to cede Lombardy to Sardinia–Piedmont, nucleus of the emergent kingdom of Italy, and could do nothing to prevent the same power from dispossessing the Habsburgs of Tuscany and of Modena. Next, the Seven Weeks' War of 1866, in which Prussia, exploiting German nationalism, was in alliance with Italy, forced Austria both to renounce its hopes of reviving its ancient hegemony in Germany and to cede Venetia. After this disaster the Habsburg emperor Francis Joseph took a step intended to consolidate his "multinational empire": in 1867, to conciliate Hungary, he granted to that kingdom equal status with the Austrian Empire in what was henceforth to be the Dual Monarchy of Austria–Hungary. The result, however, was that the Magyars, jealous of their unique parity with the Germans and of their superiority over the non-Magyar peoples of their kingdom, rejected any suggestion of conciliating the Slavs and the Romanians of the Dual Monarchy by similar measures. The ardent German nationalists of the Austrian Empire, as opposed to the Germans who were simply loyal to the Habsburgs, took the same attitude as did the Magyars.

Remote from Austria's national concerns but still wounding to the House of Habsburg was the fate of Francis Joseph's brother Maximilian: set up by the French as emperor of Mexico in 1864, he was executed by a Mexican firing squad in 1867. No less grievous to the dynasty and of more concern to Austria–Hungary was the suicide of the crown prince Rudolf in 1889, though his fitness for the imperial and royal succession was questionable; and the scandalous misconduct of certain archdukes and archduchesses, in the imperial and in the Tuscan lines alike, further impaired the Habsburgs' personal prestige. The assassination of Francis Joseph's Wittelsbach consort Elizabeth in 1898 was to be followed in less than two decades by an assassination of far greater consequence.

In 1878 Austro-Hungarian forces had "occupied" Bosnia–Hercegovina, which belonged to decadent Turkey. In 1908 that territory had been formally annexed to Austria–Hungary, in a manner that was outrageous not only to Serbia (which coveted Bosnia for itself) but also to Serbia's patron, Russia. Visiting the Bosnian capital, Sarajevo, in 1914, the archduke Francis Ferdinand, heir presumptive to the Dual Monarchy (and incidentally legatee, from 1875, of the rights of the House of Austria–Este to Modena), was shot to death by a nationalist Serb. A month later the First World War was beginning.

World War I and after

World War I led to the dismemberment of the Habsburg Empire. While Czechs, Slovaks, Poles, Romanians, Serbs, Croats, Slovenes, and Italians were all claiming their share of the spoil, nothing remained to Charles, the last emperor and king, but "German" Austria and Hungary proper. On November 11, 1918, he issued a proclamation recognizing Austria's right to determine the future form of the state and renouncing for himself any share in affairs of state, and on November 13 he issued a similar proclamation to Hungary. Even so, he did not abdicate his hereditary titles either for himself or for the Habsburg dynasty. Consequently the national assembly of the Austrian Republic passed the "Habsburg Law" of April 3, 1919, banishing all Habsburgs from Austrian territory unless they renounced all dynastic pretensions and loyally accepted the status of private citizens. In Hungary, however, the collapse of the republican regime at the end of 1919 raised strong royalist hopes of a Habsburg restoration, and after the conclusion of the Treaty of Trianon (June 1920) Charles twice tried to return (March and October 1921). Under pressure from the other European powers, especially those of the Little Entente (Czechoslovakia, Yugoslavia, and Romania), the Hungarian parliament on November 3, 1921, decreed the abrogation of Charles's sovereign rights and of the Pragmatic Sanction.

Habsburg property rights in Austria, forfeited under the law of 1919, were restored in 1935 but withdrawn again by the German chancellor Adolf Hitler in 1938. After World War II the Allied Control Council in Austria in January 1946 declared that it would support the Austrian government in measures to prevent any return of the Habsburgs, and the law of 1919 was written into the Austrian State Treaty of 1955. In June 1961 the Austrian government rejected an application by the archduke Otto, head of the House of Habsburg, to be allowed to return to Austria as a private citizen, but in 1963 the administrative court of Austria ruled that Otto's application was legal. Because of Socialist opposition to his return, however, he was not granted a visa until June 1966 after the People's Party had won a majority in that year's general election.

BIBLIOGRAPHY. For the English reader, the most comprehensive introduction to the subject remains ARCHDEACON WILLIAM COXE, *History of the House of Austria*, 3rd ed., 4 vol. (1847–53), a neglected work covering Habsburg history from 1218 to 1848; A. WANDRUSZKA, *Das Haus Habsburg* (1956, 2nd ed. 1959; Eng. trans., *The House of Habsburg*, 1964), covers 600 years more briefly but most authoritatively. Works in English on the Habsburgs in the 19th and 20th centuries vary in scope and in quality: C.A. MACARTNEY, *The Habsburg Empire, 1790–1918* (1968), provides a masterly survey by an expert; A.J.P. TAYLOR, *The Habsburg Monarchy, 1809–1918*, rev. ed. (1948, paperback 1964), is rather severely critical of

the dynasty's failings; E. CRANKSHAW, *The Fall of the House of Habsburg* (1963, paperback 1969), makes a brilliant and easily readable vindication of the last emperors; A.J. MAY, *The Habsburg Monarchy, 1867–1914* (1960), is fair and scholarly, with a good bibliography; Z.A.B. ZEMAN, *The Break-up of the Habsburg Empire 1914–1918* (1961), gives a just account of a subject usually misrepresented. For genealogical detail, see W. MERZ, *Die Habsburg* (1896), which contains 19 tables; and M. DUGAST-ROUILLE, *Les Maisons souveraines de l'Autriche* (1967), also well tabulated and with illustrations. For the physical heredity, see O. RUBBRECHT, *L'Origine du type familial de la maison de Habsbourg* (1910); W. STROHMAYER, *Die Vererbung des Habsburger Familientypus* (1937); and J. NADA, *Carlos, the Bewitched* (1963; U.S. title, *Carlos, the King Who Would Not Die*, 1963). For original research, the *Inventare* (1933–) of the Haus-, Hof- und Staatsarchiv in Vienna (known as the Austrian State Archives from 1945) are an indispensable guide.

(J.R.-S.)

Ḥadīth

Ḥadīth in Islām is the record of the traditions or sayings of the Prophet Muḥammad, revered and received as a major source of religious law and moral guidance, second only to the authority of the Qur'ān, or scripture of Islām. It might be defined as the biography of Muḥammad perpetuated by the long memory of his community for their exemplification and obedience. The development of Ḥadīth is a vital element during the first three centuries of Islāmic history, and its study provides a broad index to the mind and ethos of Islām.

NATURE AND ORIGINS

The term Ḥadīth derives from the Arabic root *ḥdth*, meaning "to happen," and so, "to tell a happening," "to report," "to have, or give, as news," or "to speak of." It means tradition seen as narrative and record. From it comes the Sunnah (that is, tradition taken as precedent and authority or directive), to which the faithful conform in submission to the sanction that Ḥadīth possesses and that legalists, on that ground, can enjoin. Tradition in Islām is thus both content and constraint, Ḥadīth as the biographical ground of law and Sunnah as the system of obligation derived from it. In and through Ḥadīth, Muḥammad may be said to have shaped and determined from the grave the behaviour patterns of the household of Islām by the posthumous leadership his personality exercised. There were, broadly, two factors operating to this end. One was the unique status of Muḥammad in the genesis of Islām; the other was the rapid geographical expansion of the new faith in the first two centuries of its history into various areas of cultural confrontation. Ḥadīth cannot be rightly assessed unless the measure of these two elements and their interaction is properly taken.

Sunnah

The experience of Muslims in the conquered territories of west and middle Asia and of North Africa was related to their earlier tradition. Islāmic tradition was firmly grounded in the sense of Muḥammad's personal destiny as the Prophet—the instrument of the Qur'ān and the apostle of God. The clue to tradition as an institution in Islām may be seen in the recital of the *Shahādah* or "witness" ("There is no god but God and Muḥammad is the apostle of God"), with its twin items as inseparable convictions—God and the messenger. Islāmic tradition follows from the primary phenomenon of the Qur'ān. Muḥammad had come through a variety of factors to the sense of prophetic vocation and to the assurance of finality within it. The Qur'ān, received by him in revelation-inspiration (*waḥy*), had never been a corporate activity but always solitary and personal. It had been accompanied by striking features of poetic eloquence and intensity and even of physical "signs," all of which contributed to the drama of its accumulation, recited as literature and acclaimed by disciples. Its sequences developed with the biographical movement of the Prophet's career of leadership, to which its contents gave form and occasion. Its 114 *sūrah*s, or chapters, synchronized with the events of 23 years of Muḥammad's life (from AD 609 to 632). It was thus inextricably bound up with his person and the agency of his vocation. Acknowledgment of the Qur'ān as scripture by the Islāmic community was inseparable from acknowledgment of Muḥammad as its appointed recipient. In that calling, Muḥammad had neither fellow nor partner, for God, according to the Qur'ān, spoke only to Muḥammad. When he died, therefore, in AD 632, the gap thus created in the emotions and the mental universe of Muslims was shatteringly wide. It was also permanent. That death had also terminated the revelation embodied in the Qur'ān. By the same stroke scriptural mediation had ended, as well as prophetic presence.

The Prophet's death was said to have coincided with the perfection of revelation. But the perfective closure of both the book and the Prophet's life, though in that sense triumphant, was also onerous, particularly in view of the new changing circumstances, both of space and time, in the gathering geographical expansion of Islām. In all the new pressures of historical circumstance, where was direction to be sought? Where, if not from the same source as the scriptural mouthpiece, who by virtue of that consummated status had become the revelatory instrument of the divine word and could therefore be taken as an everlasting index to the divine counsel? The instinct to and the growth of tradition making are thus integral elements in the very nature of Islām, Muḥammad, and the Qur'ān. Ongoing history and the extending dispersion of Muslim believers provided the occasion and spur for the compilation of Ḥadīth.

HISTORICAL DEVELOPMENT

The appeal of the ordered recollection of Muḥammad to the Islāmic mind did not become immediately formalized and sophisticated. On the contrary, there is evidence that the full development of Ḥadīth was slow and uneven. Time and distance had to play their role before memory became stylized and official.

Tradition in pre-Islāmic Arabia. The first generation had its own immediacy of Islāmic experience, both within the life span of the Prophet and in the first quarter century afterward. It had also the familiar patterns of tribal chronicle in song and saga. Pre-Islāmic poetry celebrated the glory of each tribe and their warriors. Such poetry was recited in honour of each tribe's ancestors. The vigour and élan of original Islām took up these postures and baptized them into Muslim lore. The proud history of which Muḥammad was the crux was, naturally, the ardent theme, first of chronicle, and then of history writing. Both needed and stimulated the cherishing of tradition. The lawyers, in turn, took their clues from the same source. While the Qur'ān was being received, there had been reluctance and misgiving about recording the words and acts of the Prophet, lest they be confused with the uniquely constituted contents of the scripture. But the very traditions that reveal Muḥammad disapproving of the practice of recording his words are evidence enough that the practice existed. With the Qur'ān complete and canonized, those considerations no longer obtained; and time and necessity turned the instinct for Ḥadīth into a process of gathering momentum.

Relation to early Arabic poetry

Developments of the 1st and 2nd centuries AH. Within the first century of the Prophet's death, tradition had come to be a central factor in the development of law and the shape of society. Association by Ḥadīth with Muḥammad's name and example became increasingly the ground of authority. The 2nd century brought the further elaboration of this relationship by increasing formalism in its processes. Traditions had to be sustained by an expert "science" of attestation able to satisfy rigorous formal criteria of their connection with the person of Muḥammad through his "companions," by an unbroken sequence of "reportage" (see below). This science became so meticulous that it is fair (even if also paradoxical) to suspect that the more complete and formally satisfactory the attestation claimed to be, the more likely it was that the tradition was of late and deliberate origin. The developed requirements of acceptability that the tradition boasted simply did not exist in the early, more haphazard and spontaneous days.

It is clear that many customs and usages native to non-Arab societies prior to their Islāmization found their way into Islām in the form of reputed or alleged traditions of Muḥammad, though always on the condition of their general compatibility with the Islāmic religion. Implicit in this sense in Muḥammad's personal example and genius, tradition inferred an elasticity and an embrace large enough to comprehend and anticipate all that Islām in its wide geographical experience was to become.

Qur'ānic commentary

Qur'ānic commentary, as it developed in the wake of these other factors of law and custom, also leaned heavily on traditional material, for the incidents of the Qur'ānic narrative and the occasions of revelation could best be understood by what tradition had to say in its reporting of them. Further, since the patterns of Qur'ānic commentary were largely hortatory, Ḥadīth was a ready mine of word and story calculated to exemplify and reinforce what exhortation commended. Except in rare and controversial cases (the so-called Ḥadīth Qudsī, or Holy Tradition), these traditional factors in Qur'ānic interpretation were only elucidatory, and the substance of tradition could in no way dispute or displace the essential, primary, authority of the Qur'ānic text. For the *obiter dicta* (incidental observations) of Muḥammad, though sacrosanct, did not have the hallmark of revelation that belonged solely to the Qur'ān. Among earliest developed examples of Ḥadīth are the narratives of the biographer Ibn Isḥāq (died AH 150 [AD 767]) and the compilation of laws by Mālik ibn Anas, known as al-Muwaṭṭa' (died AH 179 [AD 795]). But they preceded by less than half a century the success of the theory that made tradition indispensable to the valid development of Islāmic law.

3rd century AH and subsequent developments. The chief protagonist of the view correlating tradition and law was Muḥammad ash-Shāfi'ī (died AH 204 [AD 820]) who claimed for tradition a divine imprint as an extension of the revelation of the Qur'ān. It was in line with this conviction that the phrase "the Qur'ān and the Sunnah" became current to describe the fount of authority in Sunnī Islām (the major traditionalist sect). By this mandate and out of the needs and inventiveness of lawyers, the mass of tradition grew apace. When virtually no issues could be argued, still less settled, except by connection with cited acts and opinions of Muḥammad, the temptation to require or to imagine or to allege such traditions became irresistible. Supply approximated to demand, and the growth of both made more ingenious and pretentious the science of supporting attribution. The increasing volume and complexity of the material contained in Ḥadīth necessitated larger compilations and more detailed classification. These factors worked together to inspire a critical editorial activity that in the course of the 3rd century generated what have come to be regarded as the six canonical collections of Ḥadīth by Sunnī Muslims. The first two of them have acquired a status of great sanctity. Before noting these it is convenient to describe the editorial task and the editorial procedures that constitute the developed science of Ḥadīth criticism.

THE SCIENCE OF HADITH

Matn and *isnād*

The study of tradition distinguishes between the substance, or content, known as the "gist" (*matn*) of the matter, and the "leaning" (*isnād*) or chain of corroboration on which it hangs.

Form of Ḥadīth and criteria of authentication. That Muḥammad observed, "Seek knowledge, though it be in China" or "Beware of suspicion, for it is the falsest of falsehoods" reveals the *matn* or "the meat of the matter." The formula introducing such a Ḥadīth would speak in the first person: "It was related to me by A, on the authority of B, on the authority of C, on the authority of D, from E (here a companion of Muḥammad) that the Prophet said. . . ." This chain of names constituted the *isnād* on which the saying or event depended for its authenticity. The major emphases in editing and arguing from tradition always fell on the *isnād*, rather than on a critical attitude to the *matn* itself. The question was not, "Is this the sort of thing Muḥammad might credibly be imagined to have said or done?" but "Is the report that he said or did it well supported in respect of witnesses and transmitters?" The first question would have introduced too great a danger of subjective judgment or independence of mind, though it may be suspected that issues were in fact often decided by such critical appraisal in the form of decisions ostensibly relating only to *isnād*. The second question certainly allowed a theoretically objective and reasonably precise pattern of criteria.

If the adjacent names in the chain of transmission overlapped in life, there was certainty that they could have listened to one another. Their travels were also investigated to see if their paths could have really crossed. Biographies could be built up to show that they were honest men and spoke truly. Comparative study could be made of their reputations for veracity as acknowledged by their contemporaries or indicated by their traditions when compared. The frequency of currency through several sources was yet another element in the testing of traditions. Most important of all was the final link with the "companion," who in the first instance had the tradition from his or her contact with the Prophet.

Classifications. In all these ways, and others involving more minutiae, it was possible to establish categories of Ḥadīth quality. Traditions might be sound (*saḥīḥ*), good (*ḥasan*), or weak (*ḍā'īf*). Other terms, such as healthy (*ṣāliḥ*) and infirm (*saqīm*), were also current. Each of the three classifications was liable to subdivisions, depending on refinements of assessment and, later, on their standing with the classic compilers. Distinctions were less rigorously seen if the traditions were cited not for legal definitions but merely for moral purposes. A *ḍā'īf* tradition, for example, might well be salutary for exhortation, even if lawyers were required to exclude or ignore it. Traditions also varied in strength according to whether one or more "companions" could be adduced, whether the *isnād* had parallels, whether they were continuous back to Muḥammad (*muttaṣil*), or intermitted (*mawqūf*). The subtleties in these and other questions were part of the active competence that attended the whole science.

Weight of traditions

The repute and authority of the canonical collections did much to stabilize the situation, but only because their emergence demonstrated that the zest for tradition had overreached itself. By the end of the 3rd century AH it was sorely necessary to solidify Ḥadīth into a stable corpus of material to which no new element could credibly be added and from which the extravagances of the process of development had been purged. The Ḥadīth tradition within the various traditions had by then become a permanent and disciplined element in the authority structure of Islām—the second great source of law and practice, complementary to the Qur'ān and available for analogical handling (*qiyās*) and for consensus (*ijtihād*) as further sources of legislation, arguing from the Qur'ān and the Sunnah as primary. Shī'ah tradition (see below) stands apart from this structure of authority.

THE COMPILATIONS

The most revered of all traditionalists was Muḥammad ibn Ismā'īl al-Bukhārī (AH 194–256 [AD 810–870]), whose *Kitāb al-Jāmi' aṣ-Ṣaḥīḥ* (*The Book of the Authentic Collection*) has a place that is unique in the awe and esteem of Muslims as a work of great historical import and of deep piety. While still a boy he made the pilgrimage to Mecca and gathered traditions in wide travels. According to tradition, he was inspired to his editorial task by a vision of Muḥammad pestered by flies while asleep—flies that he (al-Bukhārī) fanned away from the face of the Prophet. The flies represented the cloud of spurious traditions darkening the true image, and the fan was its tireless rescuer. Whatever the truth of this narrative, it captures the temper of al-Bukhārī's vocation. His *Ṣaḥīḥ* occupied 16 years of editorial pains and scrutiny. He included 7,397 traditions with full *isnād*. Allowing for repetitions, the net total was 2,762, gathered, it is said, from over 600,000 memorized items. He arranged the whole

into 97 books and 3,450 chapters or topics, repeating the traditions that bore on several themes.

Of comparable stature was the *Ṣaḥīḥ* of Muslim ibn al-Ḥajjāj (AH 202–261 [AD 817–875]), to which the compiler prefaced a discussion of the criteria of Ḥadīth. The material largely confirms his contemporaries, and all such traditions common to these two authorities are known as agreed (*muttafaq*). It became characteristic to give freer rein to prevailing or communal assent in matters of *isnād*.

There are four other classical collections of tradition, all belonging within the 3rd century AH, and interdependent in part. Abū Dā'ūd al-Sijistānī (AH 202–275 [AD 817–889]) produced his *Kitāb as-Sunan* ("Book of traditions"), containing 4,800 traditions relating to matters of jurisprudence (as the term *Sunan* indicates, in contradistinction to a *Jāmiʿ*, or collection embracing all fields). Abū ʿIsā Muḥammad at-Tirmidhī (died AH 279 [AD 892]) edited the *Jāmiʿaṣ-Ṣaḥīḥ*, adding notes on the distinctive interpretations of the schools of law (*madhāhib*). Abūʿ Abd ar-Raḥmān an-Nasā'ī (AH 216–303 [AD 830–915]) produced another *Kitāb as-Sunan* with special concern for the religious law relating to ritual acts. Abūʿ Abdallāh ibn Mājā (AH 210–273 [AD 824–886]), a pupil of Abū Dā'ūd, compiled another with the same title but tended to a readier tolerance of less than satisfactory traditions. Preferences shifted between these four, and some were slower of recognition than others. Nor did they oust the earlier collection of Mālik ibn Anas, which maintained, if intermittently, its wide appeal. But they formed the increasing reliance of generations of Muslims, within the unique eminence of the master "pair," and formed the sources of later popular editions, intended to conflate material for didactic purposes. One such was the work of Abū Muḥammad al-Baghawī (died AH 516 [AD 1122]) called *Maṣābīḥ as-Sunnah* ("The Lamps of the Sunnah"). Commentaries on all these classical *musannafāt*, or compilations, were many, and important in education and piety.

SECTARIAN VARIATIONS

Shīʿah differences

The Shīʿah tradition (distinguished from the Sunnī tradition by belief in the special role of the Prophet's cousin ʿAlī and his descendants) diverges sharply from a very early date, though the emphasis on the personality of Muḥammad was identical. The Shīʿah minority broke away from the (to be) dominant Sunnī stream of Islām for deep reasons of politics, emotion, and theology. There was the dispute about caliphal succession and the role of ʿAlī, cousin and son-in-law of Muḥammad and fourth caliph, and bitter cleavage because of the tragic fate of his two sons and especially of Ḥusayn in the massacre of Karbalāʿ, from which there ultimately evolved the theology of vicarious suffering epitomized in Shīʿah devotion and ritual. All these factors inevitably involved the business of tradition. The schism read the origins according to the divided loyalties, and there was little that was not potentially contentious, apart from obvious matters; *e.g.*, Muḥammad's intentions for ʿAlī and the caliphate. The issues were fought out in rivalry for the mind of the Prophet, the authority of which was the sole agreement in the very disputing of it. The Shīʿah thus rejected the Sunnah of the Sunnīs and developed their own corpus of tradition (though there is evidence that an-Nasā'ī, at least, among the classical compilers, had sympathy with aspects of their cause). They also questioned the Sunnī notions of *isnād* and of the community as a locus of authority and evolved their own system of submission to their *imām*s (Shīʿah leaders). This altered the whole role that tradition might play. The major Shīʿah compilations date from the 4th and 5th centuries and allow only traditions emanating from the house of ʿAlī. The first of them is that of Abū Jaʿfar Muḥammad al-Qulīnī (died AH 328 [AD 939]), *Kāfī fī ʿIlm ad-Dīn*, which might be translated: "All You Need About the Science of Religious Practice."

SIGNIFICANCE OF HADITH

To browse in the canonical collections of Ḥadīth is to be initiated into a fascinating world of faith, of behaviour and authority, a world of almost encyclopaedic inclusiveness. Provisions of law are the primary element, enlarging Qur'ānic legislation. They contain a whole array of moral, social, commercial, and personal matters, as well as the themes of eschatology. All reaches of public and private conduct may be found there, from the disposal of a date stone to the crisis of the deathbed, from the manner of ablution to the duties of forgiveness, from the physical routines of digestion to the description of the day of judgment. There is a Talmudic capacity for detail and scrupulousness in legal and ethical prescriptions and precepts. There are many delightful stories of integrity and right action like that of the purchaser of a plot of ground who subsequently unearthed in it a pot of gold, which he brought back to the former owner, protesting that it was not within his bargain. The vendor, likewise, refused to claim it since he had not known it was there when he sold his field. An arbitrator solved their dilemma of honesty by proposing the marriage of the son of one with the daughter of the other so that, after alms, the gold might be settled on the couple. Through and in tradition, Islām aligned itself authoritatively with all it found compatible in local usages and brought hospitably and masterfully within its purview the continuity of many cultures. There is wide evidence of the impact of Jewish and Christian elements, notably in the realm of eschatology, elaborating the stark and urgent Qur'ānic doctrine of the last judgment. But always the imprint of Islām is clear. Tradition is at once a mine and a kind of currency, the source and the circulation of the values it makes and preserves.

BIBLIOGRAPHY. J. ROBSON, "Hadīth," in *The Encyclopaedia of Islam*, new ed., vol. 3, pp. 23–28 (1965), and T.W. JUYNBOLL, "Hadīth," in *The Encyclopaedia of Islam*, vol. 2, pp. 189–194 (1927), two important summaries with extensive bibliographies; ALFRED GUILLAUME, *The Traditions of Islam* (1924), a general introduction serviceable for a first study; I. GOLDZIHER, *Études sur la tradition islamique*, ed. by L. BERCHER (1952), a French trans. of the major part of vol. 2 of *Muhammedanische Studien* (1890), an indispensable work; MUHAMMAD ALI, *A Manual of Hadith* (1944), a general selection, mainly from al-Bukhārī, in Arabic and English; A.J. WENSINCK, *A Handbook of Early Muhammadan Tradition* (1927), an alphabetical arrangement by a great Dutch scholar; M.Z. SIDDIQI, *Hādīth Literature: Its Origins, Development, Special Features, and Criticism* (1961), an Asian Muslim's presentation; J. SCHACHT, *Origins of Muhammadan Jurisprudence* (1950), a very able analysis of legal traditions.

(A.K.C.)

Hadrian

Hadrian was emperor of Rome from August 11, AD 117, until July 10, 138. An ardent admirer of Greek civilization and a creative artist in his own right, Hadrian was one of the most cultivated of all the Roman emperors. His complex personality still defies analysis. Capable of cruelty and deception, he combined a longing for peace with a taste for strict military discipline. He strengthened the frontier defenses of Britain but abandoned the eastern conquests of his predecessor, Trajan. His restless spirit drove him to so much travel that he was away from Rome more than he was in it. His homosexuality culminated in an infatuation that launched a new cult throughout the Roman Empire. Almost by accident, Hadrian provoked a rebellion among the Jews and earned their bitter, unceasing hatred. This passionate and erudite philhellene is a fitting symbol of the Greco-Roman world of the 2nd century.

Early life. The family of Hadrian came from the town of Italica, near Seville, in southern Spain. They were not, however, of native Spanish origin but rather of settler stock. Hadrian's forebears left Picenum in Italy for Spain about 250 years before his birth. Hadrian himself was probably born in Rome, as one of the ancient sources for his life explicitly states. Many modern biographers have assumed without good reason that Hadrian was born in Italica and have proceeded to postulate a powerful influence of Spanish sun and temperament upon the young man. In fact, there is nothing particularly Spanish about



Hadrian, bust in the Museo Archeologico Nazionale, Naples.
Anderson—Mansell

Hadrian. He bears the stamp of education in cosmopolitan Rome.

Hadrian, whose full Roman name was Publius Aelius Hadrianus, was born on January 24, 76. When his father died in 85, he was entrusted to the care of two men: one, a cousin of his father, was the future emperor Trajan, and the other was Acilius Attianus, who was later to serve as prefect of the emperor's Praetorian Guard early in Hadrian's own reign. In the year 90, Hadrian visited the home of his ancestors in Spain probably for the first time. At Italica he received some kind of military training and also developed a fondness for hunting that stayed with him for the rest of his life. It would seem, however, that Hadrian did not much care for the small-town provincial life of Italica. He remained there for only a few years, and his subsequent relations with the place were not cordial. When he returned to Spain as emperor, he avoided Italica altogether.

Rise to power. Trajan was consul in 91, and his ward moved into several junior posts of importance. Hadrian began to follow the traditional career of a Roman Senator, which meant advancement through a conventional series of posts, and held the post of military tribune with three Roman legions in succession. In about 95 he served with the Legio II Adjutrix in the province of Upper Moesia, on the Danube River, whence he transferred in the next year to the adjacent province of Lower Moesia (with the Fifth Macedonica). Toward the end of 97 Hadrian was chosen to go west to Gaul to convey congratulations to Trajan, whom the aged emperor Nerva had just adopted and thereby designated his successor. Trajan's ward now belonged to the governing circles of the empire. Inevitably, hostility and envy awaited him in that society. In 98 Julius Servianus, his brother-in-law, attempted to prevent him from being the first to inform Trajan of Nerva's death; Servianus wanted the distinction for himself, but his plan failed. Thereafter, the two men were probably never on cordial terms, for Servianus posed a constant threat to Hadrian's position.

Plotina's protégé

The greatest single political figure behind the emperor Trajan was the man who masterminded his elevation, Lucius Licinius Sura. Hadrian was fortunate to enjoy this man's favour, and, as long as Sura was alive, Hadrian's career advanced briskly. Trajan's wife, Plotina, seems also to have been close to Sura and a partisan of Hadrian. For a time Servianus could do no harm. Through Plotina's favour, Hadrian married Trajan's grand-niece, Vibia Sabina, in 100. In the following year Hadrian held the senatorial office of quaestor and in 102 served as Trajan's companion in the Emperor's first war in the kingdom of Dacia on the Danube. In 105 Hadrian rose to the office of tribune of the plebs and, exceptionally, advanced to the praetorship in the very next year. No less exceptional than the speed of promotion was the fact that Hadrian served as praetor while in the field with the Emperor during his second war in Dacia. In 107 he had a brief tenure as governor of the Danubian province of Lower Pannonia. Then, in 108, Hadrian reached the coveted pinnacle of a Senator's career, the consulate. In the preceding year Licinius Sura had held that office for the third time, an honour vouchsafed to very few. Sura's influence must have been significant for Hadrian. It was a cruel blow when death removed so powerful a patron at an unknown date in the years immediately following Hadrian's consulate.

The career of Hadrian appears to have stopped for nearly ten years. Other promising young Romans suffered a similar retardation at about the same time. It would appear that a new political influence, opposed to Sura, Plotina, and Hadrian, dominated Trajan's court after Sura's death. It may be conjectured that Servianus played some role. Only one fact illuminates this otherwise obscure period of Hadrian's life: he held the post of archon at Athens in 112, and a surviving inscription commemorating this office was set up in the Theatre of Dionysus. Hadrian's tenure of that high Athenian office is a portent of the philhellenism that characterized his reign, and it suggests that in a time of political inactivity Hadrian devoted himself to the nation and culture of his beloved Greeks. Somehow, however, Hadrian's star rose again; and he returned to favour before the Emperor died.

One source says that Hadrian was an officer under Trajan during the Parthian wars at the end of his reign. In 117, when Trajan began his journey westward, Hadrian was left in charge of the crucial army in Syria. Friends of Hadrian, whose careers had been held up, can also be discovered in sensitive commands at the same time, probably because Plotina and her associates had regained Trajan's confidence. On August 9 Hadrian learned that Trajan had adopted him; this was the sign of succession. On the 11th, it was reported that Trajan had died on the way to Rome, whereupon the army proclaimed Hadrian emperor. The close sequence of events has always provoked suspicion of a conspiracy on Plotina's part, but the truth will never be known. Certainly, it was Trajan who had taken the fateful step of entrusting the army of Syria to Hadrian. That, at least, was not a decision made in his name by a woman and her friends.

Policies as emperor. Hadrian wrote to the Senate requesting honours for his adoptive father and ratification of the army's proclamation; all this was granted. The new emperor began a slow return to Italy. He had to make sure of the crucial provincial commands; it was also expedient to have some dissidents rounded up at home before his return and (he would be able to argue) on someone else's orders. Trajan's conquests in Armenia and Mesopotamia were quickly abandoned. The relinquishment of these territories was almost inevitable, and the process may already have begun under Trajan himself. Rome could not have managed those remote regions. Hadrian would have been pleased to recall the advice of the first Roman emperor, Caesar Augustus, not to extend the boundaries of the empire. He soon came to look upon his reign as a new Augustan age. In 123 he began to style himself Hadrianus Augustus, deliberately evoking the memory of his great predecessor; he announced a golden age on his coinage. The peace he so much cherished was a latter-day Augustan peace, and he bequeathed to posterity a public statement of his exploits that imitated the one left by Augustus.

His admiration of Augustus

Acilius Attianus, as prefect of the Praetorian Guard, directed affairs in Rome before Hadrian's return. He ordered the summary executions of four senators of exalted, consular rank: Cornelius Palma, Publilius Celsus, Avidius Nigrinus, and Lusius Quietus, all (it would seem) threats to the security of Hadrian. This bloody prelude to the new regime was unsettling, and Hadrian affirmed it was contrary to his will; he laid the blame on Attianus, just as he often blamed instructions of the dead Trajan for other unpopular acts. When Hadrian reached Rome in the summer of 118, his position was reasonably stable. He courted popular sentiment by public largesse, gladi-

atorial displays, and a formal cancellation of debts to the state. Attianus, however, was replaced; he had served his purpose. Attianus' colleague in the prefecture, Sulpicius Similis, was also dismissed. Hadrian installed as prefects the distinguished Marcius Turbo, a general to whom the Emperor owed much, and Septicius Clarus, the patron of Suetonius the biographer. Before many years had passed, both of these men had fallen into disgrace. Hadrian was mercurial or possibly just shrewdly calculating in dispensing favours.

The new emperor remained at Rome for three years. In 121 he set forth on a tour of the empire, west and east, with a view to inspecting troops and examining frontier defenses. He went to Gaul and Germany, thence to Britain in 122. From there he moved on to Spain and spent the winter in Tarraco, where he made arrangements for coping with an uprising in Mauretania (Morocco). It is not known whether he went himself to North Africa at this time. He next passed eastward, approaching Asia Minor (Anatolia) by the Aegean after an overland trip through the Balkans. He quickly negotiated some problems with the Parthians and then visited northwest Asia Minor. Returning to the west coast in 124, he sailed to Athens and finally reached Rome again in 125. This prolonged absence from the capital of the empire had its administrative justifications. There had been disturbances in some provinces, and the Parthians had to be dealt with; there was a general need for imperial supervision. Nevertheless, another motive impelled the emperor in his journeys, namely, an insatiable curiosity about everything and everybody. The Christian writer Tertullian called him rightly *omnium curiositatum explorator*, an explorer of everything interesting. That curiosity was bred of a keen intellect and an anguished spirit. These together drove him inexorably, and by a roundabout path, to the Greek East. After he left Spain early in 123, he never saw the western provinces again.

Hadrian spent another three years in Rome, but in 128 he set forth for the second time as emperor. After a visit to North Africa, he went to Athens, and from there he sailed to Asia Minor; he penetrated far eastward into Syria and Arabia. Crossing over into Egypt, he explored the Nile; then, for the third time, he went to Athens. It is not certain whether Hadrian returned to Rome in 132 or a little later; he was certainly there in May of 134, but by then a revolt in Judaea forced him abroad still another time. He went to Palestine, not as a tourist but as a commander. That journey was Hadrian's last.

Hadrian's wall

The Emperor's travels show the man better than anything else, and some of his most memorable achievements are associated with them. In northern Britain he initiated the construction of the tremendous frontier wall that bears his name from Wallsend-on-Tyne to Bowness-on-Solway. At Lambaesis, in Algeria, his rigorous inspection of the troops and his severe standards of discipline can be seen in a long inscription preserving an address he made to the soldiers in 128. In Athens, the city he loved so well, the Emperor's benefactions were numerous. At the Athenians' request, he had their laws professionally redrafted, and he brought to completion the massive temple of Olympian Zeus that the Peisistratid tyrants had begun over five centuries before. He created a federation of Greeks, based at Athens, that gave equal representation to all Greek cities. This was the Panhellenion, which thereafter played a conspicuous part in the history of Roman Greece. At the shrine of Delphi, Hadrian gave his support to a building renaissance that ushered in a new era of vitality. The impact of all this, on Hadrian personally, cannot be exaggerated. Like Augustus before him, he was initiated into the Greek mystery religion at Eleusis, and, after the temple of Olympian Zeus was dedicated, he assumed the title Olympius.

The irrational element in Hadrian was important. He was an adept in astrology, like many intelligent Romans of the time. He was also an aesthete who ascended Mt. Etna, in Sicily, and Jabal al-Ḥamrāʾ, near Syrian Antioch, simply to watch the sunrise. He had a lively sense of the past, preferring older writers to more recent ones, favouring archaism for its own sake. He revolutionized style in the empire by wearing a beard and setting a precedent for generations of emperors. The philosophic aura of the beard may have pleased Hadrian, but at the same time there is reason to think that he was concealing a facial blemish. Hadrian did not like blemishes; he adored beauty.

In the town of Bithynium-Claudiopolis (modern Bolu) in northwest Asia Minor, Hadrian encountered a languid youth, born about 110, by the name of Antinoüs. Captivated by him, Hadrian made Antinoüs his companion. When, as they journeyed together along the Nile in 130, the boy fell into the river and drowned, Hadrian was desolate and wept openly. In antiquity a report circulated and was widely believed that Antinoüs had cast himself deliberately into the river as a part of some sacred sacrifice. Although Hadrian himself denied this, the sober 3rd-century historian, Dio Cassius, thought it was the truth. The religious character, if such there was, of the relation between Hadrian and the boy is totally elusive. The emotional involvement is, however, quite clear. Seeing Hadrian's grief, the Greek world strove to provide suitable consolation for the bereaved and honour for the deceased. Cults of Antinoüs sprang up all over the East and then spread to the West. Statues of the boy became a common sight; his unheroic shape introduced a new and important element into the art of the Roman Empire. In Egypt the city of Antinoöpolis commemorated his death.

Artistic achievements. The artistic temperament of Hadrian manifested itself in his poetry, his architectural designs, his very style of life. Four complete poems of his composition survive; they illustrate an exceptional technical mastery of versification. Although the manner of expression is often artificial and the subjects are slight, Hadrian's control of difficult metres and poetic vocabulary is impressive. His most famous verses are no exception: these are the lines addressed to his soul and reportedly uttered as he lay dying. In architecture, the Emperor had a notorious quarrel with a leading contemporary architect, Apollodorus of Damascus, whom it is even alleged Hadrian had put to death. It is easy to believe at the least that the Emperor was confident of his talents and not likely to tolerate criticism. His ultimate artistic achievement was undoubtedly the villa he created for himself at Tivoli, outside Rome. Here the Emperor surrounded himself with elegant evocations of his travels; by landscaping and superior reproductions, he re-created the sights he most loved and thereby managed in his last years to experience the satisfactions of travel without ever leaving the shores of Italy.

As a creative artist himself, Hadrian was not the best of patrons. The authoritarianism displayed in the case of Apollodorus was probably not unique, and it is clear that Latin literature did not progress during his reign. The greatest Hadrianic authors, Suetonius the biographer, Juvenal the satirist, and Tacitus the historian, were all, in a sense, only survivors of the Trajanic age. All three had begun writing before Hadrian's accession and finished their work relatively early during his reign. They had no immediate literary heirs. Suetonius, although elevated to the important literary post of *ab epistulis* in the court during Hadrian's first years, was summarily dismissed about 122. An ancient source alleges excessive familiarity with Hadrian's wife, Sabina; but that would not have been something to worry the Emperor. More probably there had been a literary quarrel. The fortunes of two eminent orators, Dionysius of Miletus and Favorinus of Arelate (in Gaul), are instructive. The former Hadrian openly favoured and advanced; he then tried to overthrow him. Favorinus is found living in exile toward the end of Hadrian's reign. The Emperor's tastes dominated the world. His penchant for archaic authors and styles is mirrored in the tastes of the entire century throughout the Greek East and the Latin West. It is difficult to ascertain whether Hadrian was in this respect more a symptom or a cause. He was probably something of both.

In the city of Rome itself, during his brief sojourns there, Hadrian left his memorial in several imposing buildings. Designs for the Temple of Rome and Venus

provoked the conflict with Apollodorus. He completely rebuilt the Pantheon, which had been destroyed by fire in the reign of his predecessor. The original structure had been Augustan; therefore, a rebuilding by a man who viewed himself as a new Augustus was no accident. His own great tomb (the modern Castel Sant'Angelo) was likewise inspired by an Augustan precedent, the Julio–Claudian mausoleum, at Rome.

Last years. When Hadrian left Rome in 134 for his final journey abroad, it was to resolve a problem of serious proportions in Judaea. The troubles had already started when he was in the last stages of his previous journey. Under the leadership of Bar Kokhba (known also as Bar Koziba), the Jews were in open revolt. What had moved them is not altogether clear. Rabbinical literature alludes to a Hadrianic persecution that caused fear and apostasy. The probable explanation of this kind of reference is a universal ban on circumcision that Hadrian issued in, it seems, the early 130s. The Emperor had an abhorrence of physical mutilation and even went so far as to declare that castration was no less a crime than murder. It was in the same spirit that he denounced and forbade circumcision, which he viewed as mutilation. There is no reason to imagine that Hadrian intended by his measure to punish or provoke the Jews. His concern was ultimately humane, and he had perhaps no notion of how important the rite of circumcision was to Jewry. The uprising came swiftly and understandably. Hadrian's visit to Athens in 131–132 and his residence at Rome until the summer of 134 suggest a reluctance to deal personally with the disturbance in Judaea. He first placed an able general, Sextus Julius Severus, in charge of the problem. In the year after Hadrian's arrival in the Near East, the revolt was over. Recent discoveries have shown that several measures hitherto connected with the close of the revolt and often cited as indications of imperial severity have to be dated at least six years earlier and, very probably, well before that. Already by 129 a second garrisoning legion had been stationed in Judaea and the area transformed into a province of high rank with a former consul as governor. This disposition was clearly neither a cause nor an effect of the uprising and may well have been envisaged by Trajan. Likewise, the establishment of a Roman colony at Jerusalem under the name of Aelia Capitolina must be detached from the context of the rebellion. Hadrian meted out no savage punishments in 135.

The Jewish revolt

In 134 Hadrian's aged rival, Julius Servianus, held the consular office for the third time. This was a great but empty honour, for the man was too old. He and others may, however, have seen in his young grandson, Pedanius Fuscus, a successor to Hadrian. In 136 both Servianus and Fuscus were executed. The Emperor had realized that it was time to face the issue of succession, and he wanted it resolved in his own way. With Fuscus eliminated, Hadrian adopted the profligate Lucius Ceionius Commodus, aged about 36. The extravagant life of Ceionius, now renamed Lucius Aelius Caesar, portended a disastrous reign. Fortunately, he died two years later, and Hadrian, close to death himself, had to choose again. This time he picked an 18-year-old boy named Annius Verus, the future emperor Marcus Aurelius.

In 138 Hadrian arranged for the succession to pass to the young Verus in due course. His arrangements were clever. An estimable and mature senator, Antoninus, was adopted by Hadrian and designated to succeed him. The Emperor, however, required that Antoninus adopt both the young Verus and the eight-year-old son of the recently deceased Ceionius. Thus, the family of his first choice was remembered, whereas an early succession for the older boy seemed assured. No one expected that Antoninus would last very long. Hadrian's scheme of imposing a double adoption upon his immediate successor looks like another imitation of the first emperor, Augustus, who had made a similar demand of Tiberius. By an irony of fate, Hadrian's expectations about the future were confounded. Antoninus, like Tiberius, lived far longer than anyone would have though possible. He did not die until 161.

On July 10, 138, Hadrian died at the seaside resort of Baiae. Death came to him slowly and painfully. He wrote a letter in which he said how terrible it was to long for death and yet be unable to find it. His reign concluded two years after a double execution; it had begun with a quadruple one. The dead man was not widely mourned. Few loved him, though many feared him. He was someone to propitiate like a god, wrote a person who knew him, but he was not one to evoke affection. Since the Senate was not inclined to raise him formally to the rank of a god after his death, Antoninus had to work hard to secure a vote of deification. His exertions won him the name Pius.

Assessment. Yet Rome had reason to be grateful to Hadrian. His buildings were not his only lasting bequest. In the realm of law he had effected profound and enduring reform. The circle of the Emperor's advisers henceforth regularly included professional jurists; the praetor's edict, according to which magistrates had administered justice by varying and sometimes novel procedures, was revised and given a permanent form. The work of revision had been brilliantly carried out by the jurist Salvius Julianus under instructions from Hadrian. The humanitarianism that went so badly astray in the case of circumcision had in other areas, such as the treatment of slaves, a beneficent effect upon society. Hadrian's legal pronouncements do him credit. Whatever his weaknesses, he was no snob. One can easily believe the ancient testimony that reports his liking of the plebs.

A man of many moods and inordinate sensitivity, Hadrian was unpredictable. A friend could suddenly become an enemy; a 20-year-old Eastern youth could become a god. Hadrian seemed unable to rest physically and intellectually, even at the end when he craved to be released. Although his wife Sabina was accorded full public honour, her marriage to Hadrian was generally accounted a failure, and scandalmongers claimed that he had poisoned her when she died in 136. The Emperor composed his autobiography so as to leave posterity with his own version of the mysteries and paradoxes of his career, but posterity knew more than Hadrian wished to be known. Historians were able to counter his own statements. A definitive biography does not and cannot exist. Only the enigma and the tragedy remain.

BIBLIOGRAPHY. Ancient evidence for Hadrian and his reign may be found in the life of Hadrian in the *Historia Augusta;* CASSIUS DIO, *Roman History*, bk. 69 (surviving only in Byzantine epitomes); life of Hadrian in AURELIUS VICTOR, *On the Caesars;* and the life of Hadrian in the anonymous *Epitome De Caesaribus.* There are two modern biographies of Hadrian in English, neither of which is wholly reliable: B.W. HENDERSON, *The Life and Principate of the Emperor Hadrian, A.D. 76–138* (1923); and S. PEROWNE, *Hadrian* (1960). The article "Hadrian" in the *Oxford Classical Dictionary*, 2nd ed., pp. 484–486 (1970), contains many errors. The fictional evocation, M. YOURCENAR, *Mémoires d'Hadrien* (1951; Eng. trans., *Memoirs of Hadrian*, 1954), is, however, remarkably successful.

(G.W.Bo.)

Haeckel, Ernst

Ernst Heinrich Haeckel, a zoologist and an enthusiastic proponent of Darwin's theory of evolution—"more Darwinist than Darwin himself"—became embroiled in controversy as he popularized the theory of descent. The concept of evolution afforded him a means of constructing a family tree for all living things. The theory of descent was the basis for his monist philosophy, in which he envisioned all of nature as a unity, while rejecting design and a Creator. It was Haeckel who gave the then-recurrent recapitulation theory its fullest expression as the "fundamental biogenetic law." Freely originating new terms as he needed them, he declared that ontogeny (the embryology and development of the individual) briefly, and sometimes necessarily incompletely, recapitulated, or repeated, phylogeny (the developmental history of the species or race). From the evidence of embryology, paleontology, and comparative anatomy he undertook to depict an entire evolutionary scheme and to trace the descent of man.



Haeckel, *c.* 1870.
The Bettmann Archive

Born at Potsdam on February 16, 1834, he was the younger son of Carl and Charlotte Sethe Haeckel and grew up at Merseburg, where his father, a lawyer, was a government official. While attending the *Gymnasium* (secondary school), he read books on nature and travel and particularly enjoyed Goethe's works and Darwin's account of his voyage aboard the "Beagle." The youth collected and classified plants for his own herbarium.

After a rheumatic attack prevented him from beginning the study of botany at the University of Jena and sent him home to his parents, it was decided that he would enter medicine. He studied at Würzburg and at the University of Berlin, where his professor, the physiologist and anatomist Johannes Müller, took him on a summer expedition to observe small sea creatures off the coast of Heligoland in the North Sea.

Education

Such experiences in marine biology strongly attracted Haeckel toward zoology, but dutifully he took his medical degree, as his family wished, at Berlin in 1857. His father then agreed to his travelling to Italy, where he painted and even considered art as a career. At Messina he studied the one-celled protozoan group Radiolaria, which are strikingly crystalline in form; not surprisingly, Haeckel later maintained that the simplest organic life had originated spontaneously from inorganic matter by a sort of crystallization.

Darwin's influence

The turning point in Haeckel's thinking was his reading of Charles Darwin's 1859 work, *On the Origin of Species by Means of Natural Selection*. Meanwhile, he completed a dissertation in zoology in 1861 at Jena and became *Privatdozent*, or lecturer. In 1862 he was appointed extraordinary professor of zoology, and that year, when he published his monograph on the Radiolaria, he expressed in it his agreement with Darwin's theory of evolution; from that time he was a proponent of Darwinism and soon was lecturing to scientific and lay audiences on the descent theory, incurring both admiration and opposition. Haeckel believed the theory was implicit in Goethe's writings. Darwin had described evolution through the natural selection of accumulated favourable variations that in time formed new species; to Haeckel, however, this was only a beginning, with consequences to be pursued further. In 1865 he was appointed full professor and he remained at Jena until his retirement in 1909.

Haeckel's views on evolution

Haeckel saw evolution as the basis for a unified explanation of all nature and the rationale of a philosophical approach that denied final causes and the teleology of the church. His *Generelle Morphologie der Organismen* (1866; "General Morphology of Organisms") presented many of his evolutionary ideas, but the scientific community was little interested. He set forth his ideas in popular writings, all of which were widely read, though deplored by many of Haeckel's scientific colleagues.

Enthusiastically attempting to explain both inorganic and organic nature under the same physical laws, Haeckel portrayed the lowest creatures as mere protoplasm without nuclei; he speculated that they had arisen spontaneously through combinations of carbon, oxygen, nitrogen, hydrogen, and sulphur. In those days of great interest in protoplasm it was believed for awhile that certain deep-sea dredgings had brought up such structureless organisms; when scientists found this to be in error, Haeckel continued to insist, throughout the years, that "monera" existed. From them he traced one-celled forms with nuclei and three kingdoms—animal, vegetable, and the neutral, borderline "protista." His artistic leanings toward ideal symmetries led him to outline numerous geneological trees, sometimes to supply missing links or branches; and he reconstructed the human ancestral tree, to show man's descent from the lower animals.

Haeckel tended to speculate and for some years pondered over the problem of heredity. Interestingly, though it was only on a theoretical basis, he suggested as early as 1866 that the cell nucleus was concerned with inheritance. He had long been thinking of "vital molecular movement" when, in 1876, he attempted to place heredity on a molecular basis in a work entitled *Die Perigenesis der Plastidule* ("The Generation of Waves in the Small Vital Particles"). Here again, he traced a branching scheme, this time to illustrate the mechanism of heredity and to show the influence of outer conditions on the inherited undulatory motion he attributed to the "plastidules," the term he adopted for the molecules making up protoplasm.

Though his concepts of recapitulation were in error, Haeckel brought attention to important biological questions. His gastraea theory, tracing all multicellular animals to a hypothetical two-layered ancestor (which he even drew in section), stimulated both discussion and investigation. His propensities to systematization along evolutionary lines led to his valuable contributions to the knowledge of such invertebrates as medusae, radiolarians, siphonophores, and calcareous sponges.

Haeckel was the teacher of noted zoologists and a member of learned societies. He was not one to change his ideas and was often involved in controversy. While extending his arguments into the theological, political, and social fields, he attracted disciples, yet alienated friends. Building the collections around his own, Haeckel founded both the Phyletic Museum in Jena and the Ernst Haeckel Haus; the latter contains his books and archives and in it many mementos of his life and work are preserved. Haeckel died at Jena on August 9, 1919.

BIBLIOGRAPHY. Biographical references are GEORG USCHMANN'S article on Haeckel in the *Dictionary of Scientific Biography*, vol. 6, pp. 6–11 (1972), with bibliography; and his *Geschichte der Zoologie und der zoologischen Anstalten in Jena 1779–1919* (1959), describing Haeckel's career and influence at Jena, with a comprehensive list of source material. WILHELM BOLSCHE, *Ernst Haeckel: Ein Lebensbild* (1900; Eng. trans., *Haeckel: His Life and Work*, 1906); JOHANNES HEMLEBEN, *Ernst Haeckel, in Selbstzeugnissen und Bilddokumenten* (1964); PETER KELMM, *Ernst Haeckel, der Ketzer von Jena* (1966); and GERHARD HEBERER (ed.), *Der gerechtfertigte Haeckel* (1968), provide further data on his life and writings. Discussions of Haeckel's thought may be found in DANIEL GASMAN, *The Scientific Origins of National Socialism: Social Darwinism in Ernst Haeckel and the German Monist League* (1971), which describes Haeckel's influence on German social and political thought under National Socialism; and P. SMIT, "Ernst Haeckel and His 'Generelle Morphologie': An Evaluation," *Janus*, 54:236–252 (1967), which reviews his work and the impact of his theories. ERIK NORDENSKIOLD in *Biologins Historia*, 3 vol. (1920–24; Eng. trans., *The History of Biology*, 1928); and EMANUEL RADL in *Geschichte der biologischen Theorien seit dem Ende des siebzehnten Jahrhunderts*, 2 vol. (1905–09; Eng. trans., *The History of Biological Theories*, 1930), present general introductions to Haeckel. SIR GAVIN DE BEER in *Embryos and Ancestors* (1954), discusses the recapitulation theory.

(G.Ro.)

Hague, The

The Hague (Dutch 's-Gravenhage, usually abbreviated Den Haag; French La Haye) is the seat of the government of The Netherlands and the capital of the province of South Holland. The third largest city in the country with a population in 1971 of 537,643, The Hague lies on a sandy plain about four miles from the North Sea. Its fine parks, its nearness to the beach and the sea, and its easy connections with the principal cities of the country make the city a convenient capital, a pleasant residential area, and a popular centre of international conferences.

History. The city's name recalls the hunting lodge and principal residency of the counts of Holland, located in a woodland area called Haghe, or "hedge" (whence 's-Gravenhage, "the counts' private dominion"). Count William II built a castle there in 1248, around which several buildings—including the famous Knight's Hall (1280)—came to be clustered. Together they form the Binnenhof ("inner courtyard") in the very heart of the city. Originally this group of buildings was surrounded by a double canal. Around 1350 an artificial lake, the Hofvijver, was dug on the north side and still forms one of the many attractions of the city. On the west side is the Buitenhof ("outer courtyard"), where once stood the workshops and stables of the court. The main gateway of the courtyard, *gevangenpoort* ("prison gate"), once used as a prison, formed the entrance. These courtyards, together with areas north and east of it, formed a complex under the Count's immediate rule and jurisdiction. He rented part of the ground to the nobles who formed his council and to high court dignitaries, who built their mansions there, giving rise to a spacious and well-laid-out distinctive district, still seen today in the Plaats, Kneuterdijk, Voorhout, and Vijverberg.

In the 14th century the counts connected the Binnenhof with the Dutch waterways by means of an entrance canal known as the Spui.

Growth of the village west of the court

West of the court area, a small settlement of farmers and craftsmen grew up to service the court, gradually forming a village, the boundaries of which were officially determined in 1370. A flourishing wool industry grew up within the village, the rapid growth of which was reflected in the construction in 1399 of the Great Church of St. James or St. Jacobs to replace the wooden parish church. The village became the centre of a larger rural district known as the Haagambacht. The governing body (Magistraat) resided in the Town Hall opposite the Great Church. Until 1795 the court area and the village comprised two separate sections, each with its separate administration and jurisdiction.

In 1436 the earldom of Holland passed into the possession of the dukes of Burgundy. They charged a *stadhouder* (lord lieutenant) with the government of the earldom, assisted by a court composed of professional lawyers. These were mostly foreigners who, in establishing themselves in the surrounding court area, proved of great financial and cultural importance for The Hague; it was probably through their influence that a new Flemish Renaissance Town Hall (Stadhuis) was built in 1565.

Because of its open location, the borough suffered much in the beginning of the Dutch war of liberation from Spain (1568–1648). But in 1585, when much of the danger had passed, the States-General, the governing body of the seven United Provinces of the Netherlands, along with other bodies of the central government, established themselves in the Binnenhof. The government of the province of Holland also returned, and *stadhouder* Prince Maurice of Orange took up residence here. It was at his instance that in 1616 The Hague was defended by a web of canals that remained the city's borders to the mid-19th century.

In the 17th century, when the Dutch Republic played a leading role in Europe, The Hague became an international city, a centre of diplomacy, and a refuge for persecuted minorities such as the French Protestants and the Portuguese Jews. A mixture of nationalities lived within the city, and all of them helped to fashion its social and cultural life.

After 1630 there was an important expansion both in the court area and in the newly constructed Prinsegracht region in the village. Many beautiful houses were built in the new French-Italian style by the architects Jacob van Campen and Pieter Post. The finest surviving examples of their work are the Noordeinde Palace (now the International Institute of Social Studies) and the Mauritshuis (now the Royal Art Gallery) on the Plein.

Canals and small harbours were constructed around the Spui, encouraging small trades and industries. The first Protestant New Church, established in 1654, contains the

Charles E. Rotkin—Photography for Industry



The old castle in the heart of The Hague. The Ridderzaal or Knights' Hall stands in the middle of the Binnenhof or Inner Court. The Binnenhof is surrounded by the Buitenhof or Outer Court of buildings. The Hofvijver pond lies to the north.

tombs of the Dutch statesmen Jacob and Johan de Witt and of the philosopher Spinoza.

Building continued in the 18th century under such capable architects as the French refugee Daniel Marot and Pieter de Swart. The medieval houses of the nobles, already partly replaced by 17th-century patrician mansions, were then reconstructed in the French style.

By about 1780 The Hague's silver and porcelain were famous. Printing and publishing prospered, and many books proscribed abroad were produced in the city.

Decline under French rule

During French rule (1795–1813), The Hague declined into a poor provincial town, but with liberation from the French it once again became the residence of the monarch and the administrative centre, now of the kingdom of The Netherlands. Immediate revival, however, was impeded by a lack of money and the narrow-mindedness of the city magistrates. It was not until after 1850, when the revenues from the Dutch East Indies started to pour in, that any real signs of improvement were seen. The Willemspark (a residential quarter, or "villa park," dominated by greenery and influenced by the architecture of London squares of the period) was then constructed, and the city started to expand outside its borders of 1616. Cheap housing was constructed for a rapidly growing middle class. Netherlanders repatriating from the Dutch East Indies showed preference for living in The Hague, thus giving the population a markedly "Indonesian" aspect still found today.

After 1870, prosperity increased, and a more progressive city administration provided modern public facilities. Many new shops were built in addition to better housing areas. Between 1880 and 1890, the population jumped from under 114,000 to 200,000. The arts flourished, and The Hague school of landscape painters became world famous. As a result of the Hague Peace Conferences held in The Hague in 1899 and 1907, The Hague became a permanent centre of international law. The American steel magnate Andrew Carnegie endowed the Peace Palace, completed in 1913.

The 20th century saw real growth and progress: industry increased, harbours were constructed, and oil and insurance companies set up their head offices in the city. Large office buildings were designed by architects such as H.P. Berlage (1856–1934), who also designed the Haggs Gemeentemuseum, the municipal museum.

During World War II the city suffered heavily from the German occupation, but afterward there was a remarkable tide of reconstruction and expansion.

Among the more interesting structures of the postwar period are the unusual United States Embassy designed by Marcel Breuer (1961) and The Netherlands Congress Centre designed by J.J.P. Oud (1969).

The contemporary city. *Physical layout and boundaries.* The Hague is situated along the shore of the North Sea: 4° east of Greenwich, at 52° north latitude and about 15 feet above sea level. Its northern climate is tempered by a branch of the Gulf Stream, resulting in an average temperature of 50° F—several degrees higher than elsewhere at the same latitude. The Hague is protected from the sea by a row of wide dunes broken only by the Scheveningen fishing harbour and the beach promenade about 2 miles long. To the south of the city is Westland, the vegetable garden of western Europe, and, to the east, the polders of The Netherlands low-middle region, where the land often lies below sea level. It is in this polder land that Zoetermeer, 10 miles east, has become a satellite town expected to grow in the next decade from a quiet township of about 20,000 inhabitants to a town accommodating about 100,000. North of The Hague is the municipality of Wassenaar, one of the richest municipalities in The Netherlands and the residence of cabinet ministers, ambassadors, and wealthier businessmen. In the direction of Haarlem, less than seven miles from The Hague centre, begins the bulb-growing area where the crocuses, daffodils, hyacinths, and tulips attract hundreds of thousands of visitors every spring.

Demography and economic life. In 1971 The Hague's population was 537,643, and that of its metropolitan area was 710,528. This included a large number of older people: nearly 41 percent of the population were above 45 years of age. In religion, 29.3 percent were Roman Catholic, 33.5 percent Protestant, and 5.6 percent other; 31.6 percent reported no religious affiliation.

There is little heavy industry in The Hague. Most of the city's 26,000 business firms are engaged in trade, banking, insurance, or services. The hosting of international conferences is big business in The Hague. The Netherlands Congress Centre, which is able to accommodate 9,000 conference delegates a day, was completed in 1969. In the same year there were over a half million overnight stays in the hotels by over 187,000 guests. Several large oil firms have their international headquarters in the city.

Political and governmental institutions. The Hague is the seat of The Netherlands government and Parliament. This means that practically all of the ministerial departments are established here.

Most of them are in or very close to the historic city centre, but a few are in the suburbs, such as Rijswijk (Ministry of Culture, Recreation, and Social Works) and Leidschendam (Ministry of Social Affairs). Many of the political parties also have their headquarters in The Hague. And because The Hague is the city of the government, it also accommodates many government and semi-government institutions such as industrial organizations, waterways departments, and advisory councils.

The Hague is also the capital of South Holland, one of the 11 provinces into which The Netherlands is divided, and the seat of the High Council of The Netherlands and of the Supreme Court of The Netherlands. It also houses the International Court of Justice, the Institute of Social Studies, and 61 embassies.

Transportation services. The Hague is situated in the heart of the city belt or agglomeration known as the Randstad Holland, or Rim-City Holland. It is less than an hour's drive (35 miles) to Amsterdam and the international Schiphol Airport, while Rotterdam, with its own Zestienhoven Airport, is less than a half hour (10 miles) away. The Hague has direct train connections with the German Ruhr area via Rotterdam, while Brussels is reached by electric train in two hours and Paris in five hours. From The Hague to the Belgian border by road takes about an hour and a half, as does the trip to the German frontier.

Since 1968 there has been ferry service from Scheveningen to the English coastal town of Great Yarmouth, while from the nearby Hook of Holland there is a fine boat service to the English steamer harbour of Harwich.

Cultural life. The Hague's 14 museums comprise a wide range of collections. The Mauritshuis Royal Art Gallery has a remarkable collection of the works of the Dutch masters: Rembrandt, Vermeer, Jan Steen, and others. The Haags Gemeentemuseum has the largest collection of paintings by Piet Mondrian in the world as well as extensive collections of modern art, old handicrafts, musical instruments, and prints.

Other notable museums are The Netherlands Costume Museum, the Mesdag Museum, the Mesdag Panorama, the Coin and Medal Cabinet, and the Postal Museum. In addition, about 25 galleries display various aspects of modern art.

Art and music

The Hague has two art academies: the Koninklijke Akademie van Beeldende Kunsten (Royal Academy for the Plastic Arts) and the Vrije Akademie (the Free Academy). The Koninklijk Conservatorium voor Muziek (Royal Conservatory of Music) graduates dozens of musicians annually. Musical life in The Hague is largely dominated and determined by the resident orchestra, The Hague Philharmonic, whose main auditorium since 1969 has been The Netherlands Congress Centre. Theatre life is centred mainly in the Koninklijke Schouwburg (Royal Theatre) for which de Haagse Comedie (the Hague Comedy Group) is the resident company. The avant-garde section of this company has been giving performances since 1969 in the HOT (Haags Ontmoetingscentrum voor Toneel), a converted church in the centre of the city. A third local company, the Nieuwe Komedie, specializes in modern plays.

The city also houses the Netherlands Dance Theatre, specializing in modern ballet and well known throughout Europe.

Recreation. Of The Hague's 16,140 acres, 1,820 consist of parks, woods, dunes, large gardens, playing fields, and sports grounds—an average of over 144 square feet of green for every citizen, more than any other large city in western Netherlands.

There are 16 large parks and a beach about five miles long, which at low tide is dozens of yards wide and provides pleasure and recreation for hundreds of thousands every year. The city has 773 acres of sports fields; several swimming pools; three sports halls; 16 tennis parks; one racecourse (Duindigt); and the largest camping park in western Europe, with 8,500 camping places.

BIBLIOGRAPHY. JACOB DE RIEMER, *Beschrijving van 's-Gravenhage*, 2 vol. (1730–39), is the most reliable of a number of early descriptions of The Hague and its history; H.E. VAN GELDER, *Zeven eeuwen Den Haag* (1937), a modern general view of the history of The Hague, published at the town's 700th anniversary; C. DE WIT, *Den Haag vroeger en nu* (1968), is the most recent general view of the history of The Hague, more condensed than that by Van Gelder but with more emphasis on the foundation of the initial Castle and its first development; *Haagsch Jaarboekje* (1889–1899); and its successor, *Die Haghe* (annual), published by the association Die Haghe, contains a large number of good articles on a variety of subjects; *Statistical Yearbook* (annual), offers a wealth of data; *The Hague, Swinging Between Tranquility, Dynamic Action and the Arts* and *The Hague, the Arts*, both published by the Public Relations Department City of the Hague, offer general information about The Hague.

(M.v.H./Ja.B.)

Hahn, Otto

The discovery of nuclear fission, at the end of 1938, by the German chemist Otto Hahn, working with the radiochemist Fritz Strassmann, was one of the most pregnant events of modern times. Hahn, the senior of the two, was awarded the 1944 Nobel Prize for Chemistry in recognition of this discovery. Already highly respected for his previous work, he was widely known as a brilliant experimentalist.

Landesbildstelle Berlin



Hahn.

Hahn was born on March 8, 1879, in Frankfurt-am-Main, the son of a glazier. Although his parents wanted him to become an architect, he eventually decided to study chemistry at the University of Marburg. Life at home was simple and abstemious, but at Marburg Hahn developed a taste for beer and cigars, and his student days were not without romantic episodes. He worked hard at chemistry, though he was inclined to absent himself from physics and mathematics lectures in favour of art and philosophy, and he obtained his doctorate in 1901. After a year of military service, he returned to the university as chemistry lecture assistant, hoping to find a post in industry later on.

In 1904 he went to London, primarily to learn English, and worked at University College with Sir William Ramsay, who was interested in radioactivity. While working on a crude radium preparation that Ramsay had given to him to purify, Hahn showed that a new radioactive substance, which he called radiothorium, was present. Fired by this early success and encouraged by Ramsay, who thought highly of him, he decided to continue with research on radioactivity rather than go into industry. With Ramsay's support he obtained a post at the University of Berlin. Before taking it up, he decided to spend several months in Montreal with Ernest Rutherford (later Lord Rutherford of Nelson) to obtain further experience of radioactivity. Shortly after returning to Germany in 1906, Hahn was joined by Lise Meitner, an Austrian physicist, and five years later they moved to the new Kaiser Wilhelm Institute for Chemistry. There Hahn became head of a small but independent department of radiochemistry.

Early work on radioactivity

Feeling that his future was more secure, Hahn married Edith Junghans, the daughter of the chairman of Stettin City Council, in 1913; but World War I broke out the next year, and Hahn was posted to a regiment. In 1915 he became a chemical warfare specialist, serving on all the European fronts.

After the war, Hahn and Miss Meitner announced the discovery of a new radioactive element, protactinium. Since nearly all of the natural radioactive elements had then been discovered, he devoted the next 12 years to studies on the application of radioactive methods to chemical problems.

In 1934 Hahn became keenly interested in the work of the Italian physicist Enrico Fermi, who found that when the heaviest natural element, uranium, is bombarded by neutral subatomic particles called neutrons, several radioactive products are formed. Fermi supposed these products to be artificial elements similar to uranium. Hahn and Miss Meitner, assisted by the young Strassmann, obtained results that at first seemed in accord with Fermi's interpretation but which became increasingly difficult to understand. Miss Meitner fled from Germany in July 1938 to escape the persecution of Jews by the Nazis, but Hahn and Strassmann continued the work. By the end of 1938, they obtained conclusive evidence, contrary to previous expectation, that one of the products from uranium was a radioactive form of the much lighter element barium, indicating that the uranium atom had split into two lighter atoms. Hahn sent an account of the work to Lise Meitner, who, in cooperation with her nephew Otto Frisch, formulated a plausible explanation of the process, to which they gave the name nuclear fission.

Discovery of nuclear fission

The tremendous implications of this discovery were realized by scientists before the outbreak of World War II, and a group was formed in Germany to study possible military developments. Much to Hahn's relief, he was allowed to continue with his own researches. After the war, he and other German nuclear scientists were taken to England, where he was profoundly affected by the announcement of the explosion of the atomic bomb at Hiroshima in 1945. Although now aged 66, he was still a vigorous man; a lifelong mountaineer, he maintained physical fitness during the enforced stay in England by a daily run.

On his return to Germany he was elected president of the former Kaiser Wilhelm Society, renamed the Max Planck Society for the Advancement of Science; and became a respected public figure; a spokesman for science; and a friend of Theodor Heuss, the first president of the Federal Republic of Germany. He campaigned against further development and testing of nuclear weapons. Honours came to him from all sides; in 1966 he, Meitner, and Strassmann shared the prestigious Enrico Fermi Award. This period of his life was saddened, however, by the loss of his only son Hanno and his daughter-in-law, who were killed in an automobile accident in 1960. His wife never recovered from the shock. Hahn died in Göttingen on July 28, 1968, after a fall; his wife survived him by only two weeks.

BIBLIOGRAPHY. OTTO HAHN, *Mein Leben* (1968; Eng. trans., *My Life*, 1970), provides information on personal experiences and views. His scientific work and views to 1945

are treated in *Vom Radiothor zur Uranspaltung* (1962; Eng. trans., *A Scientific Autobiography*, 1966), including three of Hahn's scientific papers on nuclear fission. R. SPENCE, "Otto Hahn," *Biogr. Mem. Fellows R. Soc.*, 16:279–313 (1970), describes Hahn's life and work and gives a complete list of his publications.

(R.Sp.)

Haiti

Haiti (République d'Haïti) is a republic in the West Indies, situated in the western part of the Caribbean island of Hispaniola. With an area of 10,714 square miles (27,750 square kilometres), it occupies slightly more than one third of the island. Since much of its territory consists of two peninsulas, it has a coastline that is long in proportion to its size—amounting to 672 miles. It is bounded to the east by the Dominican Republic (which occupies the remainder of the island), to the north by the Atlantic Ocean, and to the west and south by the Caribbean Sea. Its population numbers approximately 4,250,000. The capital is Port-au-Prince.

Haiti, which won its independence from France in 1804, was the first country in the Americas, after the United States, to win freedom from colonial rule. From 1957 to 1971 its president was François Duvalier, popularly known as Papa Doc; Duvalier established a personal style of government. The great majority of the people are Negro; about 5 percent, however, consists of a mulatto minority who form an élite that has played a leading role in Haiti's history; many members of this élite are now living abroad. The official language is French, but a Creole patois is more generally spoken. Coffee is the principal export, and the United States is Haiti's principal trading partner.

Strategically, Haiti commands the Windward Passage, a 50-mile-wide corridor running between the northwest extremity of Haiti and the eastern extremity of Cuba, which is situated midway on the sea-lane between New York and the Panama Canal. Important air routes between North and South America pass across Haiti, but without stopping there. (For coverage of associated physical features, see ATLANTIC OCEAN; CARIBBEAN SEA; for historical aspects see HAITI, HISTORY OF.)

The landscape. *Relief.* Haitian territory consists essentially of two peninsulas, both of which project westwards —the northwestern peninsula and the southern peninsula. They are separated by the Golfe de la Gonâve, and are joined together by the alluvial plain of Artibonite, which covers about 310 square miles, as well as by the Plateau Central, which covers 840 square miles. The backbones of the two peninsulas are formed by mountain ranges of white calcareous (chalky) rock. The northern peninsula trends from northwest to southeast, while the southern peninsula trends east to west. The mountains of the southern peninsula are higher, reaching an altitude of 8,772 feet about 15 miles west of the Dominican frontier. Short rivers flow rapidly down to the sandy shore, carving gullies in the hillslopes, and forming small coastal plains on which local administrative and trade centres have been established. Viewed from the sea, therefore, the prospect consists of succeeding valleys and hills. Most valleys are quite small, as are the towns located in them. The principal exceptions to this general pattern are the Plaine du Nord (the Northern Plain), which extends over almost 150 square miles, and the two plains—Cul-de-Sac (140 square miles) and Léogâne (40 square miles)—which unite in the region of the capital, Port-au-Prince. The plain of Cul-de-Sac extends into the interior along an east-west rift that divides the island into two distinct parts and which is subject to frequent but minor earthquakes. Two lakes, the Étang Saumâtre (literally "brackish pond"), with an area of 65 square miles, and the Étang de Miragoâne, with an area of about 10 square miles, are also situated in this rift near the Dominican frontier.

The principal mountain ranges are the Massif du Nord on the northern peninsula, the Massif du Sud (Massif de la Hotte) on the southern peninsula, the Massif de la Selle in the southeast, and the Chaîne des Montagnes Noires in the central part of the country. The principal river is the Artibonite, which rises in the mountains of the Dominican Republic and flows westward for about 174 miles to drain into the Golfe de la Gonâve. To the north of the Artibonite is the country's second river, L'Estere, which is about 28 miles long; both rivers are partly navigable.

Offshore islands

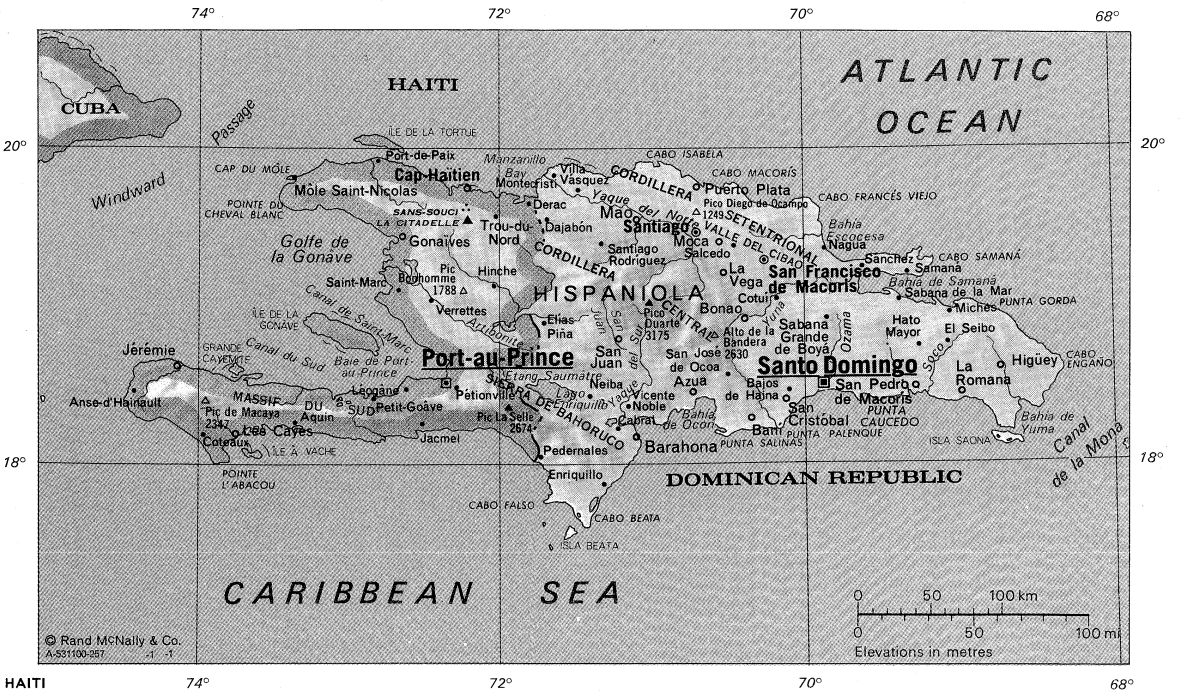
Haiti has several offshore islands, the largest of which is the island of La Gonâve, about 40 miles long and with an area of 254 square miles; it lies about 40 miles northwest of Port-au-Prince. Off the northern coast lies the small island of Tortuga ("Turtle Island")—once the haunt of French buccaneers who, in the 17th century, first colonized the western part of Hispaniola; it has an area of 69 square miles. Other small islands are Grande Cayemite, situated off the northern coast of the southern peninsula, and Île à Vache, which lies off the southern coast.

Climate. Haiti lies near the northern limit of the tropical zone. The climate is modified by the proximity of the sea as well as by the mountainous terrain. Temperatures vary little throughout the year; they remain constantly high near sea level, but are lower at the higher altitudes inland. Average temperatures range from 75° F (24° C) in January and February to 83° F (28 °C) in July and August. Port-au-Prince, at sea level, has an average annual temperature of 80° F (27° C). Inland, frosts occur from time to time at high altitudes, such as Pic la Selle (8,793 feet). There is considerable variation in the amount of rainfall experienced at different locations, as the mountain ranges intercept low-flying clouds blown by the northeast trade winds in winter, and by eastern and southeastern winds in summer. Seasonal rainfall is highest in the south from May to November, and in the north from December to April. The average annual rainfall varies greatly in different locations: at Môle Saint-Nicolas in the northwest it is only about 20 inches, while some highland areas receive over 100 inches. Hurricanes occasionally strike the island, travelling from south to north; they may occur between the months of August and November.

Vegetation and animal life. Timber was once so abundant that mahogany and guayacan (a certain type of lignum vitae, the hardwood) were exported in quantity. The demand for cultivable land created by an increasing population, however, led to deforestation, so that today only a few pine forests survive at the higher altitudes. Stands of mahogany, rosewood, and cedar may still nevertheless be found. An alpine type of vegetation occurs at altitudes of more than 4,800 feet, where the climate is misty, cold, and damp. Elsewhere, many tropical and semi-tropical plants occur, with coconut palms and trees bearing such tropical fruits as avocados, mangoes, limes, and oranges growing wild. Mangrove swamps are often found bordering tidal mud flats on the coast.

So far as is known, animal life never included poisonous snakes or mammals large enough to be dangerous to man. There are a number of rodents, and insects abound. Reptile life includes crocodiles, iguana, and lizards. There are over 200 species of birds, including ducks and guinea hens, and—on the lakes—flamingos. There are about 270 species of fish in the coastal waters, including tarpon, kingfish, barracuda, and red snapper.

Traditional regions. Five traditional regions may be recognized; they correspond closely to the five *départements* into which the country has been divided—the North, the Northwest, the Artibonite, the West and the South. The North—in reality located in the northeast—is a well-watered plain, suitable for the cultivation of tropical crops, including sisal in its northeastern corner, which receives somewhat less rainfall than the remainder. The Northwest is, on the contrary, an arid area that becomes desert in its southern half; the island of Gonâve may be grouped with this arid region. In the centre of the country the region of the Artibonite River basin includes both a lowland area and the Central Plateau; rainfall is not abundant, but the Artibonite River itself, which flows at the rate of almost 3,000 cubic feet per second, offers possibilities of hydroelectric development. The West (in effect the southeast), where the cap-



HAITI

MAP INDEX

Cities and towns

Anse-d'Hainault . 18·30n 74·27w
Aquin . . . . . 18·16n 73·24w
Cap-Haïtien . . . . . 19·45n 72·15w
Coteaux . . . . . 18·12n 74·02w
Dérac . . . . . 19·41n 71·48w
Gonaïves . . . . . 19·30n 72·40w
Hinche . . . . . 19·09n 72·01w
Jacmel . . . . . 18·14n 72·32w
Jérémie . . . . . 18·39n 74·08w
Léogâne . . . . . 18·31n 72·38w
Les Cayes . . . . . 18·12n 73·45w
Môle Saint-Nicolas . . . . . 19·48n 73·23w
Pétionville . . . . . 18·31n 72·17w
Petit-Goâve . . . . . 18·26n 72·52w
Port-au-Prince . . . 18·32n 72·20w
Port-de-Paix . . . . . 19·57n 72·50w
Saint-Marc . . . . . 19·07n 72·42w
Trou du Nord . . . . 19·38n 72·01w
Verrettes . . . . . 19·04n 72·27w

Physical features and points of interest

Artibonite, *river* . . . . . 19·15n 72·46w
Bahoruco, Sierra de, *mountains* . . . 18·20n 72·00w
Bonhomme, Pic, *peak* . . . . . 19·05n 72·15w
Caribbean Sea . . . 17·00n 73·00w
Cheval Blanc, Pointe du, *point* . . . . . 19·41n 73·27w
Gonâve, Golfe de la, *gulf* . . . . . 19·20n 73·15w
Gonâve, Île de la, *island* . . . . . 18·45n 73·00w
Grande Cayemite, *island* . . . . . 18·35n 73·45w
Hispaniola, *island* . . . . . 19·00n 72·30w
Hotte, Massif de la, see Sud, Massif du
L'Abacou, Pointe, *point* . . . . . 18·01n 73·47w
La Citadelle, *historic site* . . . . . 19·35n 72·13w
La Selle, Pic, *peak* . . . . . 18·22n 72·00w
Macaya, Pic de, *peak* . . . . . 18·25n 74·00w
Manzanillo Bay . . . . . 19·40n 71·45w
Môle, Cap du, *cape* . . . . . 19·50n 73·25w
Port-au-Prince, Baie de, *bay* . . . . 18·40n 72·30w
Saint-Marc, Canal de, *channel* . . . . . 18·45n 72·40w
Sans-Souci, ruins . . . . . 19·37n 72·12w
Saumâtre, Étang, *lake* . . . . . 18·35n 72·00w
Sud, Canal du, *channel* . . . . . 18·35n 73·00w
Sud, Massif du, *mountains* . . . . . 18·25n 73·55w
Tortue, Île de la (Tortuga), *island* . . . . . 20·01n 72·50w
Vache, Île à, island . . . . . 18·05n 73·38w
Windward Passage . . . . . 20·00n 74·00w

ital is located, is where geographic contrasts are the most noticeable; here the people of the plains depend for their subsistence on the waters that flow down from the Massif de la Selle. The South (in effect the southwest) consists of fertile valleys tucked between mountain ranges that render transport unusually difficult; in addition the region is exposed to hurricanes from the south.

Human settlement. Human settlement almost everywhere has modified the natural landscape. The countryside is settled not by villages but by scattered extended-family settlements, which take the form of small groups of dwellings sheltered from passers-by and from the winds by groves of banana and coffee trees. The small wooden-frame houses thus protected vary in design according to the region, but are everywhere enclosed within a compound of four mud-daubed wattle walls. The roof is either thatched or—when the inhabitants are prosperous—is made of corrugated-iron sheets. Some peasants achieve considerable wealth, but seldom use it to further improve their housing; the furniture inside most houses remains restricted to necessities, with tools such as hoes and machetes taking pride of place. Wealth is invested in land and cattle, as well as in lavish entertainment that all who can are expected to provide free for all. The traveller in the countryside often meets small groups of men engaged in collective work to the sound of drums and other instruments; files of women bringing produce to market are also a common sight. On Saturday night dance drums can be heard everywhere.

Family settlements

Urban growth has played a lesser part than rural population growth in changing the Haitian landscape. About 420,000 people live in Port-au-Prince, and a further 30,000 in its suburbs, but after this the principal towns are Cap-Haïtien (population 44,000), Gonaïves (29,000), Les Cayes (23,000), Jérémie (17,000), and Port-de-Paix (14,000)—fewer than 600,000 town dwellers all told. Thus out of a population of nearly 4,250,000, only about 13.6 percent is urban. The official urban figure of 18 percent includes residents of about 100 commune (local administrative district) headquarters, most of whom are farmers. Town growth, moreover, is virtually restricted to Port-au-Prince, with its Pétionville suburb, and Cap-Haïtien.

Port-au-Prince was founded in 1749. It was laid out on a gridiron plan in an almost unbelievably beautiful setting of green and red mountains that fall abruptly down to the blue waters of the Golfe de la Gonâve. The city centre has uninterrupted rows of stone houses, and several buildings that have arcades. The streets are straight and still broad enough for the existing traffic. The poorest inhabitants—constituting more than half the city's population—live in wooden frame houses that cover the hills to the north and the low ground to the south. Density reaches a peak of 284 to the acre around the Catholic cathedral.

The route to the mountains is via the Champ de Mars, the old colonial drilling ground and promenade, now turned into public gardens; it is dominated by the national palace—Haiti's White House. Relatively well-kept houses stand to the north and south of it; they are occupied by skilled manual workers or junior office employees whose children, on Sundays, crowd the streets dressed in school uniforms. Farther east are residential areas standing on slopes that become progressively steeper and are covered with thick garden vegetation from which royal palm trees emerge with dignity. The density in this area is 40 persons per acre, or even less, as houses grow fewer in number but larger in size and modern in style. Stupendous views of Baie de Port-au-Prince are seen from these slopes, particularly at sunset.

Cap-Haïtien

Cap-Haïtien has virtually remained within the limits of its gridiron plan established in the French colonial era. It differs in appearance from the capital city because fear of fires once made its governor decide that all its houses should be built of stone. Today it remains an old-fashioned agglomeration of storied and balconied stone houses.

People and population. Haiti's population was about 3,000,000 in 1950, was estimated at about 4,000,000 in 1962, but was found to have increased to only 4,244,000 at the 1971 census.

The official language is French, but workers and peasants speak Creole—a language in which many West African elements are discernible, although its vocabulary is of predominantly French origin. The bulk of population originated in the 480,000 slaves who won their freedom in a struggle that ended with Haiti's independence in 1804. Most of the Africans brought to the island before the middle of the 18th century came from the Dahomeyan town of Ouidah, in West Africa, but later on the French slave traders mostly called at Loango in the Congo. Mulattoes, who often had enjoyed the privilege of some education, also formed part of the new nation, together with a few Europeans. Since 1804 small waves of immigration have taken place—the immigrants consisting mostly of traders and mechanics, together with a few teachers and priests; they came at first from the British Isles and later from the United States, forming a part of the same repatriation movement that elsewhere gave birth to Liberia. A concordat with the pope, signed in 1860, gave Haiti an all-French clergy and helped to attract more French-speaking residents, some of whom came from France, some from Guadeloupe, and some from Martinique. Members of these three foreign communities today hardly number more than 1,000 and almost all of them are of mixed foreign and Haitian parentage. Arab traders came in the latter years of the 19th century, followed by a few Germans and Italians. The U.S. occupation, which lasted from 1915 to 1934, led to the establishment of a number of enterprises, both public and private, manned by personnel who, in contrast to earlier immigrants, migrated with their families.

Almost all Haitians are baptized as Roman Catholics, and the clergy has always consisted of ordinary priests rather than missionaries. Protocols with the pope, signed in 1966, enabled the president to appoint an archbishop, two bishops, and three auxiliary bishops of Haitian descent. There are about 160 Haitian priests and the three bishoprics still held by the French are likely to have Haitian incumbents before long. British Methodists introduced Protestantism as early as 1816, but no mass conversion movement was launched before the advent of American revivalist missionaries in the 1940s. Even so, Protestants remain few, and to place their numbers at 10 percent would seem an exaggeration, although they stand high in prestige in both rural and urban communities.

The realities of religion in Haiti, however, require a look in another direction. As the country was left without a regular clergy until as late as 1860, the peasants built up Voodoo (*vaudou*), a syncretic cult in which the Catholic god rules over an African pantheon—a far cry indeed from the voodoo of New Orleans.

The national economy. Although Haiti, like its neighbour the Dominican Republic, has a diversity of resources unknown to other West Indian islands, these resources are too limited in quantity for commercial exploitation.

Table 1: Town Growth in Haiti

	1950 census	1971 census
Port-au-Prince	134,600	420,000
Suburbs	30,000	30,000
Cap-Haïtien	24,200	44,000
Gonaïves	13,600	29,000
Les Cayes	11,600	23,000
Jérémie	11,000	17,000
Port-de-Paix	6,400	14,000

Known mineral resources are restricted to alluvial gold (discovered on the Haitian-Dominican borders as early as the 16th century), copper, bauxite (discovered comparatively recently), and small quantities of silver, lignite, and manganese.

Population pressure on the land

Vegetation and other biological resources are unusually varied in species, but have suffered as a result of overpopulation, which even at times led hungry peasants to uproot coffee trees in order to plant quick-growing foodstuffs instead. Soil erosion has resulted from such actions, and soil conservation by terracing has only occurred sporadically in recent decades. Educational work is needed before improved methods of cattle-breeding and fishing can be introduced.

Prospects for the use of hydroelectric power are limited almost everywhere except in the Artibonite River Valley, which has a potential for large-scale development.

More than 85 percent of the economically active population is employed in agriculture and fishing. Coffee, by far the most important commercial crop, is cultivated on the slopes of almost all mountains, in the shade of fruit and other trees. Haitian coffee is of an excellent mild type, but production varies considerably from year to year. Sugarcane is grown on all the plains, but depends for success on irrigation. Most of the crop is processed at a mill located near Port-au-Prince. The northeast is favourable for the cultivation of sisal (about 27,000 metric tons were produced in 1969), and the Artibonite Valley for rice, the yearly crop amounting to about 45,000 tons. Cacao, tobacco, and cotton are also grown. Aromatic plants, such as lime, vetiver (the khuskhus plant, from the roots of which an oil is derived), neroli (sour orange, used for making perfumes), and amyris (a tropical shrub from which a sweet oil is obtained), are grown in the south and are processed for their essential oils.

Table 2: Haitian Production
(in 000 of metric tons)

year	coffee	sugar
1960	25.5	60
1961	43.5	71
1962	35.4	69
1964	33.0	61
1965	34.5	65
1966	27.9	68
1968	27.9	67
1969	27.9	59
1970	28.8	66

Source: *United Nations Statistical Year Book*, 1971.

Crops grown for local consumption include maize (the staple food), manioc, several kinds of peas and beans, and all tropical fruits but especially mangoes. Bananas are mostly of the plantain species, and are eaten as a vegetable rather than as a fruit. Palm trees disappear at about 3,600 feet above sea level, and vegetables and fruits characteristic of temperate climates may be grown. Out of the total agricultural production, farmers use about three-quarters for their own consumption, export about 20 percent, and process about 5 percent.

Mining has so far yielded disappointing results. The production of bauxite was 446,000 tons in 1968. The

output of copper, which was 5,000 tons in 1964 fell to 1,800 tons in 1969, and silver production fell from 92,000 to 17,000 troy ounces during the same period. Alluvial gold, collected by Indians at the time of Columbus, still yields about 8,000 troy ounces a year.

Manufacturing. The oldest and largest industrial plant in Haiti is the Port-au-Prince sugar mill, but two others were opened more recently. Total production averages 66,000 tons, but can increase when higher international quotas permit. A cement factory has a yearly capacity of 60,000 tons, but in 1967 had an output of only 35,000 tons. There are two cotton mills, two soap factories, and a tannery, as well as some small plants producing pharmaceutical and plastic articles, paint, and aerated water.

Fuel-driven generating plants supplied about 80,215,000 kilowatts of electricity in 1968, 90 percent of which went to Port-au-Prince.

Financial services. The state operates a central bank that issues notes and coins in gourdes, the national currency, maintained at a fixed value of U.S. 20 cents (5 gourdes = $1 U.S., 12 gourdes = £1 sterling on December 1, 1970) ever since the time of the American occupation, which ended in 1934. There are state banks for development and for rural credit and three private ones —the Banque Commerciale d'Haïti, the Royal Bank of Canada, and the Banque Populaire Colombo-Haïtienne.

Foreign trade. More than one-half of the foreign trade is conducted with the United States, which took 60 percent of exports and supplied 52 percent of imports in 1967–68. Other buyers of Haitian products are France, Italy, and Japan, while Japan and France also supply goods. Coffee, the largest export item, accounts for about 41 percent of all exports. Sisal represents about 5 percent of exports; other exports are sugar, bauxite, copper, and handicrafts. Imports include cotton goods, foodstuffs, machinery, mineral oils, and motor vehicles.

Table 3: Haitian Foreign Trade
(in $000,000 U.S.)

year	imports	exports	year	imports	exports
1938	8	7	1963	39	41
1948	31	30	1964	41	40
1953	44	38	1965	36	37
1958	43	39	1967	36	34
1960	36	33	1968	38	36
1962	46	43	1970	53	40

Source: *UN Statistical Year Book*, 1969, 1971.

The economy, as is often the case in developing countries, is governed by an oligopoly of foreign merchants who maintain the state through export taxes and custom duties, the burden of which ultimately falls on the peasantry. Out of a revenue of about 140,200,000 gourdes for the fiscal year 1968–69, the government spent about 25.5 percent on defense and police, about 11.7 percent on education, about 13.6 percent on health, and about 8.1 percent on debt redemption.

Trade unions. Trade unions have been recognized since 1946, but they never had more than 7,000 paying members and their numbers declined from 56 in 1954 to 48 in 1964. Employers maintain a chamber of commerce in Port-au-Prince.

Transportation. Transport problems result from the the mountainous character of two-thirds of the country; rivers usually have to be crossed close to the seashore, where their flow is broader and deeper than farther inland. All travelling on land was on horseback until as late as 1915. Since that time about 2,000 miles of roads have been built, but they deteriorate rapidly in the rainy season. There are about 14,700 motor cars in the country. Three short railway lines, radiating from the capital, have a combined length of 187 miles, and are used to carry sugarcane and other produce. The indented coastline, however, has innumerable tiny ports that have been visited by small craft since the days of buccaneers and freebooters. Small sailboats and power boats still call there regularly, and several hundred foreign vessels visit the two main ports of Port-au-Prince and Cap-Haïtien each year.

Difficulty of travel

An airport, accommodating jet aircraft, is situated close to Port-au-Prince; there are frequent flights to and from New York, either directly or via Santo Domingo, San Juan, or Miami.

Administration and social conditions. Since 1804, Haiti has had 20 different constitutions and 41 heads of state, including two emperors and one king. Today, under the 1964 constitution, the president is elected for life. He appoints the members of his Cabinet—all styled secretaries of state—and usually directs one to two or more of the 19 departments into which his administration is divided. There is a unicameral National Assembly of 58 members, and an advisory role is played by a Resources and Development Council and a Budget Bureau. The national territory is divided into five *départements.* The *départements,* each of which has a capital, are further subdivided into communes, of which there are about 100, each one consisting of a centre called a *bourg,* which enjoys municipal authority, and five to six *sections rurales* (rural districts), each supervised by a local resident appointed by the central police authority.

All Haiti's constitutions have provided for universal suffrage, which, since 1950, has also included women, but, as almost all rural citizens are illiterate, the results of elections are always subject to discussion. The only organized party is the "Single Party for Revolutionary and Governmental Action," founded by President Duvalier. In rural areas the peasants are prompt to react when their *chefs de section* (district leaders) exceed their power, but opposition at the national level is expressed only by political groups remaining abroad.

Political activity

Justice. Justice is rendered according to the French tradition. Codes inspired by those of Napoleon were issued in 1825 together with a more peculiar Rural Code that governs the conduct of country life. There is a justice of the peace in every commune, as well as primary courts at departmental capitals, four Courts of Appeal, and a Supreme Court in Port-au-Prince. All judges are appointed by the head of state; the higher ones are appointed for terms of ten years.

The armed forces. The armed forces are divided into four services. These are, firstly, the army proper, consisting of about 1,000 men with 100 officers; the presidential guard, which is composed of 250 men and 15 officers; the air force, consisting of 140 men and 30 officers; and the coast guard, consisting of 250 sailors and 40 officers, manning eight patrol vessels. Since 1961 there has also been a militia, the National Security Volunteers, numbering about 10,000 men and women; they are popularly known as Tontons Macoutes—a name that literally means "uncles knapsacks" and derives from a folk-tale ogre.

Education. The educational tradition also is French. Public schools are free and attendance is compulsory at the primary level, but funds are lacking for the proper implementation of the law, except in Port-au-Prince, which has about as many children at school as there are throughout the rest of the country. There are about 300,000 in school altogether, of which 20,000 are in secondary school and the remainder in primary school. There is a university at Port-au-Prince with an enrollment of about 1,500.

Health and welfare. There are about 350 doctors—approximately one doctor for 13,420 people. In the towns there are 11 hospitals and 12 outpatient clinics, while in the countryside there are about 150 rural clinics and 17 sanatoriums. The United States Agency for International Development (AID) formerly conducted malaria-eradication campaigns and distributed surplus food, but the agency left Haiti in 1966. Yaws was practically eradicated by mass treatment performed with international assistance in the 1950s. No reliable data are available on health conditions but the infant mortality rate is estimated at 170 per 1,000 live births, and an alleged life expectancy of 40 years was mentioned as an objection to a 1966 law providing a pension at 65, on the basis of contributions of 3 percent of earnings over a 20-year period.

Table 4: School Population in Haiti, 1968

level	teachers	pupils
First level (primary, elementary)	6,778	275,192
Second level–general (secondary school, high school, middle school)	1,286	23,047
Vocational (trade and technical, teacher training)	383	5,892
Third level (university, teachers' college, higher professional school)	226	1,313

Regular welfare services are in an embryonic stage. They would, in any case, hardly fit the peculiar needs of the country. But 209 cooperatives may be mentioned as fostering mutual help along lines that are not without roots in Haitian tradition.

Housing. Housing is simple but adequate in the countryside, but is quite unsatisfactory in towns because of overcrowding, despite the fact that five presidents in succession have built model estates out of their private funds. A number of privately built residences in Port-au-Prince offer interesting examples of the use of cement and wrought iron in modern tropical architecture.

Police services. There are 600 policemen in the capital city, and about 2,000 more in the rest of the country, including about 500 people who, as mentioned earlier, act as part-time *chefs de section.*

Wages and cost of living. Wages are low and are hard to control by legislation; an attempt to establish a fixed minimum wage a number of years ago was virtually disregarded. Prices in recent times have risen more rapidly than in the past. Between 1963 and 1969 the cost of living rose by about 20 percent except for rents, which were "frozen" by the government, which forbade further increases.

Social and economic divisions. In sharp contrast to the rest of Latin America, Haiti never had a landed aristocracy. Thus power was for a long time held by an élite of professionals and of merchants of old standing. The contact between them and the illiterate masses was maintained through the army, which, until 1915, was an administrative as well as a military body, thus providing an opportunity for social as well as political mobility. After 1915 social and economic development resulted in the growth of two intermediate social groups. One of them, most conservative in outlook, associated itself with élite employers, but the other consisted of skilled manual and junior office workers who held entirely different views. The tension between these two groups grew more bitter than tension between the élite and the illiterates had ever been. In 1946 the majority, consisting of the poorer people, brought to power Dumarsais Estimé, who was defeated in 1950 by a reaction led by Gen. Paul Magloire; in 1957, however, the majority, most of whom are black rather than mulatto, won again with François Duvalier, who remained in power until his death in 1971. This resulted in a mass emigration of the élite and of many others associated with previous governments.

Tensions between élite and peasants

Cultural life and institutions. Haitian arts and literature have traditionally been under strong French influence, in spirit as well as in language, although the author Georges Sylvain (1865–1925) marked a new departure when he wrote fables in Creole under the title *Cric-Crac* (a traditional password and response exchanged between storyteller and audience at the beginning of a tale) as early as 1901. An interest in local folk culture became apparent in the late 1920s with the publication of a periodical, *La Revue Indigène,* and of *Ainsi Parla l'Oncle,* an essay by the internationally renowned scholar Jean Price-Mars (1876–1970). Another Haitian author, Jacques Roumain (1907–44), first won world fame writing French novels on popular Haitian subjects and on folk tradition, subsequently exercising a powerful influence in theatrical dancing and in painting as well as in literature. The works of Haitian folk painters are hung in many museums and private collections abroad.

Port-au-Prince has museums of history and ethnology, a national library, and a library of Haitiana kept in a Catholic school. Seven dailies and about 35 other periodicals are published. Three dailies are of old standing—the official *Le Moniteur* gazette, with a circulation of 4,000 and *Le Nouvelliste* and *Le Matin,* each of which has a circulation of 3,000. Public libraries are maintained at departmental headquarters.

The capital city has 14 radio stations; a television station was opened in 1965. In 1967 the television station was temporarily closed because of financial problems.

Table 5: Haiti, Area and Population

	area		population	
	sq mi	sq km	1950 census	1971 census*
Départements†				
L'Artibonite	2,625	6,800	567,000	756,000
Nord	1,583	4,100	539,000	698,000
Nord-Ouest	1,062	2,750	168,000	212,000
Ouest	3,050	7,900	1,083,000	1,611,000
Sud	2,394	6,200	740,000	967,000
Total Haiti	10,714	27,750	3,097,000	4,244,000

*Preliminary figures. †Statutory provision has been made for the creation of four new *départements:* Centre, Grand Anse, Nord Est, and Sud Est. No administrative action has yet been taken to create them.
Source: Official government figures.

Prospects for the future. Economically Haiti faces difficulties because the growth in gross domestic product is virtually absorbed by the rate of population increase. In the earlier 1960s there was a small improvement in living standards, but by the later 1960s further population growth had offset economic advancements. The best hope for a more rapid improvement in living standards resides in an irrigation project in the Artibonite River valley, which might enable some 22,000 farmers to settle on 80,000 acres there.

BIBLIOGRAPHY. Official statistical data appears in the *Bulletin trimestriel de statistiqué* (quarterly); and, occasionally, in *Guide économique de la République d'Haïti.* The results of the 1950 census were published in *Recensement général de la République d'Haïti* (1950). Only summary results of the census of 1971 are so far available. General information may be found in S. RODMAN, *Haiti: The Black Republic* (1954); and S. SIMMONDS, *Economic and Commercial Conditions in Hayti* (1956). For their relation to world events, see K. IRVINE, *The Rise of the Colored Races* (1970). Opposite views on recent developments were presented by H. COURLANDER and R. BASTIEN, *Religion and Politics in Haiti* (1966); B. DIEDERICH and A. BURT, *Papa Doc* (1969); and F. DUVALIER, *Mémoires d'un leader du Tiers-Monde* (1969). Changes in the social structure were studied by S. and J. COMHAIRE-SYLVAIN, "Urban Stratification in Haiti," in *Social and Economic Studies, University of the West Indies, Jamaica,* vol. 8 (1959); and the emergence of a new art school by S. RODMAN, *Renaissance in Haiti* (1949). Haitian folk culture was described in a general way by H. COURLANDER, *The Drum and the Hoe* (1960). The results of extensive field work were incorporated in A. METRAUX *et al., Making a Living in the Marbial Valley, Haiti* (1951); and in R.A. HALL *et al., Haitian Creole* (1953).

(E.Sy.)

Haiti, History of

Christopher Columbus discovered the island that now includes Haiti and the Dominican Republic on December 6, 1492, and named it La Isla Española. By the end of the 16th century, most of the island's original Arawak Indian population had disappeared—worked to death or slaughtered outright by the Spaniards or killed by disease. Spanish settlement was thin and restricted mainly to the eastern end of the island; French pirates, based in the Cayman Islands, had an almost unimpeded run of the western end. The pirates began to establish plantations there; in 1664 they founded Port-de-Paix in the northwest, and the French West India Company took possession (see DOMINICAN REPUBLIC, HISTORY OF THE).

Arrival of the French

The French colonial regime (to 1804). In 1697, by the Treaty of Ryswick, the western third of the island was formally ceded to France by Spain and was renamed Saint-Domingue.

Prosperity under the French. The 18th century saw a great increase in Saint-Domingue's population. It became the most prosperous New World colony, exporting sugar, coffee, cocoa, indigo, and cotton cultivated by African slave labour; most of the land and slaves were owned by a handful of families. By 1789 nearly two-thirds of France's foreign investments were based on Saint-Domingue, and in a good year its trade needed more than 700 oceangoing vessels carrying more than 80,000 seamen.

Abolition of slavery. Saint-Domingue had a population in 1789 of 556,000; of this, 500,000 were slaves, 32,000 were whites, and 24,000 were free blacks. On August 24, 1791, stimulated by the French Revolution (begun 1789), the slaves rose in rebellion. In order to maintain the island as a French possession, slavery was abolished by a decree of February 4, 1794. In 1795, by the Treaty of Basel, Spain ceded the rest of the island to France, but war in Europe precluded the actual transfer of possession.

Moves toward independence. In May 1801 Toussaint-Louverture (*q.v.*), a former slave, was declared governor general, but he would not take the final step of declaring the colony independent. Later that year Napoleon sent his brother-in-law, Gen. Charles Leclerc, with an experienced force, including several mulatto officers in exile from Saint-Domingue, to restore the old regime. After several months' resistance, Toussaint came to terms with the French expedition, early in 1802. But the French broke the arrangement and imprisoned him (he died on April 7, 1803).

In the face of a rumour that Napoleon intended to restore slavery in Saint-Domingue as he had done in other French possessions, Jean-Jacques Dessalines and Henry Christophe led a black army against the French in 1802. The French commander and a large part of his army were defeated, and on November 9, 1803, the remnant of the French expedition, under Gen. Jean-Baptiste Rochambeau, surrendered. Under the armistice signed on November 18, the French withdrew, but they maintained a presence in the eastern part of the island until 1809.

Haiti (1804–1957). On January 1, 1804, the whole island was declared independent under its original Arawak name of Haiti.

Independent Haiti (1804–1915). The war with France had utterly laid waste the country and destroyed the economy. In October 1804 Dessalines assumed the title of Emperor Jacques I; on October 17, 1806, he was killed while trying to put down a mulatto revolt and Henry Christophe took control of his kingdom. Civil war broke out between the blacks under Christophe (later Henry I) in the north and mulattoes under Alexandre Sabès Pétion, based at Port-au-Prince in the south. In 1808, with British help, Spanish rule was restored in the eastern part of the island (Santo Domingo).

Christophe managed to improve the country's economy, but he had to force peasants to work on the plantations. He built a spectacular palace, Sans Souci, where he held an equally spectacular court; he also built an imposing fortress, the Citadelle Laferrière, in the hills to the south of Cap-Haïtien—where, with mutinous soldiers almost at his door, he committed suicide in 1820.

Jean-Pierre Boyer, who had succeeded to the presidency of the mulatto south on Pétion's death in 1818, became president of the whole country after Christophe's death. In 1822 he invaded and conquered Santo Domingo, which had declared itself independent from Spain the previous year and was now engaged in fighting the Spaniards and in internal quarrels. Boyer abolished slavery and confiscated church property in Santo Domingo; it was not until 1844 that the Haitians were expelled from Santo Domingo by a popular uprising.

Haitian independence was recognized by France in 1825, in return for an indemnity of nearly 100,000,000 francs, to be paid at an annual rate until 1887. Britain recognized the state in 1833, the United States in 1862, after the secession of the Southern slave states, which had objected to recognition.

Boyer was overthrown in 1843. Between then and 1915 a succession of 20 rulers followed, 16 of whom were overthrown by revolutions or were assassinated. The 1890s saw an increase in U.S. attempts to gain military and commercial privileges in Haiti. In 1905 the U.S. took Haiti's customs into receivership, and before World War I, U.S. business interests had gained a secure financial foothold and valuable concessions.

U.S. involvement

U.S. occupation (1915–34). From 1915 to 1934, Haiti was occupied by U.S. Marines. The U.S. claimed legal justification on the grounds of humanitarian intervention and under the Monroe Doctrine. Many Haitians believed that the Marines had really come to protect U.S. investments in the country and to establish a base to protect the approaches to the Panama Canal. Haiti signed a treaty with the U.S.—originally for ten years, but eventually it lasted until 1934—establishing U.S. financial and political domination. In 1918, in an election supervised by the Marines, a new constitution was introduced under which, for the first time, foreigners were permitted to own land in Haiti.

One effect of the Marine occupation was the nominal re-establishment of the mulatto elite in control of the government. Many Haitians resented the occupation, which they believed excluded them from public office and subjected them daily to racist indignities at the hands of the Marines. The Marines revived an old law of Christophe's time, which enabled them to employ forced labour on the roads. This resulted in a revolt of *cacos* (guerrillas), which was suppressed. The program of public works undertaken by the Marines in health clinics, sewage, and roads hardly satisfied the Haitians, who felt that these efforts barely scratched the surface and left little tangible result.

In October 1930 a national assembly, the first since 1918, was elected. Controlled by nationalists, it in turn elected as president Sténio Joseph Vincent. In August 1934 the U.S. president, Franklin D. Roosevelt, withdrew the Marines; but direct U.S. fiscal control continued until 1941; and indirect control continued until 1947.

Presidential regimes (1934–57). In 1935 a plebiscite extended Vincent's term to 1941 and amended the constitution so that future presidents would be elected by popular vote.

In October 1937, troops and police from the Dominican Republic (formerly Santo Domingo), with popular support, massacred thousands of Haitian labourers living near the border. The following year the Dominican government agreed to pay $3,400,000 in compensation to relatives of those slain, but only part was actually paid. The enmity between the two countries had long historical roots, going back to the events of the 19th century. The Dominicans, with their Spanish culture and greater admixture of European blood, looked with disfavour and a certain amount of fear upon the black Haitian labourer with his "inferior" African culture and lower standard of living, which enabled him to work for far less. Haitian labour was, however, necessary to the Dominican economy.

In 1946 Haitian workers and students held strikes and violent demonstrations in opposition to the president, Élie Lescot, who had succeeded Vincent in 1941 and had altered the constitution to enable himself to serve a further term. Three military officers seized power, and under their supervision Dumarsais Estimé was elected president. In 1950 Estimé in turn sought to extend his term, and the same three officers again took control. That October one of them, Col. Paul E. Magloire, was elected president in a plebiscite. Magloire in turn sought to remain in power unconstitutionally, but in December 1956 the army forced him to resign.

Haiti since 1957. In September 1957, after a period of considerable unrest and the rise and fall of several provisional presidents, François Duvalier (called "Papa Doc") —a black physician, formerly employed on a U.S. medical-aid scheme and a student of vodun (voodoo, an animist religion of African origin, with Christian influ-

Rise of Duvalier

ence)—was elected president in a plebiscite. Duvalier promised to end domination by the wealthy mulatto elite and to put political and economic power into the hands of the black masses. Violence continued, however, and in July 1958 there was an unsuccessful attempt to overthrow Duvalier. He organized an extralegal gang of violent adherents—the Tontons Macoutes—who terrorized the population. Duvalier, by then firmly in control, had himself elected president for life.

Relations with the Dominican Republic. In April 1952 an agreement was signed between Haiti and the Dominican Republic, regulating the movement of Haitian labour. In the late 1950s an uneasy alliance came into existence between Duvalier and Rafael Trujillo, the Dominican leader, to protect themselves against the common threat of uprisings similar to the successful overthrow of the Cuban government in 1959. In the 1960s, with a more liberal government in power in the Dominican Republic, Duvalier sealed off the border, razed a zone contingent to it, and prohibited entry on pain of summary execution; he aimed to keep infiltrators out of Haiti and to prevent Haitians from leaving.

The people. Haiti in the 1970s was a country in which 95 percent of the people were black, the remainder mulatto. The official language was French, but over three-quarters of the population spoke a Creole patois; near the Dominican border a Spanish patois was also spoken. The official religion was Roman Catholicism, but Duvalier claimed to be an authority on vodun, which the majority of the rural population practiced.

The economy. The economy was predominantly agricultural. The land was mostly cultivated in small overworked plots, with primitive tools, resulting in a low level of productivity. Production was mainly for local sale and subsistence. Coffee accounted for 75 percent of exports, the other main commodities being sugar and sisal. The second major source of income was tourism, with about 100,000 annual visitors. The country had considerable untapped resources, including abundant timber, fishing, and such minerals as gold, silver, copper, bauxite, and tin. A plan was announced for the Artibonite Valley, which would irrigate 100,000 acres and also provide electricity. Industry was mainly concerned with processing coffee, sugarcane, sisal, and edible oils; but there was also some production of soap, cement, and cotton fabrics and some light manufacturing for the U.S. market. The late 1960s saw a general economic decline, due partly to the ending of aid and investment from abroad and to a drop in tourism because of an uncertain political situation.

Economic decline in the 1960s

Education. Education was free where available in Haiti, but by 1970 only a small number of children attended schools. Secondary education was limited and provided mainly by the Roman Catholic Church. About 1 percent of the gross national product was spent on education. In 1968 there were 313 urban primary schools, 508 country schools, 16 professional schools, and 15 schools of higher education. There was also a university in Port-au-Prince. The majority of professional people—doctors, dentists, writers, academics, and journalists—had left the country, most going to Africa.

Government and administration. In April 1961 the two houses of the legislature were amalgamated into one. Although the constitution required elections every six years, by the early 1970s none had been held since April 1961. All assembly deputies were members of the one official party, and the assembly was completely subordinate to Duvalier, who ruled by decree. The constitution in effect in the early 1970s was mainly the result of the 1950 military coup, but it had been amended in 1964 to enable Duvalier to be made president for life.

Duvalier's rule. Haiti under Duvalier was, in effect, a police state ruled as a dictatorship. The militia was founded originally because Duvalier mistrusted the army and thus built an armed force of supporters with rudimentary military training. The militia still existed in the early 1970s, although Duvalier had by then strengthened his control over the army. The Tontons Macoutes, Duvalier's private police force, became less brutal and arbitrary, because its earlier conduct had disturbed tourists.

Several attempts to overthrow Duvalier by force were unsuccessful, and in 1968 his enemies tried unsuccessfully to invade Haiti from outside the country. In 1969 Duvalier began to show increasing signs of ill health. His death in 1971 left Haiti's future even more uncertain than before.

BIBLIOGRAPHY

General: J.G. LEYBURN, *The Haitian People* (1941, reprinted 1966 with new introduction by S. MINTZ), an overall survey of the growth of Haiti including a very good sociological and historical analysis of the development of Haitian culture and institutions, based on the thesis that Haitian society is and always has been deeply divided between two castes, a small mulatto elite and the black peasant masses, and a supplementary bibliography dealing with the period since 1941.

Historical background: C.L.R. JAMES, *The Black Jacobins: Toussaint L'Ouverture and the San Domingo Revolution*, 2nd ed. with new appendix (1963), generally accepted as the most revealing analysis of the historical forces, social and economic, that led to the establishment of the independent state of Haiti and of the part played in this movement by Toussaint, Dessalines, and Christophe.

Sociological and cultural background: H. COURLANDER and R. BASTIEN, *Religion and Politics in Haiti* (1966); H. COURLANDER, *The Drum and the Hoe* (1960), an evocative study of the culture of the Haitian peasant as revealed in his daily life in the fields and in the village; M.J. HERSKOVITS, *Life in a Haitian Valley* (1937, reprinted 1964), a penetrating, detailed study of a Haitian peasant community at Mirebalais.

Politics: B. DIEDERICH and A. BURT, *Papa Doc: Haiti and Its Dictator, Francois Duvalier* (1969; U.S. title, *Papa Doc: The Truth About Haiti Today*, 1969), a recent and up-to-date account, sometimes superficial in conception and often clumsy in execution, of the rise of Duvalier and his consolidation of power; L.F. MANIGAT, *Haiti of the Sixties: Object of International Concern* (1964), a penetrating analysis of the present regime and the historical and political background from which it came.

(C.L.R.J.)

Ḥājj ʿUmar, al-

A century ago, ʿUmar Tal founded the empire of the Masina, in the western Sudan (in the modern Republic of Mali). But this black conqueror was hardly a man of his time. He lived, fought, and died more like a 7th-century warrior. He was a mystic, and his life resembled those of the early followers of the Prophet Muḥammad, who fought in the name of God and converted by fire and the sword. Senegalese poets, singing of ʿUmar's life, have compared it with the Prophet's. Some have glorified him and lauded his victories, citing the thousands he killed and the thousands he sold into slavery as proof of the divine character of his mission, but others to this day hate him for having shed Muslim blood.

ʿUmar Tal was born about 1797 in the village of Halvar in Fouta Toro in the upper valley of the Sénégal River and was a member of the Tukulor tribe. His father was an educated Muslim who instructed students in the Qur'ān, and ʿUmar, a mystic, perfected his studies in Arabic and the Qur'ān with scholars of a Moorish tribe who initiated him into the Tijānīyah brotherhood.

At the age of 23, ʿUmar set out on the pilgrimage to Mecca. He was already well-known for his piety and erudition and was received with honour in the countries through which he travelled. Muhammad Bello, emir of Sokoto in Nigeria, offered him his daughter Maryam in marriage. Enriched by this princely alliance, he had become an important personage when he reached Mecca about 1827. He visited the tomb of the Prophet in Medina, returned to Mecca, and then settled for a while in Cairo. On a visit to Jerusalem he succeeded in curing a son of Ibrahim Pasha, the viceroy of Egypt. In Mecca, finally, he was designated caliph for black Africa by the head of the Tijānīya brotherhood.

Pilgrimage to Mecca

Armed with his prestige as a scholar, mystic, and miracle worker, ʿUmar returned to the interior of Africa in 1833. Trained for political leadership by his father-in-law, Muhammad Bello, the emir of Sokoto, with whom he again spent several years, and his position strengthened by the title of caliph, ʿUmar now decided to obey

the voice of God and to convert the pagan Africans to Islām.

By now he was not only looked upon as a miracle worker but had also acquired a bodyguard of followers and of devoted Hausa slaves.

Upon the death of Bello, he left for his native country, hoping to conquer the Fouta with the help of the French, in exchange for a trade treaty, an agreement the French declined because of ʿUmar's growing strength. ʿUmar realized that faith without force would be ineffective and made careful preparations for his task. In northeast Guinea, where he first established himself, he wrote down his teachings in a book called *Kitāb rimāḥ ḥizb ar-raḥīm* ("Book of the Spears of the Party of God"). Deriving his inspiration from Ṣūfism—a mystic Islāmic doctrine—he defined the Tijānī "way" as the best one for saving one's soul and for approaching God. He recommended meditation, self-denial, and blind obedience to the sheikh. He gained many followers in Guinea, but, when, in 1845, he went to preach in his own country, he met with little success.

Military achievements

Having built up an army ʿUmar decided to use force. In March 1854 he issued an order for a *jihād* to sweep away the pagans and bring back the Muslims who had strayed from the fold. Starting out with about 10,000 men who lived off the land, he spread terror in order to force the pagan chieftains to submit. In 1855 he defeated the Bambara pagans of Mali, adding to his empire. He forcibly converted them, yet these conversions proved to be ineffectual. To defend his authority ʿUmar had 300 hostages executed, but revolt broke out again as soon as his armies were removed.

After an unsuccessful attack on a French fort that had refused to supply him weapons, ʿUmar again set off toward the east, but he had great difficulty subsisting in a land already ravaged. His men deserted, and his companions began to doubt his mission.

Having been unable to decisively conquer his adversaries, ʿUmar was to spend the next ten years trying to contain his empire. Repressing new revolts, he was led eastward by the resistance he stirred up. In 1860 he signed a treaty with the French general Louis Faidherbe, governor of Senegal, accepting the Sénégal River as a common boundary.

ʿUmar perennially had to defend his conquests and foil hostile coalitions without giving up the principle of the *jihād*. This proved difficult, however, when he was confronted by the Fulani people of the Masina, who were Muslims, followers of the Qādirīyah brotherhood. When ʿUmar attacked the Fulani, he no longer represented the "wrath of God"—he was a conqueror. His mission turned into a fratricidal war, and ʿUmar inspired in his adversaries a hatred that has survived to this day.

Both armies prayed to the same God before the battle. ʿUmar, recognizing the danger to his divine mission, proposed a duel with Aḥmadu III, the leader of the Fulani army. But the latter refused the judgment of God. ʿUmar won the battle, and Aḥmadu was captured and beheaded.

In 1863 ʿUmar took possession of the city of Timbuktu, but, defeated by the nomadic Tuaregs, he had to beat a retreat. In a subsequent battle, attacked by the Tuaregs, the Moors, and the Fulani, his army was destroyed. He withdrew to the city of Hamdalahi, where he was besieged. He escaped and took refuge in a cave, but on February 12, 1864, he was killed when the cave was blown up with gunpowder.

Al-Ḥājj ʿUmar Tal's empire lasted for 50 years, from 1848 to 1897, when it was annexed by the French. Few of the Mali people still remember it, except the descendants of the Tijānī initiates or the Fulani and Bambaras, who suffered the conqueror's cruelties. In order to enhance his own position, General Faidherbe described ʿUmar in his reports as the symbol of resistance to French penetration, at the same time recognizing his virtues and his courage. In fact, ʿUmar was not anxious to oppose the French. He had sought their neutrality and had hoped to buy arms from them, but they had other sources and feared his power. The mosque of Dinguiraye in Guinea is all that remains of ʿUmar's empire.

BIBLIOGRAPHY. There is no study in English that deals specifically with the life and work of al-Ḥājj ʿUmar. Additional information may be found in J. SPENCER TRIMINGHAM, *A History of Islam in West Africa* (1962); and J.D. HARGREAVES, *Prelude to the Partition of West Africa* (1963). A complete bibliography on Senegal, Masina, and the life of al-Ḥājj ʿUmar is included in YVES-J. SAINT-MARTIN, *L'Empire Toucouleur 1848–1897* (1970), an indispensable work. The *Kitāb rimāḥ ḥizb ar-raḥīm* (written 1845, published in Arabic 1927), is ʿUmar's most famous work, in which he explains his doctrine.

(J.C.F.)

Hakluyt, Richard

The geographer Richard Hakluyt persistently and imaginatively encouraged the continuation of exploration and colonization of the New World. At a time when England was on the brink of becoming a colonial nation, Hakluyt's lectures, publications, and advice proved to be as politically influential as Queen Elizabeth's closest counsellors in promoting the undertaking of overseas expansion. His immense knowledge and editorial skill set him apart from all other geographers of his time, and his three-volume work, *The principall Navigations, Voiages and Discoveries of the English Nation . . .*, provides almost everything known about the early English voyages to North America.

He was born in or near London about 1552, the third son of Richard Hakluyt. The Hakluyt family was of some standing in the Welsh Marches and held property at Eaton. His father died in 1557, leaving his family to the care of a cousin, another Richard Hakluyt, a lawyer who had many friends among prominent city merchants, geographers, and explorers of the day. Because of these connections, and his own expertise in overseas trade and economics, he was well placed to assist the future geographer in his life work.

Young Richard, with the help of various scholarships, was educated at Westminster School and Christ Church, Oxford, entering in 1570 and taking his M.A. degree in 1577. His interest in geography and travel had been aroused on a visit to the Middle Temple, one of the four English legal societies, while in his early teens. As he writes in the "Epistle dedicatorie" to *The principall Navigations*, his cousin spoke to him of recent discoveries and of the new opportunities for trade, and showed him "certeine bookes of Cosmographie, with an universall Mappe." His imagination thus stirred, the schoolboy had thereupon resolved to "prosecute that knowledge and kinde of literature" at the university. Some time before 1580 he took holy orders, and, though he never shirked his religious duties, he spent considerable time reading whatever accounts he could find about contemporary voyages and discoveries.

Public lectures

Hakluyt also gave public lectures—he is regarded as the first professor of modern geography at Oxford—and was the first to display

> both the olde imperfectly composed, and the new lately reformed Mappes, Globes, Spheares, and other instruments of this Art for demonstration in the common schooles.

He made a point also of becoming acquainted with the most important sea captains, merchants, and sailors of England. This was the time when English attention was fixed on finding the northeast and northwest passages to the Orient, and on Francis Drake's circumnavigation of the world. Hakluyt was concerned with the activities of Sir Humphrey Gilbert and Martin Frobisher, who were both searching for a passage to the East; was consulting Abraham Ortelius, compiler of the world's first atlas, and Gerardus Mercator, the Flemish map maker, on cosmographical problems; and was gaining approval for future overseas exploration from such politically prominent men as Lord Burghley, Sir Francis Walsingham, and Sir Robert Cecil. He thus embarked upon his career as a "publicist and a counsellor for present and future national enterprises across the ocean." His policy, constantly expounded, was the exploration of temperate North America in conjunction with the search for the Northwest Passage, the establishment of England's claim to possession based on the discovery of North America

by John and Sebastian Cabot, and the foundation of a "plantation" to foster national trade and national well-being. These views are first set out in the preface he wrote to John Florio's translation of an account of Jacques Cartier's voyage to Canada, which he induced Florio to undertake, and are further developed in his first important work, *Divers voyages touching the discouerie of America* (1582). In this he also pleaded for the establishment of a lectureship in navigation. In 1583 Walsingham, then one of the most important secretaries of state, sent Hakluyt to Paris as chaplain to Sir Edward Stafford, the English ambassador there. He served in Paris also as a kind of intelligence officer, collecting information on the fur trade of Canada and on overseas enterprises from French and exiled Portuguese pilots. In support of Walter Raleigh's colonizing project in Virginia, he prepared a report, known briefly as *The Discourse on the Western Planting* (written in 1584), which set out very forcefully the political and economic benefits from such a colony and the necessity for state financial support of the project. This was presented to Queen Elizabeth I, who rewarded Hakluyt with a prebend (ecclesiastical post) at Bristol Cathedral but took no steps to help Raleigh. *The Discourse*, a secret report, was not printed until 1877. In Paris Hakluyt also edited an edition of the *De Orke Novo* of Pietro Martire so that his countrymen might have knowledge of the early successes and failures of the Spaniards in the New World.

Hakluyt returned to London in 1588. The outbreak of war with Spain put an end to the effectiveness of overseas propaganda and the opportunity for further exploration so he began work on a project that he had had in mind for some time. This was *The principall Navigations, Voiages and Discoveries of the English Nation . . .*, which, by its scholarship and comprehensiveness, transcended all geographical literature to date; the first edition, in one volume, appeared in 1589. About this time he married Douglas Cavendish, a relative of Thomas Cavendish, the circumnavigator, and was appointed to the parish of Wetheringsett in Suffolk. Until after the death of his wife in 1597, little is heard of any geographical work, but he then completed the greatly enlarged second edition of the *Voyages*, which appeared in three volumes between 1598 and 1600. Shortly before its completion, he was granted by the Queen the next vacant prebend at Westminster so that he might be at hand to advise on colonial affairs. He gave information to the newly formed East India Company and continued his interest in the North American colonizing project; he was one of the chief promoters of the petition to the crown for patents to colonize Virginia in 1606 and at one point contemplated a voyage to the colony. Nor did his belief in the possibility of Arctic passages to the East fade, for he was also a charter member of the Northwest Passage Company of 1612. In 1613 appeared the *Pilgrimage* of Samuel Purchas, another clergyman fascinated with the new discoveries of the age; in spirit, it was a continuation of Hakluyt's own work, and the two editors probably became acquainted. Purchas procured some of Hakluyt's manuscripts after his death and used them in *Hakluytus Posthumus or Purchas his Pilgrimes* of 1625. Hakluyt died on November 23, 1616, and was buried in Westminster Abbey.

Major work

Works by Hakluyt in addition to those mentioned above include translations of Antonio Galvano's *Discoveries of the World . . .* (1601), and of Hernando de Soto's account of Florida, under the title *Virginia richly valued by the description of . . . Florida . . .* (1609). But it is the *Voyages* that remain his memorial. This, the prose epic of the English nation, is more than a documentary history of exploration and adventure; with tales of daring it mingles historical, diplomatic, and economic papers to establish British right to sovereignty at sea and to a place in overseas settlement. Its overriding purpose was to stimulate, guide, and encourage an undertaking of incalculable national import. Hakluyt was not blind to the profits arising from foreign trade. It has been asserted that the income of the East India Company was increased by £20,000 through a study of Hakluyt's *Voyages*.

BIBLIOGRAPHY. *The principall Navigations . . .* were reprinted with additional matter as *Hakluyt's Collection of the Early Voyages, Travels, and Discoveries of the English Nation*, new ed., 5 vol. (1809–12), and for the Hakluyt Society with a preface by WALTER RALEIGH, 12 vol. (1903–05). *The Divers voyages* was edited by the Hakluyt Society (1850). The best text of *The Discourse on the Western Planting* is in E.G.R. TAYLOR (see below). Many narratives from Hakluyt's collection have been republished by the Hakluyt Society (founded 1846). For his life, the dedications to the two early editions of *The principall Navigations* should be consulted. See also G.B. PARKS, *Richard Hakluyt and the English Voyages*, 2nd ed. (1961); E.G.R. TAYLOR, *The Original Writings and Correspondence of the Two Richard Hakluyts*, Hakluyt Society, 2nd series, vol. 76–77 (1935); J.A. WILLIAMSON, "Richard Hakluyt," in *Richard Hakluyt and His Successors*, Hakluyt Society, 2nd series, vol. 93 (1946); D.B. QUINN and R.A. SKELTON, "Introduction" (on character, production, and sources) in the facsimile edition of *The principall Navigations . . .*, 2 vol., pp. ix–lx (1965); D.B. QUINN, "Richard Hakluyt, Editor," study accompanying facsimile edition of Richard Hakluyt's *Divers voyages*, 1582 (1967); and T. GIRTIN, "Mr. Hakluyt, Scholar at Oxford," *Geogrl. J.*, 119:208–212 (1953).

(G.R.C.)

Hale, Sir Matthew

One of the greatest scholars of the history of English common law, Sir Matthew Hale first gained renown for his fair-minded conduct during the often violent political fluctuations in England's Civil War (1642–51) and his prominent role as counsel in the political trials of royalist defendants. His reputation for integrity and sagacity grew steadily throughout his career, and even though he was a judge under Oliver Cromwell, Charles II upon his restoration raised him to high office and eventually made him lord chief justice.

By courtesy of the National Portrait Gallery, London



Hale, oil painting after John Michael Wright (1623–1700). In the National Portrait Gallery, London.

Hale was born at Alderley, Gloucestershire, on November 1, 1609, the son of Robert Hale, a barrister. Orphaned at the age of five, he was educated according to Puritan principles under the direction of his guardian, until he enrolled at Magdalen College, Oxford, in 1626, with the intention of taking holy orders. He soon changed his mind and began to devote most of his time to sports, especially fencing, and to gambling and other diversions; at one time he considered enlisting as a soldier in the service of Frederick Henry, prince of Orange. A consultation with an eminent lawyer on family business so impressed him that he chose the law as his profession. In 1628 he was admitted to Lincoln's Inn, one of London's four legal schools and societies. He studied 16 hours a day, and under the guidance of such friends as John Selden, one of the leading jurists and scholars of his age, extended his studies to include Roman law, English history, mathematics, and natural philosophy. Called to the bar in 1637, he soon had a flourishing practice.

Hale remained aloof from the Long Parliament's oppo-

sition in the 1640s to King Charles and avoided taking sides during the Civil War between the King and Parliament. Nevertheless he defended many royalists, most notably Archbishop William Laud, who had persecuted Puritan churchmen; and he probably advised the Earl of Strafford, who was impeached by the House of Commons on charges of high treason, and later Charles I during his trial. In spite of his support of the royalists, in 1649 he took the oath of fidelity to the republican Commonwealth and, later in 1654, was persuaded by his royalist friends to accept a judgeship from Oliver Cromwell, now lord protector. In 1651 and 1652, he was active in the law-reform movement and contributed much to the work of the committee that advised Parliament on far-reaching improvements in the law and legal system of the time. On Cromwell's death he refused to continue as judge and was returned to Parliament as member for Oxford. He took a prominent part in the proceedings of the Convention Parliament, called after the dissolution of the Long Parliament, and in promoting the restoration of Charles II.

Role in the Civil War

In 1660 Hale was appointed chief baron of the Exchequer, the court principally concerned with matters of crown revenue, and in the same year he was knighted. Between 1666 and 1672 he spent much time on the statutory tribunal that resolved disputes between owners and tenants of property destroyed in the Great Fire of London in 1666. In 1671 he became chief justice of the King's Bench, an office he relinquished in 1676 when his health began to fail. He died on Christmas day of the same year, at Alderley.

Hale's work on the Bench—in an age when these attributes were not a matter of course even among judges—was characterized by singular personal integrity and judicial impartiality. Moreover, he acted with scrupulous fairness towards prisoners. The one point on which he was criticized by later writers was his belief in witchcraft, and he once permitted the execution of two women accused as witches. Hale was tolerant in religious matters and on numerous occasions mitigated the rigours of the law against dissenters from the Church of England. Throughout his life he retained his Puritan sympathies and numbered among his intimate friends prominent nonconformists. But, evenhanded also in his friendships, he maintained close ties with Anglican bishops as well.

Lord Chancellor Nottingham (the other outstanding legal figure of Hale's generation) wrote of him that

> as great a lawyer as he was, he would never suffer the strictness of the law to prevail against conscience; as great a chancellor as he was, he would make use of all the niceties and subtleties of law when it tended to support right and equity.

Writings

But Hale is principally remembered not as a judge but as a jurist. He was a prodigious searcher of legal records and formed an extensive collection of manuscripts and transcripts. The bulk of this collection is now deposited in the library of Lincoln's Inn. On the basis of these manuscripts and transcripts he wrote numerous books and treatises, though he published little of his own legal work during his lifetime; some of his treatises were printed posthumously, others still remain unpublished. The published work by which he is perhaps best known is his *History of the Pleas of the Crown* (the House of Commons directed in 1680 that it should be printed, though it was not published till 1736). This work remains one of the principal authorities on the common law of criminal offenses. But he also wrote widely on topics of constitutional and civil law, as his editorial talent enabled him to analyze and rearrange the jumbled collection of 17th-century and earlier law materials. When Sir William Blackstone wrote his classic *Commentaries on the Laws of England* (1765–69), he found that he could not do better than adopt Hale's "Analysis of the Civil Part of the Law."

Hale's literary talent was enhanced by his considerable critical faculty. He was both an historian and critic of the law, and his writings take stature from his talents as historian and critic. His place is undoubtedly among the principal figures in the history of English common law.

BIBLIOGRAPHY. The principal authority for Hale's life is GILBERT BURNET, *Life and Death of Sir Matthew Hale* (1682). Fuller accounts are given in SIR JOHN WILLIAMS, *Memoirs of the Life, Character and Writings of Sir Matthew Hale* (1835); and in LORD CAMPBELL, *The Lives of the Chief Justices of England*, ch. 16–18 (1849). A balanced outline of Hale's career and work is SIR WILLIAM HOLDSWORTH, *A History of English Law*, 2nd ed., vol. 6, pp. 574–595 (1924); and a modern study of Hale's political philosophy is J.G.A. POCOCK, *The Ancient Constitution and the Feudal Law*, ch. 7 (1957). The biographical account by J.M. RIGG in the *Dictionary of National Biography*, vol. 24, pp. 18–24, is of merit, and contains a catalogue of Hale's writings, both published and unpublished.

(D.E.C.Y.)

Ḥallāj, al-

Abū al-Mughīth al-Ḥusayn ibn Manṣūr al-Ḥallāj was a celebrated but controversial representative of Islāmic mysticism (Ṣūfism). Because he represented in his person and works the experiences, causes, and aspirations of many Muslims, arousing admiration in some and repression on the part of others, the drama of his life and death has been considered a reference point in Islāmic history.

Early life and education

He was born around AD 858 in the southern Iranian community of Ṭūr in the province of Fars. According to tradition, his grandfather was a Zoroastrian and a descendant of Abū Ayyūb, a companion of Muḥammad. At an early age al-Ḥallāj went to live in the city of Wāsiṭ, an important Iraqi centre for textiles, trade, and Arab culture. His father had become a Muslim and may have supported the family by carding wool.

Al-Ḥallāj was attracted to an ascetic way of life at an early age. Not merely satisfied with having learned the Qur'ān (the Islāmic scripture) by heart, he was motivated to understand its deeper and inner meanings. During his adolescence (*c.* 874–894), at a time when Islāmic mysticism was in its formative period, he began to withdraw from the world and to seek the company of individuals who were able to instruct him in the Ṣūfī way. His teachers, Sahl at-Tustarī, ʿAmr ibn ʿUthmān al-Makkī, and Abū al-Qāsim al-Junayd, were highly respected among the masters of Ṣūfism. Studying first under Sahl at-Tustarī, who lived a quiet and solitary life in the city of Tustar in Khuzistan, al-Ḥallāj later became a disciple of al-Makkī of Basra. During this period he married the daughter of the Ṣūfī Abū Yaʿqūb al-Aqṭaʿ. He concluded his instruction in the mystical way under al-Junayd of Baghdad, a brilliant intellect, under whom al-Makkī had likewise studied.

During the next period of his life (*c.* 895–910), al-Ḥallāj undertook extensive travels, preaching, teaching, and writing. He made a pilgrimage to Mecca, where he followed a strict discipline for a year. Returning to such regions as Fars, Khuzistan, and Khorāsān, he preached and wrote about the way to an intimate relationship with God. In the course of his journeys he attracted a large number of disciples, some of whom accompanied him on a second pilgrimage to Mecca. Afterward, he returned to his family in Baghdad and then set out by sea for a mission to a territory hitherto not penetrated by Islām—India and Turkistan. Following a third pilgrimage to Mecca, he again returned to Baghdad (*c.* 908).

Factors leading to his arrest

The milieu in which al-Ḥallāj preached and wrote was filled with social, economic, political, and religious tensions—all factors that contributed to his later arrest. His thought and activity had been provocative and had been interpreted in various ways, some of which left him highly suspect in the eyes of civil and religious authorities. The Ṣūfī movement, in general, had aroused considerable opposition, and its thought and practice had yet to be coordinated with developments in jurisprudence, theology, and philosophy.

Al-Ḥallāj's propensity for travel and his willingness to share the profundity of his mystical experiences with all who would listen were considered breaches of discipline by his Ṣūfī masters. His travel for missionary purposes was suggestive of the subversive activity of the Qarmaṭians, a 9th-century movement with Ismāʿīlī affiliations that was founded by Ḥamdān Qarmaṭ in Iraq, whose

acts of terrorism and whose missionaries were undermining the authority of the central government. Through his wife's family, he was suspected of having connections with the destructive Zanj rebellion in southern Mesopotamia that was carried out by oppressed Negro slaves inspired and led by outside dissidents. The alleged involvement of al-Ḥallāj in an attempt at political and moral reform upon his return to Baghdad was an immediate factor in his arrest, and it did nothing to improve his image in the eyes of the political leaders.

Al-Ḥallāj has been identified as an "intoxicated" Ṣūfī in contradistinction to a "sober" one. The former are those who, in the moment of ecstasy, are so overcome by the presence of the Divine that awareness of personal identity is lost and who experience a merging with Ultimate Reality. In that exalted state, the Ṣūfī is given to use extravagant language. Not long before his arrest al-Ḥallāj is said to have uttered the statement, "Anā al-ḥaqq" ("I am the Truth"—*i.e.*, God), which provided cause for the accusation that he had claimed to be divine. Such a statement was highly inappropriate in the view of most Muslims. Furthermore, this was the kind of theosophical (divine wisdom) idea that was associated with the Qarmaṭians and the supporters of the Zanj slaves. There was no consensus about al-Ḥallāj, however. The long, drawn-out trial proceedings were marked by indecision.

Death and influence

After his arrest in Sūs (*c.* 911–922) and a lengthy period of confinement in Baghdad, al-Ḥallāj was eventually crucified and brutally tortured to deatn on March 26, 922. A large crowd witnessed his execution. He is remembered to have endured gruesome torture calmly and courageously and to have uttered words of forgiveness for his accusers. In a sense, the Islāmic community (*ummah*) had put itself on trial, for al-Ḥallāj left behind revered writings and supporters who courageously affirmed his teachings and his experience. In subsequent Islāmic history, therefore, the life and thought of al-Ḥallāj has been a subject seldom ignored.

BIBLIOGRAPHY. LOUIS MASSIGNON, *Passion d'al-Ḥallāj*, 2 vol. (1922; Eng. trans., *The Passion of al-Ḥallāj*, 3 vol., 1974), the definitive work on al-Ḥallāj, but not strictly a biography; *Encyclopaedia of Islam*, new ed., vol. 3 (1971), contains a comprehensive survey of the life, thought, and times of al-Ḥallāj.

(J.W.F.)

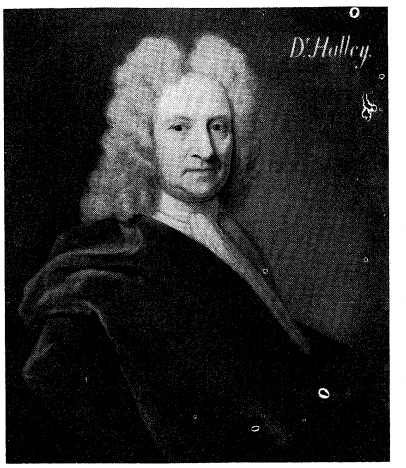
Halley, Edmond

Edmond Halley, English astronomer and mathematician, is best known for his role in the publication of Newton's *Philosophiae Naturalis Principia Mathematica* and for accurately calculating the orbit of the comet that bears his name.

Born on November 8, 1656, Halley began his education at St. Paul's School, London. Halley had the good fortune to live through a period of scientific revolution that established the basis of modern thought. He was four years old when the monarchy was restored under Charles II; two years later the new monarch granted a charter to the informal organization of natural philosophers originally called the "invisible college," which then became known officially as the Royal Society of London. He entered Queen's College, Oxford, in 1673 and there was introduced, by letter, to John Flamsteed, who was appointed astronomer royal in 1676. On one or two occasions, Halley visited the Royal Greenwich Observatory, where Flamsteed did his work, and there was encouraged to study astronomy.

Observations in the Southern Hemisphere

Influenced by Flamsteed's project of using the telescope to compile an accurate catalog of northern stars, Halley proposed to do the same for the Southern Hemisphere. With financial assistance from his father and, from King Charles II, an introduction to the East India Company, he sailed in November 1676 in a ship of that company (having left Oxford without his degree) for the island of St. Helena, the southernmost territory under British rule, in the South Atlantic. Bad weather frustrated his full expectations. But, when he embarked for home in January 1678, he had recorded the celestial longitudes and latitudes of 341 stars, observed a transit of Mercury across the Sun's disk, made numerous pendulum observations, and noticed that some stars apparently had become fainter since their observation in antiquity. Halley's star catalog, published late in 1678, was the first such work to be published containing telescopically determined locations of southern stars, and it established his reputation as an astronomer. In 1678 he was elected a fellow of the Royal Society and, with the intercession of the King, was granted the M.A. degree from Oxford University.



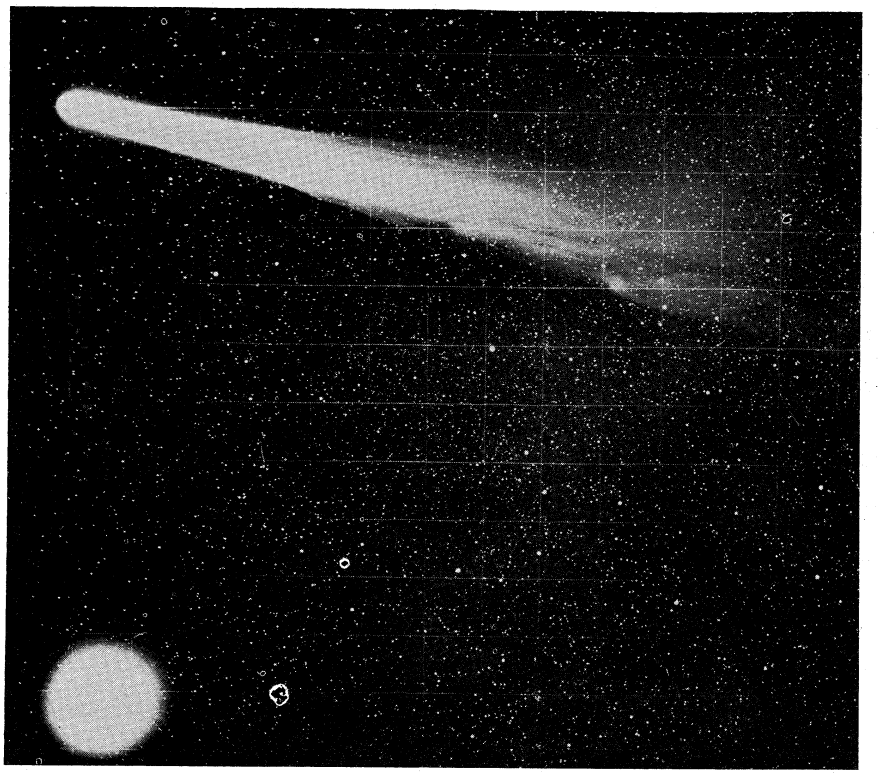
Halley, oil painting by R. Phillips, *c.* 1720. In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

In 1684 Halley made his first visit to Sir Isaac Newton in Cambridge, an event that led to his prominent role in the development of the theory of gravitation. Halley was the youngest of a trio of Royal Society members in London that included Robert Hooke, the inventor and microscopist, and Sir Christopher Wren, the famous architect, both of whom, with Newton at Cambridge, were attempting to find a mechanical explanation for planetary motion. Their problem was to determine what forces would keep a planet in forward motion around the Sun without either flying off into space or falling into the Sun. Since these men were dependent upon their scientific stature for both livelihood and sense of achievement, each had a personal interest in being the first to find a solution. This desire for priority, a propelling motive in science, was the cause of much lively discussion and competition between them.

Halley and Newton

Although Hooke and Halley had calculated that the force keeping the planets in orbit decreased as the inverse of the square of the distances between them, they were not able to deduce from this hypothesis a theoretical orbit that would match the observed planetary motions, despite the incentive of a prize offered by Wren. Halley then visited Newton, who told him he had already solved the problem—the orbit would be an ellipse—but that he had mislaid his calculations to prove it. Encouraged by Halley, Newton then expanded his studies on celestial mechanics into one of the greatest masterpieces produced by the mind of man, the *Principia*. The Royal Society decided that "Mr. Halley undertake the business of looking after it, and printing it at his own charge," which he proceeded to do. He consulted with Newton, tactfully subdued a priority dispute between Newton and Hooke, edited the text of the *Principia*, wrote laudatory verse in Latin for the preface to honour its author, corrected the proofs, and saw it through the press in 1687.

Halley had the ability to reduce large amounts of data to a meaningful order. In 1686 his map of the world, showing the distribution of prevailing winds over the oceans, was the first meteorological chart to be published. His mortality tables for the city of Breslau, published in 1693, comprised one of the first attempts to relate mortality and age in a population; as such, it influenced the future development of actuarial tables in life insurance. Under instructions from the Admiralty, he commanded the war sloop "Paramour Pink" in 1698–1700 on the first



Halley's Comet, with the planet Venus at the lower left.
Camera Press—Publix

sea voyage undertaken for purely scientific purposes, this one to observe variations in compass readings in the South Atlantic and to determine accurate latitudes and longitudes of his ports of call. In 1701 he published the first magnetic charts of the Atlantic and Pacific areas, showing curved lines that indicated positions in the oceans having the same variation of the compass. Such charts, compiled from all available observations augmented with many of his own made on sea voyages, were of great practical value in navigation and were used for many years after his death. Notwithstanding opposition from Flamsteed, Halley in 1704 was appointed Savilian professor of geometry at Oxford, where he continued to investigate the reckoning of accurate longitude at sea.

Halley's comet

Continuing his pioneering work in observational astronomy, Halley published in 1705 *A Synopsis of the Astronomy of Comets*, in which he described the parabolic orbits of 24 comets that had been observed from 1337 to 1698. He showed that the three historic comets of 1531, 1607, and 1682 were so similar in characteristics that they must have been successive returns of the same visitant—now known as Halley's comet—and accurately predicted its return in 1758.

In 1716 he devised a method for observing transits of Venus across the disk of the Sun, predicted for 1761 and 1769, in order to determine accurately, by solar parallax, the distance of the Earth from the Sun. In 1718 he published data to show that stars actually move in space. In 1720 Halley succeeded Flamsteed as astronomer royal at Greenwich, where he made observations, such as timing the transits of the Moon across the meridian, that he hoped would eventually be useful in determining longitude at sea.

Halley was an outstanding 17th-century virtuoso who, in the early days of the Royal Society, examined many different scientific questions. His concern with practical applications of science, such as problems of navigation, reflects the influence on the Royal Society of Francis Bacon, who held that science should be for the "relief of man's estate." Though wide ranging in his interests, Halley displayed a high degree of professional competence that foreshadowed scientific specialization. His wise assessment of Newton's work and his persistence in guiding it to completion earned for him an important place in the emergence of Western thought. He died at Greenwich on January 14, 1742.

BIBLIOGRAPHY. All that is now known of Halley's life is contained in a collection of contemporary materials by EUGENE FAIRFIELD MACPIKE, *Correspondence and Papers of Edmond Halley* (1932) and *Hevelius, Flamsteed and Halley: Three Contemporary Astronomers and Their Mutual Relations* (1937); and in the biographies by C.A. RONAN, *Edmond Halley: Genius in Eclipse* (1969); and A. ARMITAGE, *Edmond Halley* (1966).

(O.J.E.)

Hallucinogen

Psychopharmacological agents are drugs that act predominantly on psychological processes, such as perception, thinking, feeling, and acting. They are divided into two main classes: the psychotherapeutic drugs, which may be tranquillizers, antidepressants, or anti-anxiety agents, and the psychotomimetic drugs, which produce such psychotic reactions as depression, mania, or schizophrenia in normal subjects. Those that reproduce schizophrenic reactions are termed hallucinogens.

Characteristically hallucinogens produce changes in perception, which vary from sensory illusion to outright hallucinations. Illusions are mere distortions of what is sensed, while hallucinations are objects sensed where there are, in fact, none to be sensed. The human response to these perceptual changes depends upon a large number of factors that include personality, education, dose, reason for taking the drug, and previous experience. The experience may mimic psychosis (and is then termed psychotomimetic); under different circumstances, it may produce a quasi-mystical or "psychedelic" experience, which has been defined as one that enriches the mind and enlarges the vision.

Influence on psychiatry

The hallucinogens have also influenced the development of a new, more biologically oriented psychiatry, in which there is a renewed interest in experience and in the role perception plays in producing symptoms and behavioural changes in the schizophrenic. Doctors and nurses who have experienced model psychoses have found themselves able to communicate more effectively with patients, and the millions of young people who have experienced transient model psychoses have helped to make schizophrenia

less mysterious. Recent hypotheses concerning the role of biochemical factors in the development of schizophrenia have arisen from studies with d-lysergic acid diethylamide (LSD). It is possible that ancient man used his experience with natural hallucinogens to create his religions. Mescaline, an active ingredient of the peyote cactus, *Lophophora williamsii*, is used as a sacrament by the Native American Church of North America. Lysergic acid amide, much less potent a hallucinogen than LSD, is present in many varieties of morning glory plants and is used extensively in magical ointments and potions.

A number of the naturally occurring hallucinogens are indole derivatives, including N,N-dimethyltryptamine (DMT), bufotenine, psilocybin, harmine, ibogaine, and lysergic acid amide. All of these compounds have been used in one way or another by primitive peoples for their hallucinogenic qualities. Two other naturally occurring hallucinogens, mescaline and tetrahydrocannabinol (the active ingredient of marijuana), are not indole derivatives, but the former at least has some structural resemblance to indole. Lysergic acid diethylamide (LSD) is a synthetic derivative of the naturally occurring lysergic acid, and diethyltryptamine (DET) is a synthetic compound that closely resembles DMT. Many drugs used primarily for their other pharmacological effects, particularly the anticholinesterase agents, also are hallucinogenic.

Sites and nature of action. It is clear that hallucinogens must interfere with normal chemical reactions in the central nervous system, chiefly in the brain. But there is no generally accepted theory that explains how they do so. The hypotheses currently being investigated derive from the structure and function of the brain as an organ for receiving, processing, and transmitting information. The main transmitting elements of the brain are the nerve cells, or neurons, that are separated from one another by small gaps called synapses. When an impulse is transmitted along the neurons, it cannot directly jump these gaps, but substances called "transmitters" are believed to do so.

Transmitter substances

These so-called neurohormone transmitters are of three main types: (1) sympathomimetic amines derived from tyrosine, such as dopamine, epinephrine (adrenalin), norepinephrine (noradrenalin), and perhaps isoproterenol (isopropylnoradrenalin); (2) indoles, such as serotonin, derived from tryptophan; and (3) acetylcholine and similar compounds. The last two groups are believed to be closely related to substances found in the parasympathetic nervous system. It is generally supposed that the hallucinogens exert their major effect by interfering in some way with the action of the transmitters. The best evidence for this point of view arises from the remarkable similarity in structure between the main classes of hallucinogens and the three classes of neurohormone transmitters. Mescaline is similar in structure to the sympathomimetic amines. (The amphetamines, similar structurally, are euphorients and, in large quantities or over long periods of time, they are capable of producing psychoses.) The large number of indole hallucinogens are similar in structure to serotonin and could be imagined to interfere with its role as a transmitter. Finally, substances like atropine, which interfere with the transmitter action of acetylcholine, are also potent hallucinogens.

The resemblance between the hallucinogens and the transmitters could have any one or more of a number of effects. Thus, it has been proposed that the hallucinogens could act: (1) by causing anatomical changes in neurons; (2) by interfering directly in chemical reactions within the cell; (3) by altering blood-flow relationships in the brain; (4) by affecting the blood-brain barrier (and allowing nontoxic substances in the blood to penetrate into the brain where they become toxic); and (5) by displacing in the synapses the transmitters they resemble. (The hallucinogens could also be interfering with chemical reactions outside the brain; chemicals normally present in low concentration might then build up to the point where they would penetrate into the brain and cause abnormal changes.)

It now seems likely that psychological response to the hallucinogens may yield sooner to investigation than physiological response. Because of the dramatic nature of the perceptual changes brought about by hallucinogens, it is obvious that these changes must play a major role in shaping the psychological response. John Conolly in 1830 defined psychosis as a disease of perception combined with an inability to judge that the perceptual changes involved are unreal. In a sense, this definition is appropriate today, for it takes into account the factors known to shape personality (*e.g.*, inheritance, training, education, culture). In a similar way, a subject responding to a hallucinogen will observe certain perceptual changes (usually visual; less often auditory); however, his final experience will be influenced not only by the nature of the change but also by his judgment of it.

Biological disposition. Hallucinogens are readily absorbed by all the usual routes of administration, but the most common methods are by ingestion or inhalation. As a rule, administration by injection accelerates the response.

Metabolic transformation

The metabolism of mescaline and LSD has been studied most thoroughly. Whereas some mescaline is excreted unchanged in the urine, it is mainly oxidized. In rabbits a chief excretion product is 3,4,5,-trimethoxyphenylacetic acid. In humans some 3,4-dihydroxy-5-methoxyphenylacetic acid is excreted. With LSD most of the drug is degraded and excreted within several hours, although small quantities may be present for as long as 12 hours. Only about 1 percent is found in the brain. One of the metabolic derivatives is 2-OXY-LSD, which is psychologically inactive.

DMT and DET are metabolized very quickly in the body; ten minutes after injection they are no longer present. These compounds are mainly degraded in the body into two inactive derivatives, indole-3-acetic acid and a 6-hydroxy substance. After psilocybin is injected into rats, the highest concentration is reached in 30 minutes, after which it decreases rapidly. In 24 hours 94% is excreted, two-thirds of it in the urine. About 25% is excreted unchanged; 15% is converted to 4-hydroxyindoleacetic acid; 10% is demethylated and deaminated (with the rest unaccounted for). The parasympathomimetic hallucinogens also are excreted very quickly; up to 95% is excreted in urine within two hours. Again, very little is found in the brain after one hour (less than 0.1% of the dose).

All the hallucinogens are relatively nontoxic, in the sense that there is a wide range between the mean dose required to produce the specific psychological reaction and the dose that is lethal for one-half the animals tested (LD_{50}). In rats LD_{50} for mescaline given intraperitoneally is 370 mg per kg, whereas the active dose is about 10 mg per kg. No deaths in humans have yet been attributed to mescaline (or any of the hallucinogens), so the LD_{50} in man cannot be determined.

Members of the Native American Church of North America have taken peyote weekly for many years with no evidence of physical harm. LD_{50} for LSD was 46, 16.5, and 0.3 mg per kg for mice, rats, and rabbits respectively. Comparable values are not known for higher mammals. In chronic trials using about 100 μg per day for autistic children over periods of up to 1½ years, LSD has been nontoxic. Psilocybin at a dose of 250 mg per kg given intravenously proved fatal to a number of mice.

On the basis of studies on subjects who frequently took drugs obtained from illegal sources, it was concluded that LSD, especially, produced chromosomal aberrations in leucocytes and was responsible for congenital defects in offspring of the users; however, every study recorded by mid-1970, which used pure LSD on known subjects and was carried out with controls, failed to show chromosomal changes. Other findings suggest that the incidence of congenital abnormality among the offspring of subjects given LSD is not greater than it is in the general population.

Reactions to the hallucinogens can occur, however, which, from a psychiatric point of view, can be more dangerous than any toxic reaction. These are of two

kinds: a marked prolongation of the psychotic reaction (which may occasionally last many months) and a transient or prolonged reappearance of the psychotomimetic reaction some months after the last experience. During recurrences, serious accidents, including suicide, can occur. There is a fair amount of evidence that these reactions occur mainly in subjects who are ill before they take hallucinogens, but there is little doubt that the psychosis can be greatly aggravated by these drugs.

Dangerous reactions

In general, prolonged reactions or recurrences are best dealt with in two ways, either by personal contact with experienced friends or trained professionals who can reassure the subject, or by use of psychotropic drugs, especially tranquillizers. If parasympathomimetic drugs are consumed, tranquillizers must be used with caution, however. Massive doses of vitamin B_3 (nicotinic acid and nicotinamide) are used on a large scale by youthful drug users to terminate psychotomimetic reactions (called freak outs, bad trips, etc.).

The principal hallucinogens. *Mescaline.* Chemically mescaline is 3,4,5-trimethoxyphenethylamine, a colourless, alkaline oil, soluble in water and alcohol. The sulfate and chloride salt are white crystalline powders, very soluble in water. The three adjacent oxygen functions on the benzene ring seem vital for the hallucinogenic activity. The drug is usually prepared from peyote cactus by extraction and purification, but it can be synthesized. When given by mouth the usual dose begins to produce an effect in 1½ to 3 hours, reaches a maximum at about 8 hours, and then abates in approximately 12 hours. The prolonged action of mescaline compared to other hallucinogens may be an advantage for psychedelic therapy. A disadvantage of the drug is that a large dose is required, which often produces nausea and vomiting.

Lysergic acid diethylamide. Lysergic acid diethylamide can exist as four stereoisomers, but only the dextro form is active psychologically. Active LSD is prepared from lysergic acid, the main source of which is grain (usually rye) infected with the fungus ergot. The pure fungus in culture can also be used to synthesize lysergic acid, however. Impure, black-market LSD may contain small or large quantities of ergot alkaloids, which are psychologically inert but physiologically active and perhaps capable of modifying the LSD experience.

The usual dose given by mouth begins to act in one to two hours, reaches its maximum effect in two to four hours, and is usually gone in about six to eight hours. Usually, all LSD effects are terminated by one normal period of sleep. LSD is preferred in psychedelic therapy because the reaction is initiated more predictably, is easier to control, and disappears more regularly than is the case with other hallucinogens. All hallucinogens suffer from the disadvantage that, once the reaction has started, it must proceed through the usual time sequence unless terminated by administration of tranquillizers or massive doses of vitamins.

Cohoba. Cohoba was a narcotic snuff used by natives of Haiti and of Central and South America for many centuries. It was snuffed as deeply as possible into the lungs using bifurcated tubes, although among some tribes individuals blew snuff into one anothers' nostrils using single tubes.

Cohoba was prepared from parts of the tree *Piptadenia peregrina* and contained bufotenine, DMT, and other indole derivatives. Bufotenine is also found in fly agaric mushrooms used as a hallucinogen by natives in northeast Siberia. And it is also present in the skin of certain toad species.

Bufotenine is a weak hallucinogen, active chiefly by intravenous injection. Since it is a powerful vasoconstrictor, the dose must be small. DMT is a much more powerful hallucinogen, and it too must be given by injection (or sniffed). The reaction begins in about five minutes and is over in about an hour. Diethyltryptamine (DET), a related synthetic, begins to act in about 15 minutes, and the experience continues for about 3 hours. The hallucinogenic properties of DMT are particularly important because the substance has been found in the body fluids of schizophrenics, but not yet in normal subjects. DMT could be formed by abnormal transmethylation of normal tryptophan metabolites.

Psilocybin. Chemically, psilocybin is O-phosphoryl-4-hydroxy-N,N-dimethyltryptamine; it is the psychoactive ingredient in the hallucinogenic mushroom *Psilocybe.* Removal of the phosphate group leaves the substance psilocin. The mushroom has been used in the southern United States, Mexico, and Central America—not regularly, as is the case with peyote by the Native American Church of North America, but on special occasions and to serve special needs.

Psilocybin and psilocin produce identical hallucinogenic experiences that are indistinguishable from those produced by mescaline and LSD. In addition, these hallucinogens are cross tolerant—that is, when tolerance is developed to any one of them, increased dosage of any of the others is needed to invoke a response. In contrast, however, DMT is not cross tolerant with the major plant hallucinogens.

Psilocybin is usually taken sublingually. The reaction appears in about half an hour, reaches its maximum in an hour, and is usually gone in about four hours. Psilocybin is thus somewhere between DMT and LSD in its duration of effect. It may be given parenterally, when of course it reacts much more quickly.

Ibogaine. This drug is an indole present in the woody plant *Tabernanthe iboga.* It was used in West Africa and the Congo in small doses as a stimulant and in large doses as a narcotic. Natives chewed the root of the plant. It has been studied largely in animals, where it is shown to be a stimulant.

Harmine. Harmine is also an indole found in the plant *Banisteriopsis caapi.* It was used by natives in the area drained by the tributaries of the upper Amazon. The drug was extracted from small pieces of wood by boiling with water for from 2 to 24 hours. As was usually the case with native hallucinogens, harmine was used for religious purposes or as a medicine. In experimental studies, harmine has produced the same kind of experience as LSD or mescaline.

Marijuana. Δ^1-Tetrahydrocannabinol, found mainly in the resin of unripened fruit and adjacent leaves of the hemp plant *Cannabis sativa,* is one of the oldest known and most widespread hallucinogens. It is native to Central Asia but is grown around the world. The more tropical the area in which it is grown the more active is the material. The drug is used in a variety of ways and under many different names throughout the world. In the Middle East and North Africa resin from the plant, called hashish or kif, is ingested. In Western countries the plant is dried and incorporated into cigarettes for smoking. In this form it is called marijuana or any of a number of colloquial names.

In one of the best studies of the effects of the drug, 250 to 400 milligrams of resinous extract of powdered *Cannabis sativa* were given to ten normal subjects. The experience began suddenly, within the first hour, and was comparable to that induced by LSD; characteristic perceptual changes occurred, including illusions and visual hallucinations when the eyes were closed. Temporal disorientation was marked in all subjects, as were distortions of the body image. There were disorders in thought content (delusions) and thought processes (*e.g.*, blocking, slowing, inability to grasp meanings, loss of memory). The moods induced were chiefly euphoria and a sense of detachment, although in some cases anxiety and depression were felt. Physical changes included suffusion of conjunctiva of the eye after one to two hours, increase in pulse rate, moderate coldness of the extremities, and nausea after three hours.

As with all hallucinogens, the effect of marijuana depends upon the potency of the preparation, the mode of administration, the degree of experience of the user, and other factors. Pure tetrahydrocannabinol is the most potent form of the drug. Generally, marijuana is not addictive in a narcotic sense, but many subjects come to depend on it, and marijuana psychoses have been reported.

Drugs with hallucinogenic properties. Cholinesterase inhibitors are not necessarily hallucinogenic; but several hallucinogens have strong cholinesterase-blocking activity, and most of the hallucinogens already described have moderate to weak anticholinesterase activity. Known anticholinesterase drugs may have undetected hallucinogenic properties. Thus, physostigmine (also known as eserine), found in Calabar beans, is not a hallucinogen, but this may be due to its intense activity, which limits the dose. In an analogous way, adrenalin is such a powerful vasopressor that its psychological properties cannot be readily studied. Prostigmin is normally not hallucinogenic, but it has sometimes produced psychotic episodes in patients treated for myasthenia gravis.

The belladonna alkaloids are the best known parasympatholytic hallucinogens. The chief members of the family, atropine and scopolamine, are found in deadly nightshade, black henbane, and Jimsonweed; they have been used as poisons and hallucinogens since the dawn of history. In modern times they have found a variety of pharmacological uses in anesthesia, opthalmology, and the management of respiratory and other diseases.

Sufficient atropine has been absorbed from instillations in the eye to produce hallucinations, especially in children. Atropine psychoses are infrequent now but were fairly common at one time. A person admitted to the hospital in a psychotic state with fever and widely dilated pupils often turned out to have smoked excessive quantities of anti-asthma cigarettes, which contained belladonna alkaloids.

An interesting hallucinogen in larger doses is benactyzine, which was originally introduced as an antidepressant. The hallucinogenic dose is about 20 times the antidepressant dose; the effect produced is similar to that of LSD or mescaline.

A series of piperidyl compounds are potent hallucinogens, but the psychological experience tends to be more of a delirium than that induced by the other hallucinogens. They are typical anticholinergic chemicals and produce large pupils, dry mouth, flushed skin, and slight hypertension.

Metabolites with hallucinogenic properties. The amines dopamine, norepinephrine (noradrenalin), and epinephrine (adrenalin) are readily oxidized in the body to dopachrome, noradrenochrome, and adrenochrome. They may be further changed to adrenolutins, yellow fluorescent compounds that have been identified in the body but have not been isolated from tissues or body fluids. These are potentially very important as intermediates in schizophrenic pathology—*i.e.*, they could provide the link between vitamin B_3 dependency and schizophrenia.

Only synthetic preparations of adrenochrome and adrenolutin have been studied for hallucinogenic properties. The experience they induce resembles that of LSD. In low doses there are minor perceptual changes, alterations in thought, and depression. There have been no instances of euphoria. The experience has always been psychotomimetic. In large doses they produce LSD-like perceptual changes. Activity depends upon the mode of administration.

BIBLIOGRAPHY. R. ALPERT, S. COHEN, and L. SCHILLER, *LSD* (1966), a discussion of whether hallucinogens should be used on a larger scale (because of their potential benefit to society) or limited to a clinical setting; T.A. BAN, *Psychopharmacology* (1969), one of the most informative and best balanced textbooks on psychopharmacology; A. HOFFER and H. OSMOND, *The Hallucinogens* (1967), a work written primarily for scientists, but readable by the layman; H. KLUVER, *Mescal: The "Divine" Plant and its Psychological Effects* (1928), the first thorough study of the psychological effects of mescaline; L. LEWIN, *Phantastica: Narcotic and Stimulating Drugs* (1964), the first work on the hallucinogens, published in German in 1924 and in English in 1931; H. OSMOND, "A Review of the Clinical Effects of Psychotomimetic Agents," *Ann. N.Y. Acad. Sci.*, 66:418–434 (1957), a thorough statement of the author's concepts; R.E. SCHULTES, "Hallucinogens of Plant Origin," *Science*, 163:245–254 (1969), an excellent brief review of plant hallucinogens; L.P. SOLURSH, "Nonmedical Use of Drugs," *Canad. Med. Assn. J.*, 101:72–88 (1969), a brief presented by the Canadian Medical Association to the Government of Canada's Commission of Inquiry into the nonmedical use of drugs, providing an excellent general review of many of the hallucinogens, particularly marijuana; V.P. and R.G. WASSON, *Mushrooms, Russia and History* (1957), a limited printing of an extraordinary book, which provided the first major descriptions of *Psilocybe* and other psychologically active mushrooms and their impact on culture.

(A.Ho.)

Halogen Elements and Their Compounds

Fluorine (chemical symbol F), chlorine (Cl), bromine (Br), iodine (I), and astatine (At) are the members of the family of halogen elements, and they constitute Group VIIa of the periodic table (see Figure). They were given the name halogen from the Greek roots *hal-*, "salt," and *gen*, "to produce," because they all produce sodium salts of similar properties, of which sodium chloride, table salt, is the best known. Because of their great reactivity, the free halogen elements are not found in nature. In combined form, fluorine is the most abundant of the halogens in the earth's crust. The percentages of the halogens in the igneous rocks of the Earth's crust are 0.06 fluorine, 0.031 chlorine, 0.00016 bromine, and 0.00003 iodine. Astatine does not occur normally on Earth because it consists only of short-lived radioactive isotopes (forms of the element that differ in mass).

The halogen elements show great resemblances to one another in their general chemical behaviour and in the properties of their compounds with other elements. There is, however, a progressive change in properties from fluorine through chlorine, bromine, and iodine to astatine—the difference between two successive elements being most pronounced with fluorine and chlorine. Fluorine is the most reactive of the halogens and, in fact, of all elements, and it has certain other properties that set it apart (see below *General properties*).

Chlorine is the best known of the halogen elements. The free element is widely used as a water purification agent, and it is employed in a number of chemical processes. Sodium chloride, of course, is one of the most familiar of chemical compounds. Fluorides are known chiefly for the controversy over their addition to water to prevent tooth decay, but organic fluorides are also used as refrigerants and lubricants. Iodine is most familiar as an antiseptic, and bromine is used chiefly to prepare a gasoline additive (ethylene dibromide) that prevents deposits of lead in engines.

History

Rock salt (common salt, sodium chloride) has been known for several thousand years; it is the main constituent of the salts dissolved in seawater, from which it was obtained in ancient Egypt by evaporation. In 1648 the German chemist Johann Rudolf Glauber obtained a strong acid, which he called spirit of salt, by heating moist salt in a charcoal furnace and condensing the fumes in a receiver. Later he obtained the same product, now known to be hydrochloric acid, by heating salt with sulfuric acid.

In 1774 the Swedish chemist Carl Wilhelm Scheele treated powdered black oxide of manganese with hydrochloric acid and obtained a greenish-yellowish gas, which he failed to recognize as an element. The true nature of the gas as an element was recognized in 1810 by the British chemist Sir Humphry Davy, who later called it chlorine.

In 1811 the French chemist Bernard Courtois obtained a violet vapour by heating seaweed ashes with sulfuric acid. This vapour condensed to a black crystalline substance, which he called "substance X." In 1813 Davy, who was passing through Paris on his way to Italy, recognized substance X as an element analogous to chlorine; he suggested the name iodine.

Discovery of bromine

Bromine was discovered in 1826 by the French chemist Antoine-Jérôme Balard in the residues from the manufacture of sea salt at Montpellier. He liberated the ele-

ment by passage of chlorine through aqueous solution of the residues, which contained magnesium bromide. Distillation of the material with manganese dioxide and sulfuric acid produced red vapours, which were condensed to a dark liquid. The similarity of this procedure to that for making chlorine suggested to Balard that he had obtained a new element similar to chlorine. (The German chemist Justus von Liebig appears to have obtained the element before Balard, but he wrongly considered it to be iodine chloride.)

The fluorine-containing mineral fluorspar (fluorite) was described in 1529 by Georg Bauer, the German physician and mineralogist better known as Agricola. It appears likely that crude hydrofluoric acid was first prepared by an unknown English glassworker in 1720. In 1771 Scheele obtained hydrofluoric acid in an impure state by heating fluorspar with concentrated sulfuric acid in a glass retort, which was greatly corroded by the product; as a result, vessels made of metal were used in subsequent experiments with the substance. The nearly anhydrous acid was prepared in 1809, and two years later the French physicist André-Marie Ampère suggested that it was a compound of hydrogen with an unknown element, analogous to chlorine, for which he suggested the name fluorine. Fluorspar was then recognized to be calcium fluoride.

The isolation of fluorine was for a long time one of the chief unsolved problems in inorganic chemistry, and it was not until the year 1886 that the French chemist Henri Moissan prepared the element by electrolyzing a solution of potassium hydrogen fluoride in hydrogen fluoride. The difficulty in handling the element and its toxic properties contributed to the slow progress in fluorine chemistry. Indeed, up to the time of World War II the element appeared to be a laboratory curiosity. Then, however, the use of uranium(VI) fluoride in the separation of uranium isotopes and the development of organic fluorine compounds of industrial importance made fluorine an industrial chemical of considerable use. Astatine was prepared for the first time in 1940 by bombarding bismuth with alpha particles.

Comparative chemistry of the halogen compounds

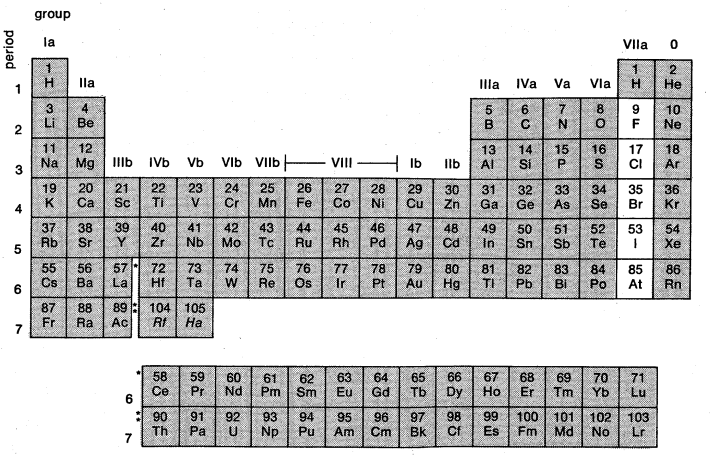
GENERAL PROPERTIES

Oxidizing property. Probably the most important generalization that can be made about the halogen elements is that they are all oxidizing agents; *i.e.*, they raise the oxidation state, or oxidation number, of other elements —a property that used to be equated with combination with oxygen but that is now interpreted in terms of transfer of electrons from one atom to another. In oxidizing another element, a halogen is itself reduced—*i.e.*, its oxidation number is reduced from 0, the typical state for a free element, to −1. In this state the halogens combine with other elements to form compounds known as halides—namely, fluorides, chlorides, bromides, iodides, and astatides. Many of the halides may be considered to be salts of the respective hydrogen halides, which are colourless gases at room temperature and atmospheric pressure and which are known to exhibit acidic properties. Indeed the general term salt is derived from rock salt, or table salt (sodium chloride). The tendency of the halogen elements to form saltlike compounds increases in the following order: astatine, iodine, bromine, chlorine, and fluorine.

Halides

This order is also that of increasing oxidizing properties of the free halogen elements. Fluorides are usually more stable than the corresponding chlorides, bromides, or iodides. (Often astatine is omitted from general discussions of the halogens because it is less well known than the other elements.) Conversely, of the halogen elements, fluorine is prepared in the free state with the greatest difficulty and iodine with the least. As a class, the halogen elements are nonmetals, but astatine shows certain properties resembling those of the metals.

Electronic structure. The chemical behaviour of the halogen elements can be discussed most conveniently in terms of their position in the periodic table of the elements (see Figure) In the periodic table the halogens



Halogen elements in the periodic table.

comprise Group VIIa, the group immediately preceding the noble gases. It is characteristic of the atoms of the halogen elements that each carries seven electrons in its outermost orbitals (an orbital is the space suitable for electrons). The seven outermost electrons of an atom of a halogen element are in two different kinds of orbitals, designated *s* (with two electrons) and *p* (with five). Potentially, a halogen atom could hold one more electron (in a *p* orbital), which would give the electrons the same arrangement (configuration) as that of the noble gas next to it in the periodic table, the noble gases all having exceptionally stable configurations of electrons in their atoms. It is in acquiring such an electron that an atom of a halogen element acts as an oxidizing agent (or is itself reduced). In the process, the atom acquires an electrical charge (in this case a negative one) and becomes a charged particle or ion. Halogen elements exist in their salts as halide ions, which are relatively large ions, colourless and extremely stable.

At room temperature and atmospheric pressure the halogen elements in their free states exist as diatomic molecules—molecules composed of two atoms each. In molecular fluorine the atoms are held together by a bond made from the union of a *p* orbital from each atom, such a bond being classed as a sigma bond. Beginning with chlorine, however, the halogen atoms are able to make use of different orbitals, called *d* orbitals—an option that is not open to the fluorine atoms. By forming partial bonds with their *d* orbitals, chlorine and the other halogens (with the exception of fluorine) form diatomic molecules held in part by multiple bonds (called pi bonds), which are stronger than single bonds, such as those that hold together the atoms in a fluorine molecule.

Diatomic halogen molecules

RELATIVE REACTIVITY

The great reactivity of fluorine largely stems from the relatively low energy of the fluorine–fluorine bond (37.7 kilocalories per mole, a standard measure for bond energies). The chlorine–chlorine bond is considerably stronger (57 kilocalories per mole) because of its partial multiple-bond character.

Fluorine and chlorine are gases at room temperature. Bromine is a reddish-brown liquid at room temperature, and is—apart from mercury—the only element that is liquid at 20° C (68° F) and atmospheric pressure. Iodine forms dark-violet crystals under these conditions. In the solid state the halogen elements form molecular lattices —extended weblike structures including a great many separate units. The lattice energies increase with increasing size of the molecules. The principal properties of the halogen elements are noted in Table 1.

The energy released in the formation of an ion from a free atom and an electron (brought up from an infinite distance) is called the electron affinity. The electron affinities for the halogen elements are all high and show only slight differences from one another. It is known, however, that the oxidizing properties (ability to take up an electron) increase from astatine to fluorine, and this

Electron affinities

Table 1: Properties of the Halogen Elements

	fluorine	chlorine	bromine	iodine	astatine
Atomic number	9	17	35	53	85
Atomic weight	18.9984	35.453*	79.904*	126.904	(210)
Colour of element	light greenish yellow	greenish yellow	brown-red	dark violet	—
Melting point (°C)	−219.62	−100.98	−7.25	113.5	~ 300
Boiling point (°C)	−188.14	−34.05	58.8	184.4	~ 370
Density (760 mm Hg)					
Gas (g/l)	1.696 (0° C)	3.214 (0° C)	5.6 (175.5° C)	6.75 (185° C)	—
Liquid (g/cm³)	1.108 (−188° C)	1.655 (−70° C)	3.12 (20° C)	3.960 (120° C)	—
Solid (g/cm³)	1.32 (0° K)	2.17 (−195° C)	4.17 (0° K)	4.930 (20° C)	—
Solubility in water	(reacts)	2.3 (vol/vol 20° C) carbon tetrachloride (reacts) 8.48% (by weight, 20° C) chloroform (280 g/1,000g, 0° C)	3.41 (g/100 g, 20°) completely miscible	0.293 (g/1,000 g aq. iodide solution completely miscible 19.2 (g/1,000g) completely miscible 49.7 (g/1,000g)	—
Oxidation numbers	−1	−1, +1, +3, +4, +5, +6, +7	−1, +1, +3, +4, +5, +7	−1, +1, +3, +4, +5, +6, +7	−1, +1, +3, +5, +7
Electronic configuration	(He) $2s^22p^5$	(Ne) $3s^23p^5$	(Ar) $3d^{10}4s^24p^5$	(Kr) $4d^{10}5s^25p^5$	(Xe)$4f^{14}5d^{10}6s^26p^5$
Isotopic abundance (terrestrial percent)	^{19}F (100)	^{35}Cl (75.53) ^{37}Cl (24.47)	^{79}Br (50.54) ^{81}Br (49.46)	^{127}I (100)	—
Radioactive isotopes (mass numbers)	17, 18, 20–22	32–34, 36, 38–40	74–78, 80, 82-90	117–126, 128–139	200–219
Heat of fusion (kcal/mol)	0.244	1.53	2.58	3.65	—
Heat of vaporization (kcal/mol)	1.561	4.863	7.159	9.889	—
Heat of hydration of X^- (kcal/mol)	120.8	88	80.3	70.5	—
Specific heat (kcal/g/°C at 25°C)					
Constant pressure	0.1842 (17° C)	0.114	0.0537	0.034	—
Constant volume		0.0849	0.04257	0.02697	—
Critical temperature (°C)	−129	144	302	512	—
Critical pressure (atm)	56.8	77.3	126	116†	—
Critical density (g/cm³)	0.63	0.567	1.064	1.336	—
Crystal structure	—	orthorhombic	orthorhombic	orthorhombic	—
Electrical resistivity (microohms-cm)	—	~ 10^{10} (near freezing point)	6.5 × 10^{10} (25° C)	5.85 (25° C)	—
Magnetic susceptibility (cm³/g)					
Gas	−3.4 × 10^{-7}	−18.7 × 10^{-10} (0° C, 760 mm Hg)	−73.5 × 10^{-6}	−3.9 × 10^{-7} (118° C)	—
Liquid		−5.9 × 10^{-7} (−16° C)	−56.4 × 10^{-6}	−88.7 × 10^{-6} (sol., 28° C)	—
Radius					
Ionic (Å)	1.33	1.81	1.96	2.20	~ 2.27
Covalent (Å)	0.71	0.99	1.14	1.33	~ 1.4
Bond energy (kcal/mol)	37.7	58.0	46.1	36.1	—
Ionization energy (kcal/mol)	402	300	273	241	—
Electron affinity (kcal/mol)	79.5	83.3	77.5	70.6	—
Electronegativity (Allred-Rochow)	4.10	2.83	2.74	2.21	1.96

*Variation of isotopic abundance in terrestrial samples limits the precision of the atomic weight given. †Estimated.

increase is reflected in the position of the respective oxidation–reduction systems in the electromotive series (a standard arrangement of oxidation–reduction systems in water in order of oxidizing strength).

The fluoride ion is extremely stable toward oxidation; the iodide ion is a mild reducing agent. This difference between the ions is caused by two factors: (1) a decreasing tendency for the halide ion to act as an electron-pair donor (a so-called Lewis base) in the series fluoride, chloride, bromide, iodide, and astatide, leading to a decrease in stability of the hydrated form of the ion (*i.e.*, the free ion surrounded by water molecules) in the same order; and (2) a decrease in electron pair acceptor (Lewis acid) properties of the free halogens in the reverse order, namely, astatine, iodine, bromine, chlorine, and fluorine.

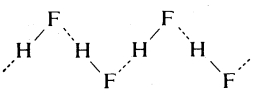
Within a molecule in which atoms are held together by a shared electron pair—*i.e.*, by a covalent or nonionic bond—the tendency of an atom to attract the shared electrons may be expressed by an electronegativity value. Fluorine has the greatest electronegativity value (is the most electronegative) of all elements (not just the halogens), and there is a decrease in electronegativity within the family of the halogen elements from fluorine through chlorine, bromine, and iodine to astatine.

Fluorine replaces any other halide ion from its compounds, as shown in the following equations (in which atoms of the elements are represented by their chemical symbols; number of atoms of a given kind in a molecule are indicated by subscripts; and positive and negative charges on ions are indicated by superscript plus and minus signs; as is customary in chemical equations the reactants are written to the left and the products to the right; the double arrows indicate that both the forward reaction and the reverse reaction occur).

$$F_2 + 2Cl^- \rightleftharpoons 2F^- + Cl_2$$
$$F_2 + 2Br^- \rightleftharpoons 2F^- + Br_2$$
$$F_2 + 2I^- \rightleftharpoons 2F^- + I_2$$
$$F_2 + 2At^- \rightleftharpoons 2F^- + At_2.$$

Chlorine, however, replaces only bromide, iodide, and astatide ions, and bromine only iodide and astatide ions. Free fluorine, chlorine, bromine, and iodine are expected to replace astatide ions.

The halogen elements all form compounds with hydrogen—the hydrogen halides. The energy of the hydrogen–halogen bond increases strongly from iodide to fluoride. Hydrogen fluoride in the crystalline state consists of infinite zigzag chains, as shown in the diagram



in which H represents the hydrogen atoms and (as before) F the fluorine atoms; the solid lines represent covalent bonds between the hydrogen and fluorine atoms within the molecules, and the dotted lines represent secondary bonds, called hydrogen bonds. The hydrogen bonds between hydrogen fluoride molecules are considerably weaker (seven kilocalories per mole) than those within the molecules (135 kilocalories per mole), yet they are retained to a great extent in the liquid state. Similar hydrogen bonding exists in the other hydrogen halides, but it is considerably weaker. The large difference in hydrogen bonding between hydrogen fluoride and the other hydrogen halides accounts for the relatively high melting and boiling points of hydrogen fluoride as compared

Table 2: Comparison of Properties of Hydrogen Halides

hydrogen halide	formula	melting point (°C)	boiling point (°C)	homolytic dissociation energy (kcal/mol)	heterolytic dissociation energy (kcal/mol)	dipole moment (D)	H–X distance (angstrom)	K_a* in water at 20° C
Fluoride	HF	−83.1	+19.5	135	340	1.91	0.92	7×10^{-4}
Chloride	HCl	−114.8	−84.9	103.1	331	1.03	1.28	$\sim 10^{7}$
Bromide	HBr	−88.5	−67.0	87.4	322	0.78	1.41	$\sim 10^{9}$
Iodide	HI	−50.8	−35.4	71.4	312	0.38	1.61	$\sim 10^{11}$

*K_a = acid dissociation constant.

to those of hydrogen chloride and the other hydrogen halides. The hydrogen–halogen bond energies also decrease considerably in going from hydrogen fluoride to hydrogen iodide. These and other properties of the hydrogen halides are compared in Table 2.

Hydrogen halides as acids

In aqueous solution the hydrogen halides are acids, as reflected in their common names; such as hydrofluoric acid for hydrogen fluoride, and so on. The acidity increases dramatically from hydrofluoric acid to hydrochloric acid. Further increases occur with the other hydrogen halides, and hydriodic acid is not only the most acidic of the hydrogen halides but one of the most acidic substances known in aqueous solution. On the other hand, the oxidizing properties of the hydrogen halides decrease markedly from hydrogen fluoride to hydrogen iodide (thereby accounting for the extreme reactivity of hydrofluoric acid toward certain substances).

The ionization potentials of the halogens are generally high, but they fall markedly with atomic number. Fluorine is the only halogen that does not form compounds with positive oxidation states—*i.e.*, states in which it has lost, rather than gained, electrons. This property is related to the inability of the fluorine atom to use *d* orbitals in bonding (see above), an inability not shared by the other halogen elements. Iodine forms an appreciable number of compounds in which a unipositive iodine ion is stabilized by coordination (surrounding with other groups); for example, with the coal-tar base, pyridine. The same situation has also been found with astatine.

Whereas fluorine exhibits only the single oxidation state of −1, the principal oxidation states of chlorine, bromine, and iodine are −1, +1, +3, +5, and +7. The oxo acids are compounds in which halogen atoms are joined to oxygen atoms and show large positive oxidation numbers. The oxo acids are all powerful oxidizing agents, being reduced to the corresponding aqueous hydrogen halides—the oxidation numbers passing from +5 or +7 to −1 in the process. The oxidizing properties of the oxo acids, however, fall with increasing oxidation number of the halogen because the coordination number increases with increasing oxidation number and the ion is stabilized by the coordination.

All the molecules and ions in which halogen atoms employ four electron pairs in bonding are basically tetrahedral in shape—as is, for example, the perchlorate ion $(ClO_4)^-$. Those employing five electron pairs are square pyramidal in structure, such as iodine(V) fluoride (IF_5)—the roman numeral in the name being used to indicate the oxidation number of the element it follows. Those with six electron pairs are octahedral; *e.g.*, the paraperiodate ion, $(IO_6)^{5-}$. The unique binary compound iodine(VII) fluoride (IF_7) is believed to have a pentagonal bipyramidal arrangement of fluorine atoms.

The maximum coordination number of chlorine(VII) toward oxygen is 4 (*i.e.*, the chlorine atom is surrounded by four oxygen atoms) in the perchlorate ion, $(ClO_4)^-$, whereas that of iodine(VII) is 6 in the paraperiodate ion, $(IO_6)^{5-}$. Analagous behaviour is found in the elements preceding the respective halogen in the periodic table, sulfur and tellurium.

Individual halogen group elements and their compounds

FLUORINE

Occurrence and distribution. The fluorine-containing mineral fluorspar has been used for centuries as a flux (cleansing agent) in various metallurgical processes. The name fluorspar is derived from the Latin *fluere,* "to flow." The mineral subsequently proved to be a source of the element, which was named fluorine accordingly. The colourless, transparent crystals of fluorspar exhibit a bluish tinge when illuminated, and this property is accordingly known as fluorescence.

Fluorspar —source of fluorine

Fluorine is found in nature only in the form of its chemical compounds, except for trace amounts of the free element in fluorspar that has been subjected to the action of radiation from radium. The principal fluorine-containing minerals are: (1) fluorspar (fluorite, CaF_2), deposits of which occur in Illinois, Kentucky, Derbyshire, southern Germany, the south of France, and the Soviet Union; (2) cryolite (Na_3AlF_6), chiefly from Greenland; (3) fluoroapatite ($Ca_5[PO_4]_3[F,Cl]$), widely distributed and containing variable amounts of fluorine and chlorine; (4) topaz $Al_2SiO_4[F,OH]_2$), the gemstone; and (5) lepidolite, a mica as well as a component of animal bones and teeth.

Production and use. Fluorspar is the most important source of fluorine. In the manufacture of hydrogen fluoride, powdered fluorspar is distilled with concentrated sulfuric acid in a lead or cast-iron apparatus. The hydrogen fluoride is obtained in a fairly anhydrous state by fractional distillation in copper or steel vessels and stored in steel cylinders. The usual impurities in commercial hydrogen fluoride are sulfurous and sulfuric acids as well as fluorosilicic acid, arising from the presence of silica in the fluorspar. Traces of moisture may be removed by electrolysis with platinum electrodes.

The preparation of the free element is carried out by electrolytic procedures in the absence of water. Generally these take the form of electrolysis of a melt of potassium fluoride–hydrogen fluoride (in a ratio of 1 to 2.5–5) at temperatures between 30° and 70° C (86° and 158° F), 80° and 120° C (176° and 248° F), or at 250° C (482° F). During the process the hydrogen fluoride content of the electrolyte is decreased and the melting point rises; it is therefore necessary to add hydrogen fluoride. In the high-temperature cell the electrolyte is replaced when the melting point reaches over 300° C (572° F). Fluorine can be safely stored under pressure in cylinders of stainless steel if the valves of the cylinders are free from traces of organic matter.

The element is used for the preparation of various fluorides, such as chlorine trifluoride, sulfur(VI) fluoride, or cobalt(III) fluoride. The cobalt compound and certain other metal fluorides are important fluorinating agents for organic compounds. With appropriate precautions, the element itself may be used for the fluorination of organic compounds. Fluorine derivatives of hydrocarbons (compounds of carbon and hydrogen) are useful as refrigerants and as lubricants. The element is also used for the preparation of uranium(VI) fluoride, which is important in the process of separation of uranium-235 from natural uranium. Its outstanding oxidizing properties make elemental fluorine of possible interest as an oxidizer in rocket propulsion. Hydrogen fluoride and boron(III) fluoride are produced commercially because they are good catalysts for the alkylation reactions used to prepare organic compounds of many kinds. The addition of sodium fluoride to drinking water to reduce the incidence of dental caries in children is a matter of some controversy.

Uses in fluorination

Chemical properties. At room temperature fluorine is a pale-yellow gas with an irritating odour. Inhalation of

the gas is dangerous except in very low concentration. There is only one stable isotope of the element—fluorine-19.

Because fluorine is the most electronegative of the elements, atomic groupings rich in fluorine are negatively charged. Methyl iodide (CH_3I) and trifluoroiodomethane (CF_3I) have different properties for that reason. The charge distributions in these molecules are shown in the following formulas, in which the Greek symbol δ (delta) indicates a partial charge:

$$\overset{\delta+\ \delta-}{CH_3I} \qquad \overset{\delta-\ \delta+}{CF_3I}.$$

The ionization potential of fluorine is very high (402 kilocalories per gram atom), giving a standard heat formation for fluoride ion of 420 kilocalories per mole.

The small size of the fluorine atom makes it possible to pack a relatively large number of fluorine atoms or fluoride ions around a given coordination centre—*i.e.*, around a central atom. Fluorine is the most powerfully oxidizing element. No other substance, therefore, is able to reduce the fluoride ion to the free element, and for this reason the element is not found in the free state in nature. Furthermore, all chemical methods have failed to produce the element, success having been achieved only by the use of electrolytic methods. The high oxidizing properties of fluorine allow the element to produce the highest oxidation numbers possible in other elements, and many higher fluorides of elements are known for which there are no other corresponding halides—*e.g.*, silver(II) fluoride, cobalt(III) fluoride, rhenium(VII) fluoride, bromine(V) fluoride, and iodine(VII) fluoride.

Fluorine reacts with nearly all other elements at room temperature. Some metals, such as nickel, are quickly covered by a fluoride layer, which prevents further attack of the metal by the element. When dry, certain metals, such as mild steel, copper, aluminum, or Monel metal, are not attacked by fluorine to any extent at ordinary temperatures. For work with fluorine at temperatures up to 600° C (1,100° F) Monel metal is suitable; sintered alumina is resistant up to 700° C (1,300° F). When lubricants are required, fluorocarbon oils are most suitable.

Fluorine reacts violently with organic matter (such as rubber, wood, and cloth), and controlled fluorination of organic compounds by the action of elemental fluorine is not possible unless special precautions are taken.

Principal compounds. Because fluorine has a high electron affinity, it forms many ionic compounds. In contrast to aluminum(III) chloride, which is covalent, aluminum(III) fluoride is composed of an ionic lattice. In general, a metal–fluorine bond is considerably stronger than the corresponding metal–chlorine bond. This strength is one reason why a fluoride compound is dissociated to a smaller extent than the corresponding chloride in a given solvent, the other reason being the "hard" properties (small polarizability) of the fluoride ion.

Small size of the fluorine ion

Because of the small size of the fluoride ion—nearly the same as that of the oxide ion—and the resulting tendency to acquire a high coordination number, hexafluorides of nonmetals—sulfur(VI) fluoride—or of metals—molybdenum(VI) fluoride and uranium(VI) fluoride—are readily formed. They are rather inert and volatile compounds, which form molecular lattices in the solid state. Three heptafluorides are known; namely, technetium (VII) fluoride, rhenium(VII) fluoride, and iodine(VII) fluoride. In certain complex fluorides even higher coordination numbers are found; for example, potassium octafluorotantalate(V) (K_3TaF_8).

The tendency of fluorine to form unique compounds with various metals in their highest possible coordination numbers (caused by its great oxidizing ability) is further assisted by the small size of the fluorine atom (and the fluoride ion). Bismuth(V) fluoride, for example, is the only known compound of bismuth with all the bismuth atoms in the +5 oxidation state; and cobalt(III) fluoride, a very useful fluorinating agent, is the only binary halide of cobalt in the +3 state. Cesium hexafluorocobaltate (IV) (Cs_2CoF_6) is readily formed by oxidizing cesium tetrachlorocobaltate(II) (Cs_2CoCl_4) with elemental fluorine at 300° C (572° F). Potassium hexafluoronickelate(IV) (K_2NiF_6) may be obtained in a similar fashion. Although silver(II) fluoride is the only known binary halide of silver(II), it may be oxidized further with fluorine in the presence of a fluoride able to form complexes. Thus in the presence of a potassium salt, silver(II) fluoride gives the compound potassium tetrafluoroargentate(III) ($KAgF_4$). Further examples are the fluoro complexes of the general composition Cs_3MF_7, in which $M = Pr^{4+}$ (praseodymium), Nd^{4+} (neodymium), Tb^{4+} (terbium), and Dy^{4+} (dysprosium). Fluorine reacts even with the relatively inert noble gas xenon to give the compounds xenon(II) fluoride, xenon(IV) fluoride, and xenon(VI) fluoride.

The high electronegativity of fluorine suggests that boron(III) fluoride should behave as a stronger Lewis acid (electron-pair acceptor) than boron(III) chloride. In fact, however, the fluoride acts as a weaker Lewis acid than the chloride, and this fact has been attributed to pi bonding between boron and the fluorine atoms. On coordination of boron(III) fluoride with a fourth fluoride ion, the tetrafluoroborate ion, $(BF_4)^-$, is formed, in which pi bonding is virtually impossible; and, indeed, in this complex ion the boron–fluorine distances are considerably greater than the simple trifluoride.

The "hard" fluoride ion is a fairly strong ligand (coordinating substance) toward similarly "hard" metal ions, whereas it acts as a weaker ligand toward "soft" metal ions, such as mercury(I) ion. In octahedral complexes, fluorine atoms exert a stronger ligand field (coordination effect) than any other halogen atoms.

Hydrogen fluoride. Pure hydrogen fluoride is prepared in the laboratory by heating dry potassium hydrogen fluoride in a copper or platinum vessel. In the liquid state hydrogen fluoride molecules are associated through hydrogen bonds. The pure liquid is almost a nonconductor of electricity; it dissolves many substances to yield conducting solutions. It has a high dielectric constant—*i.e.*, an ability to separate electric charges—and hence acts as a dissociating medium for ionic compounds. Precipitation reactions, such as that shown below between thallium fluoride (TlF) and potassium perchlorate ($KClO_4$), may be carried out in anhydrous hydrogen fluoride:

$$TlF + KClO_4 \rightleftharpoons TlClO_4 + KF.$$

Alcohols, ethers, ketones, and acids give conducting solutions containing fluoride ions; methyl alcohol, for example, reacts according to the following equation:

$$CH_3OH + HF \rightleftharpoons (CH_3OH_2)^+ + F^-.$$

Acetic acid is protonated (takes up a hydrogen ion) in the solvent, and thus acts as a base in it:

$$CH_3COOH + HF \rightleftharpoons (CH_3COOH_2)^+ + F^-.$$

In concentrated solution there is found the $(HF_2)^-$ ion, which is also a constituent of potassium hydrogen fluoride (KHF_2).

Most inorganic chlorides, bromides, and iodides are converted into the respective fluorides by reaction with hydrogen fluoride; for example, titanium chloride ($TiCl_4$) reacts according to the following equation:

$$TiCl_4 + 4HF \rightarrow TiF_4 + 4HCl.$$

These last reactions go to completion because the other hydrogen halides are more volatile than hydrogen fluoride, and they are nearly insoluble in the latter.

Hydrogen fluoride as a solvent

Hydrogen fluoride is used as a solvent for the reaction of fluorine with various organic compounds, and some of them may be electrolyzed in this solvent to give fluorocarbons. Hydrogen fluoride is much used as a catalyst in organic chemistry. Butadiene, isoprene, and styrene, for example, polymerize (*i.e.*, individual molecules join together to form long chains, or polymers) rapidly when treated with hydrogen fluoride to give useful products. One of the main industrial uses of hydrogen fluoride is as a catalyst for the alkylation (addition of alkyl groups) of aromatic compounds by alcohols, olefins, alkyl halides, esters, and ethers. Hydrogen fluoride has also been used extensively for the low-temperature alkylation of

paraffinic hydrocarbons, particularly for the production of branched-chain compounds for aviation gasoline.

Hydrogen fluoride reacts with silica glass to give silicon(IV) fluoride (SiF_4). Silica and aqueous hydrofluoric acid yield fluorosilicic acid (H_2SiF_6). Etchings made on glass with hydrogen fluoride gas are opaque; those made with the aqueous solution are transparent.

The binary fluorides may be formally considered to be salts of hydrofluoric acid. They fall into two main groups: (1) those of metals, which have ionic lattices and are characterized by low volatility and low solubility; and (2) those of the nonmetals and the transition elements (metals in the centre of the periodic table), which are volatile and frequently hydrolyzed by water and decomposed by organic solvents. Within a group of ionic fluorides, such as the alkali metal fluorides, certain properties vary in a regular manner. The lattice energy, and hence the melting point, decreases as the size of the cation (positive ion) increases; thus, cesium fluoride has a relatively low melting point. Because the free energies of solvation (a measure of ease of solubility) of the ions decrease less rapidly than the lattice energies, the solubility is higher for cesium fluoride than for rubidium fluoride. These properties of the fluorides of the alkali metals are summarized in Table 3.

Table 3: Properties of the Fluorides of the Alkali Metals

fluoride	formula	radius of cation (angstrom)	melting point (°C)	solubility at 20° C (g/100 ml water)
Sodium	NaF	0.97	988	4.22
Potassium	KF	1.33	846	49.6
Rubidium	RbF	1.47	775	130
Cesium	CsF	1.67	682	very soluble

In a series of fluorides of nonmetals in any single group in the periodic table the size–volatility relationship is the reverse of that for compounds of the alkali metals, as shown in Table 4. Intermolecular attractions (van der Waals forces) increase and volatility decreases with increasing molecular weight.

Table 4: Properties of the Fluorides of Carbon Group Elements

fluoride	formula	melting point (°C)	boiling point (°C)
Carbon	CF_4	−184	−128
Silicon	SiF_4	−97	−65
Germanium	GeF_4	—	−15 (4 atm)
Tin	SnF_4	sublimes at 700	

In gaseous fluorides the observed bond lengths are usually shorter than expected. The carbon–carbon distance in ethane, for instance, is 1.54 angstroms (one angstrom unit equals 10^{-10} metre) and the fluorine–fluorine distance in molecular fluorine is 1.42 angstroms, whereas the observed carbon–fluorine distance in carbon(IV) fluoride is only 1.36 angstroms. Similarly, the nitrogen–nitrogen and phosphorus–phosphorus distances in hydrazine and white phosphorus, respectively, are 1.48 and 2.21 angstroms, but the bond length in nitrogen(III) fluoride is about 1.40 and that in phosphorus(III) fluoride is 1.54 angstroms.

Halogen fluorides. Fluorine forms several binary compounds with the other halogens, in which the other, more electropositive, halogen may be in the oxidation state of +1, +3, +5, or +7; whereas the fluorine atom always has the oxidation number −1. The known compounds of this family are shown in Table 5.

No fluoride of astatine has been reported, but astatine(V) fluoride and astatine(VII) fluoride are thought to be capable of existence.

Halogen fluorides are made by direct combination of the halogen elements under suitable conditions. Increases in temperature favour the formation of higher fluorides, those with higher oxidation number. There is a gradual change in properties from fluorine through chlorine(I) fluoride and bromine(I) fluoride to iodine(I) fluoride, the last being unstable. On the other hand, iodine forms a heptafluoride, one of the very few known binary compounds with a coordination number of 7. Bonding in halogen fluorides is covalent. The structure of iodine(VII) fluoride has not yet been confirmed, but there are indications that the molecule has a pentagonal bipyramidal shape. The trifluorides have T-shaped molecules, and the pentafluorides appear to have square pyramidal structures.

Table 5: Halogen Fluorides

F_2	—	—	—
ClF	ClF_3	ClF_5	—
BrF	BrF_3	BrF_5	—
IF*	IF_3	IF_5	IF_7

*Unstable.

Halogen fluorides as fluorinating agents

All halogen fluorides are strong fluorinating agents, their reactivity increasing in order from the iodine to the chlorine compounds and also with increasing number of fluorine atoms in the molecule. Bromine(III) fluoride, for example, reacts explosively with water, wood, and rubber, and with incandescence even with asbestos and most organic solvents. Most metals, including many noble metals—such as gold—are converted into their highest fluorides upon reaction with halogen fluorides. With molybdenum, for example, the reaction is as follows:

$$Mo + 2BrF_3 \rightarrow MoF_6 + Br_2.$$

Gold(III) fluoride has been obtained by the thermal decomposition of the adduct formed from the metal and bromine(III) fluoride, as shown in the following equations:

$$Au + 2BrF_3 \rightleftharpoons BrF_2AuF_4 + \tfrac{1}{2}Br_2$$

$$BrF_2AuF_4 \rightleftharpoons BrF_3 + AuF_3.$$

Pure liquid bromine(III) fluoride is pale yellow in colour and may be used as an ionizing solvent for fluorides and as a reaction medium for the preparation of numerous complex fluorides. A self-ionization equilibrium has been assumed to exist in the pure liquid, as follows:

$$2BrF_3 \rightleftharpoons (BrF_2)^+ + (BrF_4)^-.$$

It has been suggested that fluoride bridging is a characteristic structural feature in the liquid. Compounds have been isolated that contain the planar tetrafluorobromate(III) ions $(BrF_4)^-$. One example—with silver fluoride—is:

$$AgF + BrF_3 \rightleftharpoons Ag^+ + (BrF_4)^-.$$

Hydrogen fluoride yields fluoride ions in bromine(III) fluoride solution, and—like alkali metal fluorides and silver fluoride—it must be considered as a base in this solvent.

Lewis acids such as antimony(V) fluoride (SbF_5) give adducts with bromine(III) fluoride, which ionize to form difluorobromine(III) ions, $(BrF_2)^+$, as shown below:

$$BrF_3 + SbF_5 \rightleftharpoons (BrF_2)^+(SbF_6)^-.$$

Reactions of the acid–base type yield complex fluorides; for example:

$$Ag^+(BrF_4)^- + (BrF_2)^+(SbF_6)^- \rightleftharpoons Ag^+(SbF_6)^- + 2BrF_3.$$

Chlorine(III) fluoride behaves in an analogous manner. Iodine(V) fluoride gives hexafluoroiodate(V) with alkali metal fluorides such as potassium fluoride:

$$KF + IF_5 \rightarrow K^+(IF_6)^-.$$

Oxides. As stated above, fluorine is the only halogen element that occurs solely in the −1 oxidation state. Thus, no oxo acid of fluorine is stable, and the hypofluo-

rites, such as the compound pentafluorosulfur hypofluorite (F_5SOF), are covalent and cannot be considered as salts of a hypofluorous acid. Binary oxides of fluorine, however, do exist, and this fact is not in conflict with the above statement since oxygen has a lower electronegativity than fluorine and assumes the positive oxidation state in the compounds. Oxygen difluoride (OF_2) is obtained by passing fluorine into a fairly concentrated solution of alkali hydroxide, as shown below:

$$2F_2 + 2OH^- \rightarrow 2F^- + OF_2 + H_2O.$$

The colourless gas oxygen difluoride has strong oxidizing properties, and it is decomposed by alkali hydroxide:

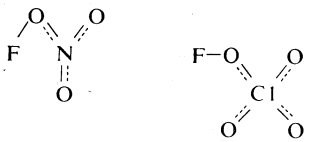
$$OF_2 + 2OH^- \rightleftharpoons 2F^- + O_2 + H_2O.$$

The highly unstable dioxygen difluoride (O_2F_2) is obtained in an electric-discharge tube cooled in liquid air. Reaction of dioxygen difluoride with antimony(V) fluoride gives a stable dioxygenyl compound $[O_2^+(SbF_6)^-]$. By admitting different ratios of oxygen and fluorine to an electric discharge, a number of unstable compounds can be obtained with molecular formulas, as follows: trioxygen difluoride (O_3F_2), tetraoxygen difluoride (O_4F_2), pentaoxygen difluoride (O_5F_2), and hexaoxygen difluoride (O_6F_2).

The compound known as fluorine nitrate (FNO_3) is prepared by the action of fluorine on concentrated nitric acid:

$$HNO_3 + F_2 \rightleftharpoons FNO_3 + HF.$$

The compound is a powerful oxidizing agent and, like the closely related compound fluorine perchlorate ($FClO_4$), is considered to be derivative of oxygen difluoride. These compounds are thought to be as follows:



in which the dotted lines indicate partial double-bond character.

Analysis. The accurate quantitative determination of the amount of fluorine in compounds is difficult. Free fluorine may be detected by its oxidizing action on chlorides such as sodium chloride, as shown:

$$F_2 + 2NaCl \rightleftharpoons 2NaF + Cl_2.$$

Qualitative tests for fluorine

The chlorine evolved may be determined by measurement of the volume of gas produced. The principal qualitative tests for the presence of fluorine are: (1) liberation of hydrogen fluoride by the action of sulfuric acid, (2) formation of a precipitate of calcium fluoride upon addition of a calcium chloride solution, and (3) decoloration of a yellow solution prepared from titanium(IV) oxide and hydrogen peroxide in sulfuric acid. Quantitative methods for analyzing fluorine are: (1) precipitation of calcium fluoride in the presence of sodium carbonate and treatment of the precipitate with acetic acid; (2) precipitation of lead chlorofluoride by addition of sodium chloride and lead nitrate; and (3) titration (determination of concentration of a dissolved substance) with thorium nitrate $[Th(NO_3)_4]$ solution using sodium alizarin sulfonate as indicator, according to the equation

$$Th(NO_3)_4 + 4KF \rightleftharpoons ThF_4 + 4KNO_3.$$

CHLORINE

Occurrence and distribution. When Sir Humphry Davy in 1810 recognized the elementary nature of the yellowish-green gas that had first been obtained by Scheele in 1774, he suggested the name chlorine from the Greek *chloros*, meaning "yellowish-green." Apart from very small amounts of free chlorine in volcanic gases, chlorine is found only in the form of chemical compounds. It constitutes 0.031 percent of the Earth's crust. The most common compound of chlorine is sodium chloride, which is found in nature as crystalline rock salt, often discoloured by impurities. Sodium chloride is also present in seawater, which has an average concentration of about 3 percent of that salt. Certain landlocked seas, such as the Dead Sea, contain up to 7.2 percent of dissolved salt. Besides sodium chloride, other metal halides, such as magnesium chloride, magnesium bromide, and—in small amount—certain sulfates, are contained in seawater. Small quantities of sodium chloride are present in blood and in milk. Other chlorine-containing minerals are sylvite (potassium chloride, KCl), bischofite ($MgCl_2 \cdot 6H_2O$), carnallite ($KCl \cdot MgCl_2 \cdot 6H_2O$), and kainite ($KCl \cdot MgSO_4 \cdot 3H_2O$). Free hydrochloric acid is present in the stomach.

Present-day salt deposits must have been formed by evaporation of prehistoric seas, the salts with the least solubility in water crystallizing first, followed by those with greater solubility. Because potassium chloride is more soluble in water than sodium chloride, certain rock salt deposits—such as those at Stassfurt, East Germany—were covered by a layer of potassium chloride. In order to gain access to the sodium chloride, the potassium salt, important as a fertilizer, is removed first.

Mining of rock salt deposits

Production and use. Rock salt deposits are usually mined; occasionally water is pumped down and brine, containing about 25 percent sodium chloride, is brought to the surface. When the brine is evaporated, impurities separate first and can be removed. In warm climates salt is obtained by evaporation of shallow seawater by the sun, to give bay salt.

Chlorine is produced on a large scale by any of a number of different methods:

1. By electrolysis of a concentrated solution of sodium chloride in water. Hydrogen is evolved at the cathode and chlorine at the anode. At the same time, an alkali metal hydroxide is produced in the electrolyte, and hence this process is often referred to as chlorine-alkali-electrolysis.

The chemical reactions occurring at each electrode and the overall cell process are given in the equations below:

cathode process: $2H_2O + 2e^- \rightarrow 2OH^- + H_2$
(cathode of iron)
anode process: $2Cl^- \rightarrow Cl_2 + 2e^-$
(anode of graphite)
cell process: $2H_2O + 2Cl^- \rightarrow 2OH^- + H_2 + Cl_2$

in which the symbol e^- represents a single electron. In the reaction vessel free chlorine and hydroxide ions must not come in contact with each other because chlorine would be consumed according to the reaction:

$$Cl_2 + 2OH^- \rightarrow (ClO)^- + Cl^- + H_2O.$$

To accomplish the separation of chlorine gas and the hydroxide ion, a porous wall is inserted between the electrodes (diaphragm process), or the iron cathode is replaced by a cathode consisting of liquid mercury (mercury cathode process) which avoids the production of hydroxide ions at the electrode. Instead, free sodium is discharged at the cathode, and this metal is readily dissolved in the mercury forming an amalgam, as follows:

$$2Na^+ + 2e^- \rightleftharpoons 2Na(amalgam).$$

The amalgam is allowed to react with water outside the cell:

$$2Na(amalgam) + 2H_2O \rightarrow 2Na^+ + 2OH^- + H_2.$$

The overall process is equivalent to the cell process given above.

2. By electrolysis of fused sodium chloride, which also produces metallic sodium; chlorine is again evolved at the anode.

3. By electrolysis of fused magnesium chloride, in which chlorine is formed as a by-product in the manufacture of metallic magnesium.

4. By oxidation of hydrogen chloride. In this process gaseous hydrogen chloride mixed with air or oxygen is passed over pumice in contact with cupric chloride as a catalyst, as shown in the following equation:

$$4HCl + O_2 \xrightarrow{\text{catalyst}} 2H_2O + 2Cl_2.$$

The equilibrium constant for this reaction decreases with increase of temperature; *i.e.*, the reaction proceeds less extensively at higher temperatures. In practice, however, a temperature of 400° C (750° F) is required to achieve a reasonable rate of conversion.

5. Of historical interest is the process in which a mixture of almost any solid chloride and manganese dioxide (MnO_2) yields chlorine when heated with concentrated sulfuric acid (H_2SO_4). The reaction occurs, as follows:

$$2NaCl + 3H_2SO_4 + MnO_2 \rightarrow MnSO_4 + 2NaHSO_4 + 2H_2O + Cl_2.$$

In the laboratory chlorine is frequently prepared by the oxidation of concentrated hydrochloric acid with permanganate or dichromate salts:

$$2MnO_4^- + 10Cl^- + 16H^+ \rightarrow 2Mn^{2+} + 5Cl_2 + 8H_2O$$

$$Cr_2O_7^{2-} + 6Cl^- + 14H^+ \rightarrow 2Cr^{3+} + 3Cl_2 + 7H_2O.$$

Laboratory preparations of chlorine

Similarly, chlorine may be prepared on a small scale by addition of any dilute acid to bleaching powder, a mixture of dry calcium salts, including the hydroxide, chloride, and hypochlorite.

Chlorine is marketed in steel cylinders in the liquid state under pressure. The total world production of chlorine was estimated at 17,000,000 tons in 1968, and more than 7,000,000 tons are produced annually in the United States. American production has shown an annual rate of increase of about 8 percent per annum since the early 1950s.

Most of the chlorine produced is used for chemical processes involving the introduction of chlorine into organic compounds, yielding carbon tetrachloride (used as a solvent, a fire extinguisher, and a dry-cleaning agent) and glycols (used as antifreeze), and other organic compounds for the manufacture of plastics (polyvinyl chloride) and synthetic rubber. Sulfur chloride, made by the action of chlorine on carbon disulfide or by combining sulfur and chlorine, is used in the vulcanization of rubber and as a chlorinating agent in organic synthesis. Chlorine and carbon monoxide form carbonyl chloride, or phosgene, which has been employed as a war gas and is used in metallurgy to transform certain oxides into chlorides. Much chlorine is used to sterilize water and wastes; and the substance is employed either directly or indirectly as a bleaching agent for paper or textiles. Chlorine is applied in the manufacture of hydrochloric acid, the extraction of titanium with formation of titanium(IV) chloride, the removal of tin from old tinplate, and in the production of antiknock materials for gasolines, such as tetraethyl lead. Anhydrous aluminum chloride is made by the reaction of chlorine with scrap aluminum or with aluminum oxide and carbon. Chlorine is also used to prepare silicon(IV) chloride and methyl chloride, which are employed in the synthesis of silicon materials. Chlorine enters directly, or indirectly as an intermediate, into many organic syntheses of industrial importance.

Physical and chemical properties. Chlorine is a greenish-yellow gas at room temperature and atmospheric pressure. It is considerably heavier than air. It has a choking smell, and inhalation causes suffocation, constriction of the chest, tightness in the throat, and—after severe exposure—edema (filling with fluid) of the lungs. As little as 2.5 milligrams per litre in the atmosphere causes death within a few minutes, but less than 0.0001 percent by volume may be tolerated. Chlorine was the first gas used in chemical warfare in World War I. The gas is easily liquefied by cooling or by pressures of a few atmospheres at ordinary temperature.

Toxicity of chlorine

Chlorine has a high electronegativity and a high electron affinity, the latter being even slightly higher than that of fluorine. The affinity of chlorine for hydrogen is so great that the reaction proceeds with explosive violence in light, as in the following equation:

$$H_2 + Cl_2 \xrightarrow{h\nu} 2HCl.$$

In the presence of charcoal, the combination of chlorine and hydrogen takes place rapidly (but without explosion) in the dark. A jet of hydrogen will burn in chlorine with a silvery flame. Its high affinity for hydrogen allows chlorine to react with many compounds containing hydrogen. Chlorine reacts with hydrocarbons, for example, substituting chlorine atoms for the hydrogen atoms successively. If the hydrocarbon is saturated, however, chlorine atoms readily add to the double or triple bond.

Chlorine reacts with many elements of both metals and nonmetals to give chlorides. Only toward carbon, nitrogen, and oxygen is it fairly inert. The products of reaction with chlorine usually are chlorides with high oxidation numbers, such as iron(III) chloride ($FeCl_3$), tin(IV) chloride ($SnCl_4$), or antimony(V) chloride ($SbCl_5$), but it should be noted that the chloride of highest oxidation number of a particular element is frequently in a lower oxidation state than the fluoride of highest oxidation number. Thus, vanadium forms a pentafluoride, whereas the pentachloride is unknown, and sulfur gives a hexafluoride but no hexachloride. With sulfur, even the tetrachloride is unstable.

Chlorine displaces the less electronegative halogens from compounds. The displacement of bromides, for example, occurs according to the following equation:

$$2Br^- + Cl_2 \rightleftharpoons Br_2 + 2Cl^-.$$

Furthermore, it converts several oxides into chlorides. An example is the conversion of iron(III) oxide to the corresponding chloride:

$$2Fe_2O_3 + 6Cl_2 \rightleftharpoons 4FeCl_3 + 3O_2.$$

With carbon monoxide chlorine gives carbonyl chloride; and with sulfur dioxide, sulfuryl chloride:

$$CO + Cl_2 \rightleftharpoons COCl_2$$

$$SO_2 + Cl_2 \rightleftharpoons SO_2Cl_2.$$

Chlorine water

Chlorine is moderately soluble in water, yielding chlorine water, and from this solution a solid hydrate of ideal composition $Cl_2 \cdot 7.66H_2O$ is obtained. This hydrate is characterized by a structure that is more open than that of ice; the unit cell contains 46 molecules of water and six cavities suitable for the chlorine molecules. When the hydrate stands, disproportionation takes place—that is, one chlorine atom in the molecule is oxidized and the other is reduced. At the same time, the solution becomes acidic, as shown in the following equation:

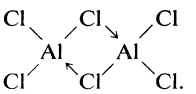
$$\overset{0}{Cl_2} + H_2O \rightleftharpoons \overset{+1}{HClO} + \overset{-1}{HCl}$$

in which the oxidation numbers are written above the atomic symbols. Chlorine water loses its efficiency as an oxidizing agent on standing, because hypochlorous acid gradually decomposes. The reaction of chlorine with alkaline solutions yields salts of oxo acids (see below).

The ionization potential of chlorine is high. Although ions in positive oxidation states are not very stable, high oxidation numbers are stabilized by coordination, mainly with oxygen and fluorine. In such compounds bonding is predominantly covalent, and chlorine is capable of exhibiting the oxidation numbers +1, +3, +4, +5, +6, and +7.

Principal compounds. The nonvolatile chlorides of the alkali and alkaline-earth metals are essentially ionic. The volatile chlorides of the nonmetals, such as boron, carbon, silicon, and sulfur, however, are covalently bonded molecular liquids at room temperature. Carbon tetrachloride is not miscible with water, but silicon(IV) chloride is readily decomposed with water (hydrolyzed).

Between the ionic and covalent extremes are the chlorides of such metals as aluminum, iron, titanium, or gold. In the chlorides of aluminum, iron(III), titanium(III), vanadium(III), chromium(III), and gold(III), a coordination number of 4 is found, and chloride bridges are present, as in the following formulation:



These chlorides are hydrolyzed by water.

The chlorides of silver and lead(II) are scarcely soluble in water. The solubility is increased in the presence of

excess chloride ions because of the formation of complex ions, namely dichloroargentate(I), $(AgCl_2)^-$, and tetrachloroplumbate(II), $(PbCl_4)^{2-}$. The increased solubility of mercury(I) chloride in the presence of chloride ions is due to disproportionation to metallic mercury and trichloromercurate(II) $[(HgCl_3)^-]$.

Preparation of anhydrous chlorides

The following methods are utilized for the preparation of anhydrous chlorides: (1) direct union of the element with chlorine, as with titanium (Ti) in:

$$Ti + 2Cl_2 \rightarrow TiCl_4;$$

(2) reaction of the element or its oxide with hydrogen chloride, as with zinc (Zn):

$$Zn + 2HCl \rightleftharpoons ZnCl_2 + H_2,$$

or magnesium oxide (MgO),

$$MgO + 2HCl \rightleftharpoons MgCl_2 + H_2O;$$

(3) chlorination with certain chlorides as in the reaction of chromic anhydride (Cr_2O_3) with carbon tetrachloride (CCl_4):

$$Cr_2O_3 + 3CCl_4 \rightleftharpoons 2CrCl_3 + 3COCl_2;$$

(4) dehydration of the hydrated chloride with hydrogen chloride or thionyl chloride as with the hydrated cobalt chloride $[Co(OH_2)_6]Cl_2$:

$$[Co(OH_2)_6]Cl_2 + 6SOCl_2 \rightleftharpoons CoCl_2 + 6SO_2 + 12HCl.$$

The chloride ion has an electronic configuration consisting of a complex octet, and hence it can act as a Lewis base (donor of an electron pair). The ion is strongly hydrated by water, which acts as a Lewis acid, and in this way the ion is stabilized. Because of its Lewis-base properties, the chloride ion is known to act as a ligand in complex compounds. The coordination numbers in such chloro complexes are usually 4 or 6; *e.g.*, tetrachloroaluminate(III) $[(AlCl_4)^-]$, hexachlorophosphate(V) $[(PCl_6)^-]$, and hexachlorotitanate(IV) $[(TiCl_6)^{2-}]$. It should be noted that several transition metal ions, such as cobalt(II), form only tetrahedral chloro complexes.

Lower chlorides of transition metal elements that have a very high heat of atomization are known only as metal clusters. Thus the dichlorides of molybdenum and tungsten, represented by M, contain a group formulated as $(M_6Cl_8)^{4+}$, in which the chlorine atoms are placed at the apices of an octahedron with the chlorine atoms lying above the centres of the eight triangular faces, thereby forming three-way bridges between the metal atoms. This structure is similar to those of certain cluster compounds of niobium and tantalum, with formulas $(Nb_6Cl_{12})^{2+}$ and $(Ta_6Cl_{12})^{2+}$, respectively, and it appears that formation of such groups of metal atoms with halogen bridges between them may be a recurring feature for these metals in their low oxidation states.

Hydrogen chloride. Hydrogen chloride is formed chiefly by three reactions: (1) reaction of the elements: $H_2 + Cl_2 \rightarrow 2HCl$; (2) hydrolysis of a covalent chloride such as thionyl chloride:

$$SOCl_2 + 2H_2O \rightarrow H_2SO_3 + 2HCl;$$

(3) reaction of an ionic chloride, such as sodium chloride, with sulfuric acid:

$$NaCl + H_2SO_4 \rightleftharpoons NaHSO_4 + HCl.$$

Hydrogen chloride is a colourless gas with a strong irritating odour. It dissolves rapidly and extensively in water, giving a strongly acidic solution, which is known as hydrochloric acid:

$$H_2O + HCl \rightleftharpoons (H_3O)^+ + Cl^-.$$

The concentrated hydrochloric acid contains about 40 percent hydrogen chloride. Many metals are dissolved in the acid giving the respective metal chloride and hydrogen, as may be seen in the equation for zinc (Zn):

$$Zn + 2HCl \rightleftharpoons ZnCl_2 + H_2.$$

This type of reaction is used for the preparation of hydrogen on a laboratory scale.

Anhydrous hydrogen chloride is less reactive. Only a few reactive metals, such as sodium (Na), react with the gas at room temperature:

$$2Na + 2HCl \rightarrow 2NaCl + H_2.$$

At higher temperatures reaction takes place with less reactive metals; in the case of a metal exhibiting several oxidation states, the lower chloride is formed preferably, as with tin (Sn) and iron (Fe):

$$Sn + 2HCl \rightarrow H_2 + SnCl_2$$

$$Fe + 2HCl \rightarrow H_2 + FeCl_2.$$

Liquid hydrogen chloride

Liquid hydrogen chloride may be used as a nonaqueous solvent. It dissolves tetraalkylammonium chloride to give a conducting (ionic) solution, in which the anionic species is an ion formulated as $(HCl_2)^-$. Hydrogen bonding in this ion is considerably weaker than in the related ion $(HF_2)^-$. Covalent chlorides, such as boron(III) chloride, behave as acids in liquid hydrogen chloride in that they are capable of accepting a chloride ion; in doing so they may be considered to be neutralized by chlorine ion donors, such as tetraalkylammonium hydrogen dichloride, which in effect are bases, the reaction being:

$$(R_4N)(HCl_2) + BCl_3 \rightleftharpoons (R_4N)(BCl_4) + HCl$$

base acid salt solvent.

Hydrochloric acid is used for the surface treatment of metals (the removal of oxide layers), for the manufacture of various chlorides, and for the preparation of dye substances. Hydrogen chloride is used for the preparation of various solvents, as well as vinyl chloride and other intermediates used in the preparation of plastics and synthetic rubber.

Oxo acids. Chlorine forms oxo acids with oxidation states of +1, +3, +5, and +7 and with coordination numbers 1, 2, 3, and 4, respectively. The names and formulas of these acids and the corresponding salts are given in Table 6. The acids are formed by reaction of chlorine with water and hydroxide ions under appropriate conditions.

Table 6: Oxo Acids of Chlorine

name	oxidation number	formula	corresponding oxide	name of the salt	anion
Hypochlorous acid	+1	$HClO$	Cl_2O	hypochlorite or chlorate(I)	$(ClO)^-$
Chlorous acid	+3	$HClO_2$	—	chlorite or chlorate(III)	$(ClO_2)^-$
Chloric acid	+5	$HClO_3$	—	chlorate or chlorate(V)	$(ClO_3)^-$
Perchloric acid	+7	$HClO_4$	Cl_2O_7	perchlorate or chlorate(VII)	$(ClO_4)^-$

Hypochlorous acid (HClO) exists only in aqueous solution. It is made by shaking precipitated mercuric oxide (HgO) in water with chlorine:

$$2Cl_2 + 2HgO + H_2O \rightleftharpoons HgO \cdot HgCl_2 + 2HClO.$$

It is also formed slowly in aqueous solutions of chlorine:

$$Cl_2 + H_2O \rightarrow HClO + HCl.$$

Hypochlorous acid is in equilibrium with its anhydride, dichlorine monoxide (Cl_2O):

$$2HClO \rightleftharpoons Cl_2O + H_2O.$$

Under the influence of light, or on heating, the acid disproportionates as in the following equation:

$$\overset{+1}{3}HClO \rightleftharpoons \overset{+5}{HClO_3} + \overset{-1}{2HCl}.$$

Hypochlorous acid is a weak acid (only slightly dissociated into ions) in water (acid dissociation constant about 10^{-8}) and a strong oxidizing agent. It is used as a bleaching agent and for the production of chlorohydrins and glycols. Chlorohydrins are formed as in the following equation:

$$HClO + CH_2 = CH_2 \rightarrow CH_2ClCH_2OH.$$

The salts of hypochlorous acid, the hypochlorites or chlorates(I), may be obtained (1) by the action of chlorine on solutions of alkali carbonate at a temperature below +40° C (104° F):

$$Cl_2 + Na_2CO_3 \rightarrow NaOCl + NaCl + CO_2;$$

(2) by passing chlorine into alkaline solutions at room temperature:

$$Cl_2 + 2OH^- \rightarrow (ClO)^- + Cl^- + H_2O;$$

or (3) by electrolysis of concentrated alkali chloride solutions in slightly alkaline medium at a high current density.

Chlorous acid

Chlorous acid ($HClO_2$) is known only in aqueous solution, in which form it is a stronger acid than hypochlorous acid (dissociation constant about 10^{-2}). The salts, called chlorites or chlorates(III), are much more stable than the free acid. The sodium salt exists in the anhydrous state as well as in the form of a trihydrate. The hydrate is white in colour and releases chlorine dioxide (ClO_2) under the influence of light; it is decomposed into sodium chloride and oxygen on heating at 200° C (392° F), according to the following equations:

$$3NaClO_2 \rightarrow 2NaClO_3 + NaCl$$

$$NaClO_2 \rightarrow NaCl + O_2.$$

Chlorites are best made by the reaction of chlorine dioxide (ClO_2) with peroxides:

$$Na_2O_2 + 2ClO_2 \rightarrow 2NaClO_2 + O_2,$$

or by disproportionation of ClO_2 in alkaline solution:

$$\overset{+4}{2ClO_2} + 2NaOH \rightleftharpoons \overset{+5}{NaClO_3} + \overset{+7}{NaClO_2} + H_2O.$$

Chlorites of some of the heavy metals may explode in the presence of certain organic materials. In alkaline solution chlorites are fairly stable, whereas in acidic solution they are converted to the free acid, which decomposes rather rapidly:

$$4HClO_2 \rightarrow 2ClO_2 + (ClO_3)^- + Cl^- + 2H^+ + H_2O.$$

Chlorites are used as bleaching agents for cellulose, oils, starch, sugar, wool, and other materials.

Chloric acid ($HClO_3$) is soluble in water, in which it acts as a fairly strong acid. The chlorate ion, $(ClO_3)^-$, is pyramidal in structure. The acid is a strong oxidizing agent, which dissolves even copper metal.

Potassium chlorate ($KClO_3$) forms white crystals, melting at 356° C (673° F), which are moderately soluble in water. When the dry crystals are heated, oxygen is released, particularly in the presence of manganese dioxide, as shown below:

$$2KClO_3 \rightleftharpoons KCl + O_2 + KClO_4.$$

Sodium chlorate ($NaClO_3$), with a melting point of 255° C (491° F), absorbs water (*i.e.*, is hygroscopic) and is soluble in water. Chlorates may explode in the presence of sulfur, phosphorus, and certain organic materials. Chlorates are produced by the action of chlorine on solutions containing hydroxide ions, as follows:

$$3Cl_2 + 6OH^- \rightleftharpoons (ClO_3)^- + 5Cl^- + 3H_2O$$

$$2HOCl + (ClO)^- \rightleftharpoons (ClO_3)^- + 2Cl^- + 2H^+.$$

Electrolysis of a hot potassium chloride solution, preferably at platinum or graphite electrodes, gives crystalline potassium chlorate on cooling. Chlorates are used to destroy weeds and in the production of chlorine dioxide, as well as in the manufacture of explosives and matches. Small amounts are used in drug products, such as mouthwash.

Perchloric acid ($HClO_4$) is the most stable oxo acid of chlorine. It contains chlorine with oxidation number +7. In the perchlorate ion, the chlorine atom is surrounded tetrahedrally by four oxygen ligands. There is a certain amount of pi bonding involved in the chlorine–oxygen bonds. The perchlorate ion has little tendency to serve as a ligand in complexes, and thus perchlorates are widely used as electrolytes in studies of complex formation in solution. The free acid is obtained by vacuum distillation of aqueous solutions. It is a colourless liquid, solidifying at −112° C (−170° F). It is commercially available as a 72 percent aqueous solution. It is miscible with water in all proportions and acts as one of the strongest acids in water. The crystalline mono- and dihydrates are known to be fully ionized in the solid state, having the formulas $(H_3O)^+(ClO_4)^-$ and

$$(H_2O{-}H{-}OH_2)^+(ClO_4)^-.$$

Perchlorates

Perchlorates are formed by heating chlorates under controlled conditions:

$$4(ClO_3)^- \xrightarrow{480°} 3(ClO_4)^- + Cl^-$$

Electrolytic oxidation of a cooled chlorate solution at a high current density also yields perchlorates. Any chlorate remaining in the electrolyte may be decomposed by hydrochloric acid, which does not react with the perchlorate ion. Perchlorate salts are converted to the free acid by reaction with fluorosilicic acid or with hydrochloric acid:

$$2NaClO_4 + H_2SiF_6 \rightleftharpoons Na_2SiF_6 + 2HClO_4$$

$$NaClO_4 + HCl \rightleftharpoons NaCl + HClO_4.$$

Alternatively, the free acid may be obtained by anodic oxidation of hydrochloric acid, followed by vacuum distillation. Sodium perchlorate is readily soluble in water, whereas the potassium, rubidium, and cesium salts are only sparingly soluble. Magnesium perchlorate forms white, hexagonal crystals, melting at 146° C (295° F); it is an excellent drying agent. Applications of perchlorates are in fireworks and explosives, and they are also employed as oxidizing agents in rockets.

Oxides. Chlorine forms binary oxides in which it has oxidation states of +1, +4, +5, and +7, as shown in Table 7.

Table 7: Oxides of Chlorine

name	oxidation number	oxide
Dichlorine monoxide	+1	Cl_2O
Chlorine dioxide	+4	ClO_2
Dichlorine hexoxide	+6	Cl_2O_6
Dichlorine heptoxide	+7	Cl_2O_7

Dichlorine monoxide (Cl_2O) is the anhydride (compound formed with loss of water) of hypochlorous acid, as shown in the following equilibrium:

$$Cl_2O + H_2O \rightleftharpoons 2HClO.$$

It is formed from chlorine and mercuric oxide:

$$2Cl_2 + HgO \rightleftharpoons HgCl_2 + Cl_2O.$$

Chlorine dioxide (ClO_2) is an orange-yellow gas at room temperature. It is made from sodium chlorate, sulfur dioxide, and sulfuric acid, according to the following equation:

$$2NaClO_3 + SO_2 + H_2SO_4 \rightarrow 2ClO_2 + 2NaHSO_4.$$

A second method of preparation involves treating silver chlorate at 90° C (194° F) with dry chlorine and condensing the resulting dioxide by cooling:

$$2AgClO_3 + Cl_2 \rightarrow 2AgCl + 2ClO_2 + O_2.$$

The chlorine dioxide molecule contains an odd number of electrons and thus is unstable. It decomposes by the following equation:

$$ClO_2 \rightarrow \tfrac{1}{2}Cl_2 + 2O_2.$$

It may be stored only when diluted by an inert gas. It is used for the preparation of chlorites and also as a bleaching agent, *e.g.*, for flour or cellulose.

Dichlorine hexoxide (Cl_2O_6) is a red liquid at room temperature. It is readily decomposed. The chief method

of preparation is the action of ozone (O_3) on chlorine dioxide:

$$2ClO_2 + 2O_3 \rightarrow Cl_2O_6 + 2O_2.$$

The substance appears to be in equilibrium with its monomeric form:

$$Cl_2O_6 \rightleftharpoons 2ClO_3.$$

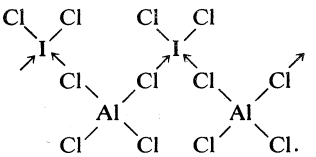
It is thermally unstable, decomposing to chlorine dioxide and oxygen.

Dichlorine heptoxide, (Cl_2O_7; boiling point −92° C [−134° F]) is the anhydride of perchloric acid. It is made by dehydrating perchloric acid with phosphorous pentoxide and distilling the product. The colourless liquid has strong oxidizing properties.

Halogen chlorides. The chlorine fluorides, in which the chlorine atoms exhibit positive oxidation numbers, have already been described. In its compounds with the other, less electronegative halogen elements, chlorine has the oxidation number −1. Of these, the iodine and bromine chlorides are chlorinating agents.

The iodine chlorides

Bromine(I) chloride is unstable, but iodine forms two stable chlorides, namely, iodine(I) chloride (ICl) and iodine(III) chloride (ICl_3). These compounds are obtained by direct union of the elements, and, when excess chlorine is used, the trichloride is obtained, in the form of yellow crystals. On addition of chloride ions the trichloride is connected to the planar ion, $(ICl_4)^-$. The trichloride also reacts with Lewis acids, such as aluminum(III) chloride ($AlCl_3$) or antimony(V) chloride ($SbCl_5$), to give crystalline compounds that contain chlorine bridges and are formulated as in the following example:



Iodine(I) chloride forms dark-brown crystals, which have an irritating smell and which melt on gentle warming. The compound may be prepared by heating iodine (III) chloride with the appropriate amount of iodine. Molten iodine(I) chloride behaves as an ionizing solvent for chlorides. By treatment with it, various elements are converted into their chlorides. The chloride ion reacts with iodine(I) chloride to give a polyhalide ion—namely, the $(ICl_2)^-$ ion—in which the atoms are arranged in a nearly linear manner. Other polyhalide anions containing chlorine have also been found in aqueous solutions.

Analysis. Free chlorine may be recognized by its smell, its colour, and its characteristic reaction with mercury to produce white mercury(II) chloride. Tests for chloride ions are:

1. The formation of a white precipitate of silver chloride on addition of silver nitrate in dilute nitric acid. (This precipitate is soluble in the presence of ammonia.)
2. The formation of chromyl chloride, a red gas, by heating a solid sample with potassium dichromate and concentrated sulfuric acid. When chromyl chloride is passed into water, a yellow chromate solution forms (bromides and iodides do not form analogous compounds).
3. The evolution of free chlorine by heating the sample with manganese dioxide and concentrated sulfuric acid.

The following methods are available for the quantitative determination of free chlorine:

Quantitative determination of chlorine

1. The chlorine-containing gas is shaken with an aqueous solution of potassium iodide, and the resulting iodine is determined by titration.
2. Chlorine is reduced in alkaline solution by an alkali arsenite. Back-titration of excess arsenite is carried out with potassium bromate.
3. In the presence of an alkali hydroxide, chlorine is reduced to the chloride ion by hydrogen peroxide, and the excess alkali hydroxide is back-titrated with acid.
4. With sulfur dioxide or sodium thiosulfate, chlorine is reduced to chloride, and the latter is analyzed as silver chloride (see below).
5. Colorimetric measurements are carried out in the presence of *o*-toluidine in hydrochloric acid.

For the determination of chloride ions, one of the following methods may be recommended: (1) gravimetric analysis (analysis by weight of a given product) as silver chloride; (2) titration of neutral chloride solution with silver nitrate in the presence of potassium chromate; and (3) potentiometric titration (measurement of voltage changes) with silver nitrate, a process that can be carried out in the presence of bromide and iodide ions.

Most insoluble chlorides can be melted with soda, and the resulting melt is then usually soluble in water. Organic compounds containing chlorine are heated with alkali peroxide, and the product is dissolved in water.

BROMINE

Occurrence and distribution. Because of the bad odour of the element, the French Academy of Science suggested the name bromine, from the Greek word *bromos*, meaning "bad smell," or "stench." Bromine occurs only in compounds and is considerably less abundant than chlorine. Apart from silver bromide (bromyrite), which is found in Mexico and Chile, the element is mainly found in seawater or in salt deposits. The bromide content of seawater is about 0.07 grams per litre, but the Dead Sea contains much more—up to five grams per litre. Natural salt deposits are the main source of bromine and its compounds in the United States. In Germany it is available as a by-product of the manufacture of potassium.

Production and use. Bromine is produced on a large scale from seawater by treatment with chlorine in the presence of sulfuric acid, according to the following equation:

Bromine from seawater

$$2Br^- + Cl_2 \rightleftharpoons Br_2 + 2Cl^-.$$

The product of the reaction is a dilute solution of bromine, from which the element is removed by blowing air through it. The free bromine is then mixed with sulfur dioxide, and the mixed gases are passed up a tower down which water is trickling. The following reaction takes place in the tower:

$$SO_2 + Br_2 + 2H_2O \rightarrow 2HBr + H_2SO_4,$$

resulting in a mixture of acids that is much richer in bromide ion than seawater. Treatment with chlorine again liberates bromine, which is freed from chlorine and purified by passage over moist iron filings. In Germany most bromine produced comes from the mother liquor obtained in the production of potassium salts. This mother liquor, which contains mainly magnesium bromide, is treated with chlorine in a continuous process to release the bromine.

Commercial bromine generally contains up to 0.3 percent chlorine. It is usually stored in glass bottles or, in the United States, in barrels coated with lead or Monel metal. In 1969, 392,000,000 pounds of bromine were produced, most of it in the United States.

Much of the bromine produced is converted to ethylene dibromide ($C_2H_4Br_2$), which is added to gasoline with tetraethyl lead to prevent deposition of lead in the engine. Bromine is also used for the preparation of silver bromide, which is employed in photography, and in the production of catalysts, such as aluminum bromide, as well as of organic dyestuffs and various other organic compounds.

Physical and chemical properties. Free bromine is a reddish-brown liquid with an appreciable vapour pressure at room temperature. Both liquid bromine and the vapour are highly toxic and produce painful burns on the skin. Like the other halogens, bromine exists as diatomic molecules in all aggregation states.

About 3.41 grams of bromine dissolve in 100 millilitres of water at room temperature. The solution is known as bromine water. Like chlorine water, it is a good oxidizing agent, and it is more useful because it does not decompose so readily. It liberates free iodine from iodide-containing solutions and sulfur from hydrogen sulfide. Sulfurous acid is oxidized by bromine water to sulfuric acid.

Bromine water

In the sunlight bromine water decomposes, with release of oxygen, as in the following equation:

$$H_2O + Br_2 \rightarrow 2HBr + \tfrac{1}{2}O_2.$$

From bromine water a hydrate can be isolated that contains 172 water molecules and 20 cavities capable of accommodating the bromine molecules. Bromine dissolves in aqueous alkali hydroxide solutions, giving bromides, hypobromites, or bromates, depending on the temperature. Bromine is readily extracted from water by organic solvents such as carbon tetrachloride, chloroform, or carbon disulfide, in which it is very soluble. In the organic solvents it gives an orange solution.

The electron affinity of bromine is high and similar to that of chlorine. It is, however, a less powerful oxidizing agent, chiefly because of the weaker hydration of the bromide ion as compared with the chloride ion. Similarly, a metal–bromine bond is weaker than the corresponding metal–chlorine bond, and this difference is reflected in the chemical reactivity of bromine, which lies between that of chlorine and that of iodine.

Bromine combines violently with the alkali metals and with phosphorus, arsenic, and antimony but less violently with certain other metals. Bromine displaces hydrogen from saturated hydrocarbons and adds to unsaturated hydrocarbons, though not as readily as chlorine does.

The ionization potential of bromine is high, and compounds containing bromine in positive oxidation numbers are stabilized by appropriate ligands, mainly oxygen and fluorine. Compounds with the oxidation numbers +1, +3, +4, +5, and +7 are known; they all contain covalent bonds.

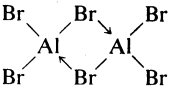
Principal compounds. The bromides of the alkali and alkaline-earth metals form ionic crystals in the solid state. The bromides of boron, carbon, silicon, germanium, and hydrogen, however, form molecular lattices in the solid state; such compounds are volatile.

The tendency to promote high oxidation numbers decreases in passing from fluorine to chlorine, and it does so even further in proceeding to bromine. Thus, for vanadium the highest fluoride is the pentafluoride (VF_5), the highest chloride is the tetrachloride (VCl_4), and the highest bromide is the tribromide (VBr_3). Similarly, sulfur gives a hexafluoride, (SF_6), a tetrachloride (SCl_4), and a monobromide (S_2Br_2) as the highest respective halides.

Properties of bromides

The bromide ion is considerably larger than the chloride ion, and this may account for the differences in structure of the solid phosphorus(V) chloride and bromide. Both compounds have an ionic lattice, which consists of tetrahedral tetrahalophosphorus $(PX_4)^+$, but the chloride contains octahedral hexachlorophosphate(V) anions, $(PCl_6)^-$, and the bromide, uncomplexed bromide ions (Br^-).

Aluminum bromide is similar in structure and properties to aluminum chloride in that it consists of dimeric, or double, molecules held together by bromide bridges, as in the following formulation:



Bromides of transition-metal elements are usually deeper in colour and have somewhat lower volatility than the corresponding chlorides. Their general chemical properties resemble those of the chlorides fairly closely, but they are more easily oxidized. The structures of some of the lower bromides are analogous to the chloride cluster compounds. Examples are the compounds formulated as $ReBr_3$, $MoBr_2$, WBr_2, Nb_6Br_{14}, and Ta_6Br_{14}.

Because the bromide ion is a weaker ligand toward hard metal ions than the chloride ion, complex bromides are in general less stable than the corresponding complex chlorides. Tetrachlorocobaltate, for example, is stable in dimethylsulfoxide solution, but tetrabromocobaltate is unstable in this solvent. The stability of the latter is increased, however, by decreasing coordinating properties of the solvent molecules, and thus it is easily formed in nitromethane or in acetonitrile.

Hydrogen bromide. Hydrogen bromide may be obtained by any of four principal methods:

1. By direct reaction of the elements, as in the following equation:

$$H_2 + Br_2 \rightleftharpoons 2HBr.$$

The reaction is much slower than that between chlorine and hydrogen, and it must be initiated by heat. It also proceeds by means of a chain mechanism involving atomic hydrogen and atomic bromine.

2. By hydrolysis of covalent bromides; *e.g.*, as with phosphorus(III) bromide (PBr_3):

$$PBr_3 + 3H_2O \rightleftharpoons H_3PO_3 + 3HBr.$$

When bromine is dropped onto red phosphorus in the presence of water, phosphorus(III) bromide is formed and—on contact with the water—immediately hydrolyzed.

3. By adding bromine to benzene containing aluminum powder, which acts as a catalyst for the following reaction:

$$C_6H_6 + Br_2 \xrightarrow{Al} C_6H_5Br + HBr.$$

4. By the action of phosphoric acid on a bromide, such as potassium bromide (KBr), as follows:

$$KBr + H_3PO_4 \rightarrow KH_2PO_4 + HBr.$$

If sulfuric acid is substituted for the phosphoric acid, the product is contaminated by free bromine, because the bromide ion is readily oxidized by sulfuric acid.

Hydrogen bromide is a colourless gas, similar in properties to hydrogen chloride. It is very soluble in water (193 grams of hydrogen bromide can be dissolved in 100 grams of water at 25° C), producing hydrobromic acid. This acid is even stronger than hydrochloric acid. Hydrobromic acid is easily oxidized when exposed to light, however, and turns brown because of the bromine liberated.

Hydrobromic acid

Hydrobromic acid may be prepared directly by hydrolysis of sulfur bromide, made by dissolving sulfur in excess liquid bromine:

$$S_2Br_2 + 5Br_2 + 8H_2O \rightleftharpoons 12HBr + 2H_2SO_4.$$

A constant boiling mixture containing about 47 percent hydrogen bromide is obtained by distillation.

Liquid hydrogen bromide has moderately good properties as a nonaqueous solvent. A solution of tetraalkylammonium bromide conducts electric current and contains anions formulated as $(HBr_2)^-$. Hydrogen bonding in this ion is similar to that in the corresponding chloro ion. Neutralization reactions occur with acidic bromides, such as boron(III) bromide. Hydrogen bromide is soluble in liquid hydrogen chloride, with which it appears to react according to the equation

$$(HCl_2)^- + HBr \rightleftharpoons HCl + (HClBr)^-.$$

Hydrogen bromide is used as a sedative and in the preparation of the various bromides. Ionic bromides may be obtained by the action of hydrobromic acid on appropriate metal hydroxide or carbonate. Silver bromide is scarcely soluble in water. The salt is used—together with silver iodide—in the production of photographic films.

Oxo acids and oxides. Hypobromous acid (HBrO) is known only in solution, which is obtained by the action of bromine water on mercury(II) oxide (HgO), as in the following equation:

$$2Br_2 + HgO + H_2O \rightleftharpoons HgBr_2 + 2HBrO.$$

Salts of the acid, the hypobromites, or bromates(I), are formed by the action of bromine on alkali hydroxide solution. They may be crystallized from the alkaline solution, as follows:

$$Br_2 + 2OH^- \rightleftharpoons (BrO)^- + Br^- + H_2O.$$

Hypobromites have oxidizing properties and are used as bleaching agents.

Bromic acid ($HBrO_3$) is prepared from barium bromate and dilute sulfuric acid, as follows:

$$Ba(BrO_3)_2 + H_2SO_4 \rightleftharpoons 2HBrO_3 + BaSO_4.$$

The salts of bromic acid, the bromates, or bromates(V), may be obtained by disproportionation of a hypobromite salt:

$$3NaBrO \rightleftharpoons NaBrO_3 + 2NaBr.$$

They also result when bromine is passed into a hot alkali hydroxide solution:

$$3Br_2 + 6NaOH \rightleftharpoons NaBrO_3 + 5NaBr + 3H_2O.$$

Bromic acid and bromates are much stronger oxidizing agents than the corresponding compounds of chlorine.

Perbromic acid

Perbromic acid ($HBrO_4$) may be obtained in various ways. The initial synthesis involved a radioactive-decay process in which the decay of selenium-83 incorporated in a selenate produced perbromate. Other modes of preparation involve oxidation of bromate solutions by xenon difluoride or by fluorine. Aqueous perbromic acid is a strong monobasic acid (releasing one hydrogen ion), which oxidizes inorganic substances, such as iodide ion and bromide ion, only slowly at room temperature. Neutralization of perbromic acid results in the formation of stable alkali metal perbromates. The perbromate ion is tetrahedral in the solid state, and it remains so in neutral and acidic solutions, thus resembling the perchlorate ion rather than the periodate ion.

All oxides of bromine are unstable. Dibromine monoxide (Br_2O) is prepared in analogy to the chloro compound. Tribromine octoxide (Br_3O_8) is obtained by the action of ozone on bromine at low pressure, and bromine dioxide (BrO_2) is prepared from the elements in a glow discharge at low temperature.

Halogen bromide. Bromine reacts with iodine to give iodine monobromide, as in the following equation:

$$\tfrac{1}{2}I_2 + \tfrac{1}{2}Br_2 \rightleftharpoons IBr.$$

The product forms dark-brown crystals at room temperature. Iodine bromide is less reactive than bromine but more reactive than iodine. It melts at 40° C (104° F), and in the liquid state it reacts with various elements to give the respective bromides. With bromide ions it forms a polyhalide ion, formulated as $(IBr_2)^-$. Like other trihalide ions, it is almost linear.

Analysis. A sensitive test for bromine is the reaction with fluorescein to give a deep red colour caused by bromination of the organic molecule, or by its reaction with fuchsine dyes in the presence of sulfurous acid, to give a deep blue colour. A more common test involves heating of the sample with dilute sulfuric acid in the presence of potassium dichromate; the bromine is then extracted with chloroform, and, upon addition of potassium iodide, the pink colour of iodine appears. The presence of bromine may also be recognized by the evolution of hydrogen bromide containing some brown bromine vapour when a solid sample is treated with concentrated sulfuric acid. Alternatively, chlorine may be added to an aqueous solution of a sample containing bromide, with development of a brown colour (free bromine).

For the quantitative determination of bromine the following methods are recommended:

1. Free bromine is titrated with sodium thiosulfate in the presence of potassium iodide:

$$Br_2 + 2(S_2O_3)^{2-} \rightleftharpoons (S_4O_6)^{2-} + 2Br^-.$$

2. Bromides may be determined either gravimetrically (by weight analysis) or by titration with silver nitrate:

$$Br^- + Ag^+ \rightleftharpoons AgBr.$$

3. In the presence of chloride and iodide, the potentiometric method may be used (as with chlorine).

4. In the absence of iodide, bromide may be oxidized to bromine, which is then determined in the distillate. Alternatively, bromide may be oxidized to bromate by hypochlorous acid. The excess of the oxidizing agent is destroyed by sodium formate, and iodine is liberated by addition of potassium iodide and acid, the free iodine being titrated by thiosulfate.

5. For the determination of bromine in an organic compound, the latter is oxidized by nitric acid, and the bromine is determined as silver bromide.

IODINE

Occurrence and distribution. Because of its violet-coloured vapours, the element was given the name iodine from the Greek word *ioeides*, "violet coloured." Iodine occurs to a small extent in seawater and is formed in seaweeds, oysters, and cod livers. Sodium iodate ($NaIO_3$) is contained in crude Chile saltpetre (sodium nitrate, $NaNO_3$). The human body contains iodine in the compound thyroxine, which is produced in the thyroid gland. It occurs in small quantities in much animal and vegetable matter.

Production and use. Iodine is produced commercially either from Chile saltpetre or from iodine-containing brines. In the former process, the salt is dissolved in hot water and the saltpetre allowed to crystallize on cooling. The mother liquor is used for further extractions until the extracts contain up to nine grams of iodine per litre. Sodium hydrogen sulfite is then added in order to reduce all iodate to iodide, and the solution is nearly neutralized with sodium carbonate. Fresh mother liquor is then added until all iodide is oxidized by the iodate to free iodine, according to the equation:

Iodine from brines

$$2NaIO_3 + 5NaHSO_3 \rightleftharpoons 3NaHSO_4 + 2Na_2SO_4 + H_2O + I_2.$$

The solid, containing up to 80 percent iodine, is collected, washed with water, and pressed into cheeselike blocks. These are heated to distill off both iodine and water.

Natural brines, or brines extracted from oil wells containing up to 150 milligrams of iodine per litre, are found in Java, California, and northern Italy. Impurities, such as clay, sand, and oil, are removed by filtration, and the solution is passed through a stream of sulfur dioxide and then through a number of containers containing bundles of copper wire. The copper(I) iodide that forms is removed by filtration, washed with water, dried, and finely ground. The product is heated with potassium carbonate to give potassium iodide, which is then oxidized to the free element with dichromate and sulfuric acid:

$$Cu_2I_2 + K_2CO_3 \rightleftharpoons 2KI + Cu_2O + CO_2$$

$$6KI + Na_2Cr_2O_7 + 7H_2SO_4 \rightleftharpoons 3I_2 + Cr_2(SO_4)_3 + Na_2SO_4 + 3K_2SO_4 + 7H_2O.$$

In an alternate process, chlorine is used as the oxidizing agent:

$$2KI + Cl_2 \rightleftharpoons 2KCl + I_2.$$

Iodine from seaweed

For a long time, iodine has been recovered on a commercial scale from seaweed. This is dried and burnt; the ash is leached with water; sodium sulfate and sodium chloride are removed by crystallization; and the remaining solution is concentrated by evaporation of water. The final solution, which contains 30 to 100 grams of iodine per litre, is treated with sulfuric acid in order to decompose any sulfite, and sulfide and manganese dioxide are added to release iodine, which is vapourized and purified by sublimation. Alternatively, addition of cupric sulfate gives cuprous iodide.

Iodine is widely used as a disinfectant and antiseptic, frequently in a solution of alcohol and water containing potassium iodide. Several compounds of iodine, such as iodoform (CHI_3), also serve as antiseptics.

Because iodine is converted to thyroxine in the thyroid gland, a small amount of iodine is essential for the body. In many places drinking water contains sufficient iodine for this purpose. In the absence of iodine in the water supply, however, goitre is prevalent, and a small quantity of iodine is frequently added to table salt in order to ensure against iodine deficiency.

Iodine and its compounds are used extensively in analytical chemistry, many analytical procedures being based on the release or uptake of iodine and its subse-

quent titration with sodium thiosulfate (iodometry). Unsaturation of fats (that is, the number of double bonds between carbon atoms) is determined by addition of free iodine (iodine number). Iodine compounds are also employed as catalysts in certain classes of organic reactions. Iodine, silver iodide, and potassium iodide are used in photography. Silver iodide is also used to seed clouds to induce rain. Iodine has been introduced into metallurgical processes for the production of certain transition metals in a high state of purity, among them titanium, zirconium, thorium, chromium, and cobalt. Electronic equipment, such as scintillation counters or neutron detectors, contain single-crystal prisms consisting of alkali metal iodides.

The annual production rate of iodine has increased from 96 tons in 1950 to 431 tons in 1955. The world's largest producer of crude iodine is Japan.

Physical and chemical properties. Iodine is solid at room temperature, dark coloured, and with a glittering crystalline appearance. The molecular lattice contains discrete diatomic molecules, which are also present in the molten and the gaseous states. Above 700° C (1,300° F) dissociation into iodine atoms becomes appreciable.

Solid iodine

Iodine has a moderate vapour pressure at room temperature and in an open vessel slowly sublimes (sublimation is vaporization of a solid—comparable to distillation of a liquid). For this reason, iodine is best weighed in a stoppered bottle; for the preparation of an aqueous solution the bottle may contain a solution of potassium iodide, which considerably decreases the vapour pressure of iodine, a complex (triiodide) being readily formed.

Molten iodine may be used as a nonaqueous solvent for iodides. The electrical conductivity of molten iodine has in part been ascribed to the following self-ionization equilibrium:

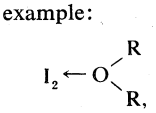
$$2I_2 \rightleftharpoons I^+ + I_3^-.$$

The alkali iodides are soluble in molten iodine and give conducting solutions typical of weak electrolytes. Alkali iodides react with compounds containing iodine with the oxidation number +1, such as iodine bromide, as in the following equation:

$$KI_3 + IBr \rightleftharpoons KBr + 2I_2.$$

In such reactions the alkali iodides may be regarded as bases.

The iodine molecule can act as a Lewis acid in that it combines with various Lewis bases. The interaction is weak, however, and few solid complex compounds have been isolated. The complexes are easily detected in solution and are referred to as charge-transfer complexes. Iodine, for example, is slightly soluble in water to give a yellowish-brown solution. Brown solutions are also formed with alcohol, ether, ketones, and other compounds acting as Lewis bases through an oxygen atom, as in the following example:



in which the R groups represent various organic groups.

Iodine gives a red solution in benzene, which is regarded as the result of a different type of charge-transfer complex. In inert solvents, such as carbon tetrachloride or carbon disulfide, violet-coloured solutions are obtained that contain uncoordinated iodine molecules. Iodine reacts also with iodide ions, because the latter can act as Lewis bases, and for this reason the solubility of iodine in water is greatly enhanced in the presence of an iodide. When cesium iodide is added, crystalline cesium triiodide may be isolated from the reddish-brown, aqueous solution. Iodine forms a blue complex with starch, and this colour test is used to detect small amounts of iodine.

The electron affinity of the iodine atom is not much different from those of the other halogen atoms. Iodine is a weaker oxidizing agent than bromine, however, because the iodide ion is a weaker Lewis base than the bromide ion and is hydrated to a smaller extent. In fact, the oxidizing properties of iodine are considerably weaker than those of bromine, chlorine, or fluorine. The following reaction—oxidation of arsenite, $(AsO_3)^{3-}$—in aqueous solution proceeds only in the presence of sodium hydrogen carbonate, which acts as a buffer:

$$(AsO_3)^{3-} + I_2 + 2OH^- \rightleftharpoons (AsO_4)^{3-} + 2I^- + H_2O.$$

In acidic solution, arsenate is reduced to arsenite, whereas, in strongly alkaline solution, iodine is unstable, and the reverse reaction occurs.

The most familiar oxidation by iodine is that of the thiosulfate ion, which is oxidized quantitatively to tetrathionate, as shown:

$$2(S_2O_3)^{2-} + I_2 \rightleftharpoons (S_4O_6)^{2-} + 2I^-.$$

This reaction is used to determine iodine volumetrically. The consumption of iodine at the end point is detected by the disappearance of the blue colour produced by iodine in the presence of a fresh starch solution.

Volumetric determination of iodine

Iodine combines directly with many elements. Silver and aluminum are easily converted into the respective iodides. White phosphorus unites readily with iodine, but no compounds of sulfur and iodine are known. The ionization potential of the iodine atom is considerably smaller than that of the lighter halogen atoms, and this is in accord with the existence of various compounds containing iodine in a positive oxidation number.

Principal compounds. As with the other halogens, iodine forms both ionic and covalent compounds in which it has the −1 oxidation state. Because of the low electronegativity of iodine as compared with those of the lighter halogens, bonding in iodides is considered to be less ionic than in other halides. Higher iodides are frequently unstable; for example, the highest iodides are the triiodide for phosphorus, the monoiodide for copper, and the triiodide for vanadium. No binary iodides are known for oxygen or sulfur. Iodides of transition metals are usually intensely coloured and less volatile than the corresponding bromides. It has been mentioned that the iodide ion is a weak Lewis base and shows a tendency to form polyiodide ions. In contrast to the fluoride ion, the iodide ion is considered a "soft" Lewis base in that it gives stable complex compounds with soft metal ions, such as palladium(II) ion, Pd^{2+}, and mercury(II) ion, Hg^{2+}. The tetraiodomercurate(II) ion, $(HgI_4)^{2-}$, is contained in a reagent used to detect small amounts of ammonia. The iodide ion can act also as a reducing agent; for example, it reduces chlorine to chloride, dichromate to chromium(III), permanganate to manganese(II), and aqueous hydrogen peroxide to water. In each case, iodine is liberated. Since iodine can be titrated with standard thiosulfate, these reactions can be used quantitatively.

Hydrogen iodide. Hydrogen iodide is obtained when iodine vapours and hydrogen are passed over a catalyst of platinum asbestos at 500° C (932° F). An alternative procedure is to allow water to react with a mixture of iodine and red phosphorus, as follows:

$$P + \tfrac{3}{2}I_2 \rightleftharpoons PI_3$$

$$PI_3 + 3H_2O \rightleftharpoons H_3PO_3 + 3HI.$$

An aqueous solution of hydrogen iodide is obtained when hydrogen sulfide is passed into a suspension of iodine in water and sulfur is filtered off:

$$H_2S + I_2 \rightleftharpoons 2HI + S.$$

Hydrogen iodide cannot be produced by the action of sulfuric acid on an iodide, because iodides are readily oxidized by sulfuric acid (see below). Hydrogen iodide is a colourless gas at room temperature and readily dissolves in water to give a strongly acidic solution. The solution, known as hydriodic acid, is one of the strongest acids in water. The salts formed from it are called iodides. Potassium iodide, a white crystalline solid melting at 686° C (1,267° F), is produced commercially to a greater extent than any other iodide. It is made by the reaction of ferrous iodide (prepared from iron and iodine) with potassium carbonate, as follows:

Hydriodic acid

$$2FeI_2 + 2K_2CO_3 + \tfrac{1}{2}O_2 \rightleftharpoons 4KI + Fe_2O_3 + 2CO_2.$$

It can also be prepared by the reaction of cuprous iodide with potassium carbonate:

$$2CuI + K_2CO_3 \rightleftharpoons 2KI + CO_2 + Cu_2O.$$

Hydrogen iodide acts also as a reducing agent; for example, it reduces sulfuric acid to a mixture of sulfur and hydrogen sulfide,

$$H_2SO_4 + 8HI \rightleftharpoons H_2S + 4I_2 + 4H_2O.$$

It also reduces arsenates to arsenites. Further examples are the reduction of dinitrogen monoxide and nitric acid:

$$N_2O + 10HI \rightleftharpoons 2NH_4I + H_2O + 4I_2$$

$$HNO_3 + 2HI \rightleftharpoons HNO_2 + I_2 + H_2O.$$

Iodides of the alkali and alkaline-earth metals are soluble in water, but iodides of silver, lead(II), mercury, and copper(I) are scarcely soluble. Mercury(II) iodide is red in colour; lead and silver iodides are yellow. In contrast to the chloride and the bromide, silver iodide is insoluble in aqueous ammonia, but all of the silver halides dissolve in the presence of thiosulfate with complex formation:

$$AgI + 2Na_2S_2O_3 \rightleftharpoons Na_3[Ag(S_2O_3)_2] + NaI.$$

Silver iodide is used in a fine crystalline form as the photosensitive constituent of photographic emulsions. Sodium thiosulfate is used to remove unreacted silver iodide. Covalent iodides, such as phosphorus(III) iodide or arsenic(III) iodide, are hydrolyzed by water with liberation of hydrogen iodide. Many organic iodides are useful in organic synthesis. Iodine is introduced into organic molecules by direct addition to double bonds or by replacement of hydrogen atoms. The latter reaction occurs less readily than with chlorine or bromine, and an organic iodide is in general more reactive than the corresponding chloride or bromide. When used as a reducing agent in organic chemistry, hydrogen iodide is more effective in the presence of phosphorus, because the liberated iodine forms the more reactive phosphorus iodide. Alkyl iodides may be used for alkylation reactions.

Oxo acids and oxides. Iodine is readily absorbed in alkaline solutions to give a hypoiodite:

$$I_2 + 2KOH \rightleftharpoons KIO + KI + H_2O.$$

The hypoiodite then undergoes hydrolysis to give hypoiodous acid:

$$KIO + H_2O \rightleftharpoons HIO + KOH.$$

The free acid is unstable and is amphoteric (acts as either an acid or a base), as shown in the following equation:

$$HIO + H_2O \rightleftharpoons (H_3O)^+ + (IO)^-$$

$$HIO + H_2O \rightleftharpoons (H_2OI)^+ + OH^-.$$

The hypoiodite ion is also unstable, because it readily disproportionates to iodate and iodide:

$$3(IO)^- \rightleftharpoons 2I^- + (IO_3)^-.$$

Iodic acid is precipitated when iodine is oxidized with concentrated nitric acid:

$$3I_2 + 10HNO_3 \rightleftharpoons 6HIO_3 + 10NO + 2H_2O.$$

It may be recrystallized from water. Because it is a strong acid with oxidizing properties, it first turns litmus red and then bleaches it. The solid acid reacts readily with certain nonmetals, such as phosphorus or sulfur, and with numerous organic compounds. The free acid is dehydrated above 110° C (230° F) to give iodine(V) oxide:

$$2HIO_6 \rightleftharpoons I_2O_5 + H_2O.$$

Iodates

The salts of iodic acid, the iodates(V), may be obtained by neutralization of the free acid or by dissolving iodine in hot alkali solutions:

$$3I_2 + 6KOH \rightleftharpoons 5KI + KIO_3 + 3H_2O.$$

This reaction is reversed in acid solution, with iodide and iodate(V) forming iodine. Iodate(V) is also produced by anodic oxidation of an iodide or by oxidizing the latter by potassium chlorate. Iodates are, in general, more stable than bromates and chlorates, and they are usually less soluble than the latter. Potassium iodate melts at 560° C (1,040° F), and by further heating it is decomposed to potassium iodide and oxygen.

When iodine is oxidized with potassium chlorate, the main product is potassium acid iodate, $KH(IO_3)_2$:

$$11KClO_3 + 6I_2 + 3H_2O \rightleftharpoons 6KH(IO_3)_2 + 3Cl_2 + 5KCl.$$

There is no known analogous compound of chlorine or bromine.

Whereas perchloric acid, $HClO_4$, contains a tetracoordinated halogen atom, orthoperiodic acid, H_5IO_6, has an iodine atom octahedrally coordinated by six oxygen atoms. It is a weak, pentabasic acid in aqueous solution.

The acid may be prepared by addition of sulfuric acid to a solution of barium periodate, the latter being prepared by heating barium iodate:

$$5Ba(IO_3)_2 \rightleftharpoons Ba_5(IO_6)_2 + 4I_2 + 9O_2$$

$$Ba_5(IO_6)_2 + 5H_2SO_4 \rightleftharpoons 5BaSO_4 + 2H_5IO_6.$$

A solution of the acid is also obtained by electrolysis of iodic acid at low temperature at a lead anode.

Periodic acid may be crystallized from aqueous solution. It is a white, deliquescent (water absorbing) solid, which is dehydrated to the composition HIO_4 by heating above 100° C (212° F) in a vacuum. Attempts to further dehydrate the material to a composition of I_2O_7 have not been successful, chiefly because oxygen is lost on further heating. Periodic acid is a stronger oxidizing agent in solution than perchloric acid, as witnessed by the fact that it oxidizes alcohols to aldehydes, ketones to carboxylic acids, and manganese(II) to permanganate. On the other hand, iodic acid is weaker in oxidizing properties than chloric acid. Most periodates are sparingly soluble in water.

Iodine oxides

Oxides with the following formulation have been prepared: iodine dioxide, I_2O_4; tetraiodine nonoxide, I_4O_9; and diiodine pentoxide, I_2O_5. The last is readily obtained by dehydration of iodic acid at 170° C:

$$2HIO_3 \rightleftharpoons I_2O_5 + H_2O.$$

The white solid is decomposed at 300° C without melting. When mixed with concentrated sulfuric acid and silica, the pentoxide is an oxidizing agent for carbon monoxide at room temperature:

$$5CO + I_2O_5 \rightleftharpoons 5CO_2 + I_2.$$

The oxide readily combines with water to give iodic acid. In spite of its name (based on a misconception), commercial I_2O_5 has almost entirely the formulation HI_3O_8. The oxide of formula I_2O_4 is prepared by treating iodic acid with concentrated sulfuric acid; this oxide has a polymeric structure composed of IO and IO_3 groups covalently bonded by oxygen bridges.

Other compounds. Because the ionization potential of iodine is lower than that of the other halogens, it shows a higher tendency to attain positive oxidation states. Iodine is found with oxidation number +1 in iodine halides, such as iodine monochloride and monobromide, as well as in hypoidous acid. A small percentage of ions in this oxidation state are believed to be present in molten iodine. The iodine cation is stabilized by coordination with strong Lewis bases, such as pyridine. Iodine gives a charge-transfer complex with pyridine, and this is partly ionized in appropriate nonaqueous solvents, as shown in the formulations below:

$$py + I_2 \rightleftharpoons py \longrightarrow I\overset{\curvearrowright}{-}I$$

$$2py \cdot I_2 \rightleftharpoons (py_2I)^+ + I_3^-$$

in which the arrow indicates movement of electrons, and py stands for a pyridine molecule. The perchlorate $(py_2I)^+(ClO_4)^-$ is also known.

Oxidation of iodine with perchloric acid gives iodine(III) perchlorate $[I(ClO_4)_3 \cdot 2H_2O]$, and oxidation of iodine by nitric acid in the presence of acetic anhydride gives io-

dine(III) acetate [$I(CH_3COO)_3$]. During electrolysis of the latter in glacial acetic acid, iodine is deposited at the cathode, as would be expected for positive iodine. All these compounds are hydrolyzed by water with disproportionation of iodine(III):

$$5I(CH_3COO)_3 + 9H_2O \rightleftharpoons I_2 + 3HIO_3 + 15CH_3COOH.$$

Analysis. Free iodine is detected (1) by the violet colour of the vapour or of its solution in carbon tetrachloride or carbon disulfide or (2) by the bright-blue colour produced in the presence of fresh starch solution in water, a very sensitive test.

Detection of iodide ions

Iodide ions may be detected in water (1) by the yellow precipitate of silver iodide, insoluble in water and ammonia solution, which is produced by addition of silver nitrate in the presence of dilute nitric acid, (2) by the formation of iodine on addition of chlorine or bromine water, (3) by the formation of iodine in the presence of other oxidizing agents, such as hydrogen peroxide or potassium dichromate, or (4) by the scarlet precipitate of mercury(II) iodide formed on addition of mercury(II) chloride, as in the equation below:

$$2I^- + HgCl_2 \rightleftharpoons HgI_2 + 2Cl^-.$$

This precipitate is dissolved by excess of iodide ions because of formation of a complex ion:

$$HgI_2 + 2I^- \rightleftharpoons (HgI_4)^{2-}.$$

Iodate or periodate is reduced by sulfurous acid to iodide and may be detected as such.

For the quantitative determination of iodine, one of the following methods may be recommended: (1) gravimetrically, by precipitation as silver iodide; (2) volumetrically, by titrating iodine with a standardized solution of sodium thiosulfate (using starch as an indicator); or (3) potentiometric titration with silver nitrate, which is applicable in the presence of both chloride and bromide. The second method is applied in the determination of many oxidizing substances. Dichromate, for example, reacts with excess potassium iodide in the presence of sulfuric acid, as shown:

$$(Cr_2O_7)^{2-} + 6I^- + 14H^+ \rightleftharpoons 2Cr^{3+} + 3I_2 + 7H_2O.$$

The iodine liberated is treated with standard thiosulfate solution.

ASTATINE

Because the element astatine has no stable or long-lived isotopes, it was given its name from the Greek word *astatos*, meaning "unstable." Minute amounts of the very short-lived isotopes, astatine-215, -218, and -219, occur in nature in radioactive equilibrium with certain long-lived, naturally occurring radioelements. In the body, astatine is concentrated in the thyroid gland. A substantial portion, however, distributes throughout the body, where it acts as an internal radiation source.

Production and use. The only practical way of obtaining astatine is by synthesizing it through nuclear reactions. The first synthesis was effected in 1940 by the bombardment of bismuth with alpha particles to obtain astatine-211.

Preparation of astatine

Astatine is usually prepared according to the general equation:

$$^{209}Bi\,(\alpha, xn)^{(213-x)}At,$$

which indicates that bismuth-209 takes up one alpha particle and emits x neutrons to form an isotope of astatine, whose atomic weight depends on the number of neutrons lost. Metallic bismuth may be used as a target material. From this, astatine may readily be removed by distillation in air from a stainless-steel tube. The free element begins to distill at 271° C (520° F, the melting point of bismuth), but the operation is best carried out at 800° C (1,470° F) with subsequent redistillation. If an aqueous solution of astatine is desired, the element may be separated by washing with an appropriate aqueous solution. Alternatively, the halogen may be removed from the target by chemical methods, such as dissolving in nitric acid, the latter being removed by boiling.

Another procedure involves the use of a metallic thorium target, which—after bombardment—is dissolved in concentrated hydrochloric acid containing hydrogen fluoride and chlorine.

Analysis. With the exception of a few spectrographic and mass-spectrographic studies, all investigations of astatine chemistry have utilized tracer techniques at extreme dilution—concentrations around or below 10^{-10} molar, molarity being the number of moles (gram molecular weight) per litre of solution. At such concentration, the effects of impurities can be very serious, especially for a halogen such as astatine, which exists in several oxidation states and can form many organic compounds. Iodine has been used as a carrier in most experiments. Techniques applied include coprecipitation, solvent extraction, ion exchange, and other forms of chromatography (separation by adsorption differences), electrodeposition (deposition by an electric current), electromigration (movement in an electric field), and diffusion. A direct identification of some astatine compounds has been made by mass spectrography.

Except for nuclear properties, the only physical property of astatine to be measured directly is the spectrum of atomic astatine. Other physical properties have been predicted from theory and by extrapolation from the properties of other elements.

Chemical properties. The astatide ion, At^-, is quantitatively coprecipitated with insoluble iodides, such as silver iodide or thallium(I) iodide. The diffusion coefficient of the iodide ion is 1.42 times that of the astatide ion, which moves more slowly toward the anode than the former under given conditions. The ion is formed by reduction of the element, using, for example, zinc and acid. It is oxidized to the zero valence state by ferric ion, Fe^{3+}, iodine (I_2), and dilute nitric acid. Thus, the astatide ion is a stronger reducing agent than the iodide ion, and free iodine is a stronger oxidizing agent than astatine.

Free astatine is characterized by volatility from solution and by extractability into organic solvents. It undergoes disproportionation in alkaline media. Astatine is coprecipitated with cesium iodide and thus appears to form polyhalide anions. Astatine extracted into chloroform has been shown to coprecipitate homogeneously with iodine when a portion of the latter is crystallized. Astatine seems to be present as the iodide, which appears to be more polar (*i.e.*, showing separation of electrical charge) in character than iodine bromide.

Astatine is known to occur in positive oxidation numbers. The astatate ion, $(AtO_3)^-$, is coprecipitated with insoluble iodates, such as silver iodate ($AgIO_3$), and it is obtained by oxidation of lower oxidation states with hypochlorite, periodate, or persulfate. So far no evidence for perastatate has been found, but this may be because the ion, $(AtO_6)^{5-}$, may show little tendency to coprecipitate with potassium iodate (KIO_4).

Astatine in the +1 state is stabilized by complexation, and complexes formulated as dipyridine astatine(I) perchlorate $[At(py)_2]\,[ClO_4]$ and dipyridine astatine(I) nitrate $[At(py)_2]\,[NO_3]$ have been prepared. Compounds with the formulas $(C_6H_5)AtCl_2$, $(C_6H_5)_2AtCl$, and $(C_6H_5)AtO_2$ have also been obtained. A variety of methods may be used to synthesize astatobenzene, C_6H_5At.

BIBLIOGRAPHY. J.W. MELLOR, *A Comprehensive Treatise on Inorganic and Theoretical Chemistry*, vol. 2 (1922, suppl. 1956), a comprehensive text in English; V. GUTMANN, *Coordination Chemistry in Non-Aqueous Solutions* (1968), a modern text emphasizing the use of certain halides as solvents and the behaviour of various halides in certain solvents, and (ed.), *Halogen Chemistry*, 3 vol. (1968), a collection of articles covering various aspects of halogen chemistry; and *International Review of Science (Inorganic)*, vol. 3 (1971), reviews recent articles on halogen chemistry; A.G. SHARPE and R.N. HASZELDINE, *Fluorine and Its Compounds* (1951), a readable, short text; Z.E. JOLLES, *Bromine and Its Compounds* (1966), a useful account of bromine chemistry; F.A. COTTON and G. WILKINSON, *Advanced Inorganic Chemistry*, 2nd rev. ed. (1966), contains a well-written chapter on the halogens; R.B. HESLOP and P.L. ROBINSON, *Inorganic Chemistry*, 3rd rev. ed. (1967), contains a well-written chapter on the halogens.

(V.G.)

Hals, Frans

One of the three or four greatest portrait painters of the 17th century, the Dutch painter Frans Hals belonged to that large group of artists who worked to order, fulfilling through their commissions a social function much as the modern photographer records contemporary life in his portraits. Hals never ventured in his art beyond the functional boundaries fixed by his clients. Yet he took far more than a mere inventory of the little world of Haarlem, its pompous notables, and its common people, of which he was the most faithful and accurate historian. In fact, with his uniquely expressive and facile style, he gave new life to the art of painting; and his sensitive observation and portrayal of the character of his subjects has rarely been excelled.

By courtesy of the Rijksmuseum, Amsterdam



"The Merry Toper," oil on canvas by Frans Hals. In the Rijksmuseum, Amsterdam. 81 cm × 66.7 cm.

Early life and works. Frans Hals left no written evidence about his life or his works, and only a brief outline of his biography is known. Though Flemish by birth, he was brought by his family to Holland at an early age. His birth date is uncertain (somewhere between 1581 and 1585), but it is definitely known that he was born in Antwerp, the son of a clothworker from Malines (Mechelen) and of a local girl. The family had at any rate settled in Haarlem by 1591 at the latest; the local townhall records give this date for the christening of Frans' younger brother Dirck, who also became a painter. Except for a brief visit to Antwerp in 1616, Hals lived all his life in Haarlem.

What he did for the first 25 or 30 years of his life is not known. The earliest indication of his activity as an artist was that, in about 1610, he joined the Guild of St. Luke of Haarlem, a body empowered to register artists as masters. Shortly afterward he married his first wife, Annetje Harmensdr. Abeel. She bore him two children before her death in 1615. Two years later, Hals married Lysbeth Reyniers, who was to survive her husband by some nine years. In all, Hals had ten children, and five of his eight sons became painters. None, however, was of note.

Formative influences

Tradition has it that Frans Hals was the pupil of Karel van Mander, a minor painter and poet who helped found a successful painting academy at Haarlem. There is no evidence either to support this claim or to refute it. From the beginning, however, Hals's work conflicted with the typical mannerisms of his presumed master. His early work is actually closer in spirit to that of Jacob Jordaens, an outstanding Baroque genre painter from Antwerp and pupil of Rubens who was younger than Hals by several years. The good humour of Hals's popular scenes recalls the joyous gatherings painted by the contemporary Dutch followers of the earthy, sensuous Italian painter Caravaggio.

Frans Hals seems, from the evidence of extant works, to have begun his career with sober portraits and with group portraits of members of the local guilds and military societies. The best of these early works—which already shows complete competence in portraiture—is a monumental painting entitled "Banquet of Officers of the Civic Guard of St. George at Haarlem," painted with a loose brushstroke technique that is unlike anything else in Dutch art of the time. It already has a sense of life and of relationship between the figures that was then unknown in this type of subject matter. By about 1620, however, Hals had begun to introduce into his paintings the jovial spirit that characterized his early works and that portrays with accuracy and enthusiasm one important aspect traditionally ascribed to Dutch character. Many of his portraits are simply pictures of merrymakers. The portrait of Hans Wurst in "The Merry Company" shows the sitter in a tall, wide-brimmed hat, wearing a necklace made of pig's feet, herrings, and eggs. The portrait of Mr. Verdonck shows the subject joyfully brandishing the jawbone of a horse. Similar in spirit are the portrait of Peeckelhaering clenching his beer mug, "The Merry Toper," and two later portraits, a picture entitled "Malle Babbe," which portrays an old madwoman laughing, with an owl perched on her shoulder, and a joyful picture in the Louvre of a laughing, carelessly dressed Gypsy girl. In Hals's group portraits, too, the spontaneous *joie de vivre* that is evident in the individual portraits is felt to a degree that revolutionizes the hitherto austere genre. One such painting is his second "Banquet of Officers of the Civic Guard of St. George at Haarlem," in which the figures take up postures normally employed for the expression of mystical religious rapture to celebrate their well-nourished contentment. In this painting, Hals displays his unmistakable genius for *mise en scène;* the dramatic effects he achieves here set him apart from most other painters. His militiamen are linked in a harmonious composition that makes the viewer aware of the cohesion of their group as a whole. Each conducts a dialogue with his neighbour, and here and there one figure is made purposely to disrupt the scheme with a gesture or a glance in the viewer's direction. Nothing is happening except a meal shared by typical members of the Dutch middle class and their conversations. Yet there is a majesty to this scene that is equal to any depiction of an incident from the life of a king. This painting also hints at the sense of mysterious spirituality, which, fostered by the artist's intimate knowledge of his subjects, came with his maturity to thread its way into his absolute realism.

Joviality of Hals's early works

By the 1620s Hals had definitively evolved a technique that was close to Impressionism in its looseness. Like the contemporary Spanish painter Diego Velázquez, he used colour to structure forms; and this use of colour is what sets the two artists apart from their contemporaries. Unique to Hals, however, is his use of quick, loose strokes of bright colour that suggest rather than enclose form and are highly expressive of movement and of the subjects' vitality. Most painters of the 17th century approached their paintings slowly, with preparatory drawings, a certain amount of underpainting, and an elaborate finish. Although there is no certain evidence of his method, Hals seems to have started directly on the canvas and painted quickly, leaving his first spontaneous expression, which is almost an oil sketch, as the finished work. Hals continued to use this technique, which gave a striking immediacy to his perceptive portrayals of character, all his life, painting with increasing freedom as he grew older.

Hals's loose painting technique

It has often been suggested that Frans Hals's life resembled the lives of the *bon-vivants* he portrayed at the beginning of his career. It is true that from 1616 he began to incur claims from creditors, and he was in financial

difficulties most of his life. He belonged, however, to the Haarlem St. George militia company and was a member of the Haarlem De Wijngaertranken (Society of rhetoricians) in 1618–19 (he was an associate of the society from 1616 to 1625); both of these facts are quite inconsistent with the romantic picture of dissipation that traditionally has been associated with the painter. Moreover, the stern preachers and theologians, the high-ranking officials, the surgeons, the admirals, the writers, and the respectable shopkeepers whose portraits Hals painted in great numbers were not likely to have posed for a dissolute high liver.

Later life and works. At any rate, the joviality began to disappear from the paintings of Hals's middle age. In the portraits painted after he reached the age of 40, the subjects seem to eye the world knowingly, with a shade of sadness in their faces. The earliest portrait that strongly shows this quality is "Man with Arms Crossed." Others follow that contain the same theme: "The Laughing Cavalier"; "Portrait of Isaac Abrahamsz. Massa"; "Pieter van den Broecke"; "Nicolaes Hasselaer"; "Willem van Heythuyzen"; and "Daniel van Aken Playing the Violin." These portraits seem to reveal a sense of foreboding; still, their mood ranges somewhere above the midpoint in the "human comedy." The period from 1630 to 1650 was Hals's most productive. He was very popular among the staid citizens of Haarlem's middle class, and during this time he painted more than 100 single portraits and six group and family portraits.

Increasing sobriety in Hals's mature works

Frans Hals lived to be very old, and it is in the paintings of his old age that his genius for portraying human character is fully revealed. The last years of his life were difficult materially, and he was harassed by discouraging family problems. Although he continued to work steadily, he received markedly fewer commissions after 1650. He had, during his long career, achieved an impressive reputation; he had been honoured by many important commissions, had become in 1644 an officer of the Guild of St. Luke, and in 1649 had painted the philosopher René Descartes. Still, although some continued to value his subtle perceptions, the public had generally begun to favour a more elegant style made popular by the English portrait painter Anthony Van Dyck. What commissions he did receive were not enough to support him, and, like his two great compatriots Rembrandt and Vermeer, he saw his possessions sold at auction for debt (1654). It was not until 1662 that his right to public assistance was recognized, and he was accorded a yearly pension by the city. In spite of this adversity, however, the portraits of Hals's last 16 years are his masterpieces. At this point, a view of the world is revealed in his painting in which the human comedy takes a tragic turn, and something breaks in the order that had kept the reasonable man and the madman separated. His portraits, no longer tempered by laughter, seem to express a realization that simply being is enough, after a certain age, for life to impress its tragic seal on everyone.

Henceforth, Hals drew gradually closer to traditional subjects and stored away his drinking glasses and his tableware. At the same time he diminished the intensity, the vividness of his themes, a greater simplicity appeared in his compositions, and he took more and more liberty with his painting. His palette lost a good deal of its lustre. But through decades of work he had evolved a remarkably broad range of blacks and whites to choose from, and these colours were sufficient for what he wanted to show.

Tragic masterpieces of Hals's old age

From 1650 on, his subjects begin almost to look awestruck, and Hals ceases to bind his compositions into powerfully articulated human masses. Instead, he strings the solitude of each figure together on a flimsy thread, with the pattern broken only here and there by some ultimate spark of vitality. The light seems to act as a nervous system in his subjects that whips their drowsy flesh back to life, and the magic of the brushwork seems to startle their faces out of a swoonlike slumber. In the two celebrated portraits of the "Governors of the Old Men's Home at Haarlem," one a group of old men and the other of old women, his men seem overcome with drunkenness and his women entranced by the obsession of death. Here he presents us with the most extraordinary reunion of senile decay ever assembled in the history of the pictorial arts; he shows us the quavering flame of dying life. It is not known whether these portraits were comprehensible to his models. Apparently, none of the regents of the home objected to the paintings hanging in their Hall of Honour. Perhaps his subjects shared the old painter's humility in the face of destiny. Thus, the harmony in the colourful blare of the early works came to be succeeded by an art that seemed to give form to elusive nervous twitches, sudden motions, and to heartbeats accelerating, only to falter and start again. All his life Frans Hals had acted as a lucid observer of Haarlem. He painted it in the loud mirth of youth, and, reflecting in the image that he made of it his own life and declining health, he remained its faithful companion until his death.

Old age fostered self-denial and a strict discipline in Hals, along with a new freedom in his painting. It most certainly was a painful time for the great painter. But the years had also sharpened his vision. There is no sign of religion in the evolution of his art; and it may be assumed that to Frans Hals, painting was a secular concern. Nevertheless, the loving compassion that permeated his art becomes, in his last years, something spiritual.

Frans Hals died on September 1, 1666, and was buried in the choir of the church of St. Bavo, in Haarlem. Like many artists whose style is unique in their own time, he left few direct followers; the closest was Adriaen Brouwer, who used Hals's techniques well to portray tavern scenes and similar subjects. Hals was for a long time regarded as a competent but limited painter whose consistent neglect of any subjects other than portraits gave him no place in the history of significant art. It was not until the 19th century that interest in his work was revived. He influenced Édouard Manet with his free style and Vincent van Gogh with his subtle range of colours. In modern times he has been appreciated for the serious and excellent realist painter that he was.

MAJOR WORKS

"Banquet of Officers of the Civic Guard of St. George at Haarlem" (1616; Frans Halsmuseum, Haarlem); "The Merry Company" (*c.* 1616–17; Metropolitan Museum of Art, New York); "Portraits of Paulus van Beresteyn and His Wife" (1620; Louvre, Paris); "Nurse and Child" (*c.* 1620; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Man with Arms Crossed" (1622; Trustees of the Chatsworth Settlement, Chatsworth, England); "The Laughing Cavalier" (1624; Wallace Collection, London); "Portrait of Jacob Pietersz. Olycan" (1625; Mauritshuis, The Hague); "Portrait of Aletta Hanemans" (1625; Mauritshuis, The Hague); "St. Matthew and the Angel" (*c.* 1625; State Museum, Odessa, U.S.S.R.); "Portrait of Isaac Abrahamsz. Massa" (1626; Art Gallery of Ontario, Toronto); "Young Man Holding a Skull" ("Hamlet"; *c.* 1626–28; Sir Richard Proby Collection, Peterborough, England); "Banquet of Officers of the Civic Guard of St. George at Haarlem" (1627; Frans Halsmuseum, Haarlem); "Verdonck" (*c.* 1627; National Gallery of Scotland, Edinburgh); "Peeckelhaering" (*c.* 1628–30; Staatliche Kunstsammlungun, Kassel, Germany); "The Merry Toper" (*c.* 1628–30; Rijksmuseum, Amsterdam); "Gypsy Girl" (*c.* 1628–30; Louvre, Paris); "Portrait of a Man" (1630; Buckingham Palace); "Malle Babbe" (*c.* 1630–33; Staatliche Museen Preussischer Kulturbesitz); "Nicolaes Hasselaer" (*c.* 1630–33); Rijksmuseum, Amsterdam); "A Burgomaster" (*c.* 1631–33; Frick Collection, New York); "Portrait of an Elderly Lady" (1633; National Gallery of Art, Washington, D.C.); "Portrait of a Man in His Thirties; Bust" (1633; National Gallery, London); "Pieter van den Broecke" (1633; Iveagh Bequest, Kenwood House, London); "Portrait of a Woman" (1635; Frick Collection, New York); "The Company of Captain Reynier Reael and Lieutenant Cornelis Michielsz. Blaeuw" or "The Meagre Company" (1637; Rijksmuseum, Amsterdam); "Willem van Heythuyzen" (*c.* 1637–39; Musée Royal des Beaux-Arts, Brussels); "The Violinist" ("Daniel van Aken Playing the Violin"; *c.* 1640; Nationalmuseum, Stockholm); "Portrait of a Young Man" (an unidentified member of the Coymans family) (*c.* 1645; National Gallery of Art, Washington, D.C.); "Portrait of Isabella Coymans, Wife of Stephanus Geraerdts" (*c.* 1650–52; private collection, Paris); "Man in a Slouch Hat" (*c.* 1660–66; Staatliche Kunstsammlungun, Kassel, Germany); "Governors of the Old Men's Home at Haarlem" (1664; Frans Halsmuseum, Haar-

lem); "Lady-governors of the Old Men's Home at Haarlem" (1664; Frans Halsmuseum, Haarlem).

BIBLIOGRAPHY

Bibliography and criticism: W. BURGER, "Frans Hals," in *Gazette des Beaux-Arts*, 24:219–232, 431–448 (1868); WILHELM VON BODE, *Frans Hals und seine Schule* (1871); FREDERIK SCHMIDT-DEGENER, *Frans Hals in Haarlem* (1901) and *Frans Hals* (1924); E.W. MOES, *Frans Hals, sa vie et son oeuvre* (1909); ANDRE FONTAINAS, *Frans Hals* (1908), in French; HERMANN KNACKFUSS, *Frans Hals* (1913), in German; F. DULBERG, *Frans Hals: Ein Leben und ein Werk* (1930); W. MARTIN, *Frans Hals en zijn Tijd* (1935); EDUARD PLEITZSCH, *Frans Hals* (1940), in German; K. BAUCH, *Frans Hals: Ein Vortrag* (1943); G.D. GRATAMA, *Frans Hals* (1943), and T. LUNS, *Frans Hals* (1946), both in Dutch; ERICH HOHNE, *Frans Hals* (1957); in German; PIERRE DESCARGUES, *Hals: étude biographique et critique* (1968; Eng. trans., *Hals: Biographical and Critical Study*, 1968); SEYMOUR SLIVE, *Frans Hals*, 2 vol. (1971), with bibliography.

Catalogues raisonnés: CORNELIUS HOFSTEDE DE GROOT, *Beschreibendes und kritisches Verzeichnis der Werke der Hervorragendsten holändischen Maler des 17. Jahrhunderts*, vol. 3 (1910; Eng. trans., *A Catalogue Raisonné of the Works of the Most Eminent Dutch Painters of the Seventeenth Century*, (1910), complete work—447 paintings; WILHELM VON BODE and M.J. BINDER, *Frans Hals: Sein Leben und seine werke*, 2 vol. (1914); W.R. VALENTINER and KARL VOLL, *Frans Hals*, 2nd ed. (1923), complete work—290 paintings; N.S. TRIVAS, *The Paintings of Frans Hals* (1941), complete work —109 paintings; SEYMOUR SLIVE, *Frans Hals*, 2 vol. (1971), complete work—approximately 220 paintings.

(P.De.)

Hamamelidales

Hamamelidales is an order of woody flowering plants that includes three families, 30 genera, and about 80 to 150 species. Knowledge is still very incomplete concerning several genera; thus the actual number of species remains uncertain. The families and most of the genera are isolated both morphologically (*i.e.*, with respect to structure and form) and geographically, although the order as a whole is distributed worldwide. Thirteen genera contain only one species each and five genera only two species each. The order Hamamelidales is especially interesting as a group from the standpoint of the phylogeny of flowering plants because it apparently forms a link between groups with extreme types of floral organization; viz., the primitive woody plants with large, petalled flowers (the magnolias, buttercups, saxifrages, and roses) and the catkin-flowered plants (Amentiferae), which include the beeches, birches, alders, and oaks.

General features. *Size range and diversity of structure.* All representatives of the order are woody: trees and shrubs, partly evergreen, partly deciduous; there are no herbaceous species. Most members of the order are medium-sized trees, but some are small shrubs (*e.g.*, *Hamamelis*, witch hazel; *Fothergilla*, witch alder); others are giant trees up to 60 metres (200 feet) high (*e.g.*, *Liquidambar*, sweet gum; *Altingia; Symingtonia;* and *Platanus*, plane tree).

Distribution and abundance. The actual distribution of the order is worldwide (except in Arctic regions), but most of the species and genera have very limited individual ranges of distribution and it is possible that some of them will soon become extinct. Moreover, there are only a few species that grow in large pure stands. Several species are widely represented in northern fossil deposits.

Economic importance. Some genera are attractive for their early blossoming; some are even truly winter blooming as is especially true of *Hamamelis* (witch hazel) and *Corylopsis*. Others are known by their brilliant yellow, orange, or crimson leaf coloration in the fall (*e.g.*, *Disanthus; Fothergilla*, the witch alder; *Hamamelis; Liquidambar*, the sweet gum; and *Parrotia*, ironwood) or for being impressive, large trees (*Liquidambar* and *Platanus*). All these genera are more or less hardy in temperate regions. *Rhodoleia*, which has showy, red, bird-pollinated flowers, can be cultivated only in tropical or subtropical regions. Parts of *Hamamelis* (witch hazel) plants have long been used as raw material for various purposes. The North American Indians used the fluid from boiled stems to stop bleeding and heal wounds. Today alcoholic extracts from leaves and bark are used for styptic (antibleeding) remedies, and watery extracts are used for various cosmetics and household liniments. Medicinal uses

The wood of several trees is used as construction material or for furniture (*e.g.*, *Altingia, Liquidambar, Parrotia, Symingtonia*, and *Platanus*). In other genera the twigs or bast (fibres) are used as a tough binding and plaiting material (*e.g.*, *Molinadendron* and *Parrotiopsis*).

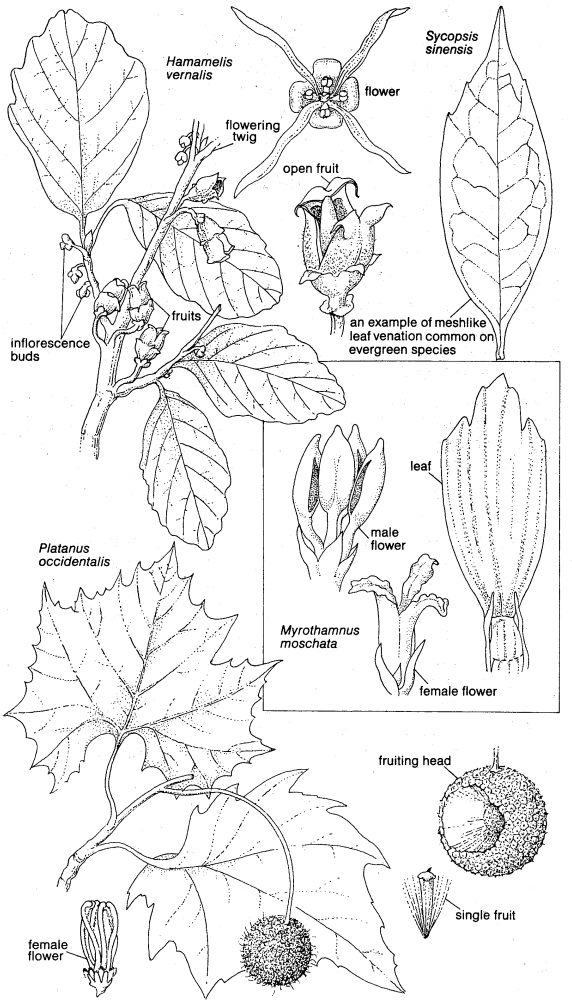
Natural history. *Life cycle.* The reproduction biology of the Hamamelidales shows some outstanding features. The range of pollination mechanisms is relatively wide. Pollination may be effected by bees or bumblebees (*e.g.*, in *Corylopsis*), by flies (in *Disanthus* and *Hamamelis*), by birds (in *Rhodoleia*), or by wind (in *Liquidambar, Parrotia, Sinowilsonia, Sycopsis*, and *Platanus*). In certain members of the family Hamamelidaceae fertilization of the ovules occurs long after pollination—*e.g.*, five to seven months in *Hamamelis virginiana*.

In the genera of Hamamelidaceae with one seed per carpel (structural unit of pistil), seed dispersal takes place by an ejecting mechanism. As the fruit opens, the bony endocarp (inner layer) bends in a characteristic manner and presses against the very hard and smooth surface of the seeds. The seeds are then forcibly discharged over distances of several metres. In contrast, the small, winged seeds of the many-seeded fruits of *Liquidambar* are distributed by wind. In *Platanus* the tufted nutlets are wind dispersed. In many species of the family Hamamelidaceae, seed germination is usually delayed because of the very hard and impervious seed coat; the seeds sometimes take one or two years to germinate.

Ecology. The family Platanaceae and a large part of the family Hamamelidaceae occur in moist woodlands, especially on stream banks, and on steep woody slopes. Other species grow best on drier rocky soils. The exact ecological range is unknown for most species. Many members of the family Hamamelidaceae grow in tropical and subtropical mountain regions, particularly in undisturbed forest areas. In general, representatives of the families Hamamelidaceae and Platanaceae are not specialized for extreme habitats, but the species of Myrothamnaceae are quite remarkable in being among the most extremely drought-resistant of all seed plants and ferns—twigs and leaves can dry out and remain in a dry condition for several years without dying. They are found in the arid regions of tropical and southern Africa and Madagascar. Adaptation to arid habitats

Form and function. *Vegetative characters.* All species of the order are woody plants. The leaves are simple and have small appendages at the base of the leaf stalks called stipules (except in *Rhodoleia*). They are usually arranged alternately along the stems. The anatomical structure of the wood is relatively uniform; the small, angular vessel elements (water-conducting cells) mostly have very oblique end plates with many crossbars, the scalariform condition.

Flower and fruit characters. The individual flowers are usually medium-sized or small, but they often form dense inflorescences (flower clusters). Several genera bear unisexual flowers, but only the family Myrothamnaceae seems to be dioecious; *i.e.*, its species bear male and female flowers on separate plants. The perianth—sepals and petals—especially the petal portion, is often reduced or absent, and the number of floral parts sometimes varies considerably within one species. The stamens (male flower structures) are usually provided with a short or long anther-connective protrusion—an appendage on the pollen sacs. The anthers open by means of valves except in the family Myrothamnaceae and in some genera of Hamamelidaceae. The female flower parts (the pistils) are composed of fused carpels (structural units of the pistil), except in the family Platanaceae, in which they are not fused, but they always show free styles, the elongate upper parts of the pistils. Each carpel usually contains a single ovule except in the family Myrothamnaceae and some Hamamelidaceae. The fruits are capsules (nutlets in Platanaceae) that split open along a ventral (inner-side) line. In the family Hamamelidaceae, other



Vegetative, floral, and fruiting structures of three families of the order Hamamelidales.
(Top) Family Hamamelidaceae, (centre right) Myrothamnaceae, (bottom) Platanaceae.
Drawing by M. Pahl based on (*M. moschata* flowers) *Beitrage zur Biologie der Pflanzen* (1966); Duncker and Humboldt, and (*P. occidentalis* flowers) G.H.M. Lawrence, *Taxonomy of Vascular Plants* (1951); The Macmillan Company

modes of opening are also common. The seeds of the families Hamamelidaceae and Platanaceae contain no starch, but protein and oil are present.

Evolution. *Fossil record.* The family Hamamelidaceae was widely distributed in the Northern Hemisphere, possibly in the Cretaceous Period (about 65,000,000 to 136,000,000 years ago); certainly in the Tertiary Period (2,500,000 to 65,000,000 years ago). In Europe the group is known at least from the Eocene Epoch (beginning about 54,000,000 years ago) until the lowest Pleistocene (about 2,500,000 years ago). *Liquidambar* is known from the Upper Cretaceous of North America. The Tertiary discoveries have been attributed with more or less certainty mainly to the genera *Corylopsis*, *Hamamelis*, *Liquidambar*, and *Parrotia*. Some other fossils have been described as separate genera; the oldest ones from the European Eocene are *Hamamelidanthium* and *Steinhauera*. The oldest remains of *Platanus* are known for America from the Upper Cretaceous; for Europe from the middle Oligocene Epoch (about 30,000,000 years ago). *Credneria*, known by leaves from the Cretaceous and long considered as ancestor of *Platanus*, has been transferred to the family Menispermaceae.

Phylogeny. The fossil record is too sparse to give sufficient information about the evolution of the group as a whole and about phylogenetic (evolutionary) relationships to other orders. Nevertheless, the characters of the living groups are interesting enough for certain phylogenetic conclusions. Floral structure, inflorescence morphology, leaf structure, wood anatomy, and chemical data show relationships to orders of the so-called "Amentiferae" (catkin bearers—catkins are elongate clusters of many small unisexual flowers), especially to the order Fagales (which includes the beech and birch families). Possible relationships are also indicated to the order Urticales (elm, mulberry and nettle families), and less probably to the orders Myricales (sweet gale family), Juglandales (walnut family), and Casuarinales, but not to the Salicales (willow order—one of the catkin-bearing plant groups). *Corylopsis* and *Sinowilsonia* (family Hamamelidaceae) are the most interesting genera with respect to the relationships shown to the catkin-bearing plant orders. *Sinowilsonia*, for example, has reduced unisexual flowers arranged in catkins. There are connections, on the other hand, to certain groups of the order Saxifragales, especially to the family Cunoniaceae. Certain superficial resemblances are also seen to the primitive woody families Cercidiphyllaceae, Eupteleaceae, Tetracentraceae, and Trochodendraceae, but they are considered to be the result of convergent evolution and not close phylogenetic affinity. Thus, the family Hamamelidaceae is most interesting in being a sort of link between two quite different main groups of flowering plants. The order Hamamelidales, therefore, represents an ancestral tropical and subtropical stock for several plant groups that have an amentiferous life form, which mainly developed in the temperate regions. Moreover, this connection lends support to the idea of a monophyletic origin (*i.e.*, one evolutionary line of descent) of the flowering plants.

Key evolutionary position

Classification. *Distinguishing taxonomic features.* Useful characters to distinguish the three families of the Hamamelidales include phyllotaxy (leaf arrangement), leaf form, hair types, structure of the inflorescences and flowers (particularly of the gynoecium—the female flower parts), and fruits.

Annotated classification.

ORDER HAMAMELIDALES

Woody plants. Leaves simple or palmately lobed, with stipules. Vessel elements in the wood narrow, with oblique end walls, perforation plates scalariform ("ladder-like") with numerous bars. Condensed inflorescences. Flowers mostly small, often unisexual, with reductions in floral structure. Calyx (sepals) lacking in few genera. Corolla (petals) often small or lacking. Wind pollination in some genera. Stamens with connective protrusion. Carpels (structural units of pistil) free or mostly fused in the region of the ovary but with free styles. Distribution centres: Middle and Eastern Asia, the Malaysian region, Central and North America, Australia, Madagascar, and tropical and southern Africa. Three families, 30 genera, and 80–150 species.

Family Hamamelidaceae (witch-hazel family)

Trees and shrubs. Leaves simple with pinnate (feather-like) or palmate (fanlike) venation or with three prominent veins or palmately lobed. Leaves either evergreen (and then with margins often smooth and lateral veins joining each other near the leaf margins) or deciduous (and then with lateral veins mostly straight and ending in marginal teeth). Stipules paired (lacking in *Rhodoleia*). Phyllotaxy (leaf arrangement) alternate, often in 2 ranks, rarely opposite. Stellate (branched in starlike pattern) hairs present in most genera (at least in the subfamily Hamamelidoideae). Inflorescences (flower clusters) mostly spicate (in spikes) or capitate; but paniculate (many-branched) in a few genera. Flowers bisexual, but partly functionally male in some genera, rarely totally unisexual; medium-sized to small, often with a hypanthium (floral cup); mostly perigynous or epigynous (*i.e.*, the ovary is half or wholly enclosed within the basal tissues of the sepals, petals, and stamens). Calyx mostly present. Petals ribbonlike in some genera, white or yellow to dark red, absent in many genera. Stamens mostly with a short or long apical protrusion. Anthers opening by 1 or 2 valves on each side or by lateral slits. Staminodia (sterile stamens) sometimes present. Carpels 2, fused at ovary level, free above, and forming 2 distinct styles. Stigma (pollen-receiving surface at tips of styles) sometimes large (in wind-pollinated genera) and warty. Ovules 1 to several per carpel, in a lateral position. Nectaries present in some genera, localized on different floral parts, depending on the genus. Calyx, corolla, and androecium (male flower parts) often present in 5s or 4s, but in some genera the number of parts is higher and variable. Fruits are capsules with leathery exocarp (outer layer) and mostly hard, bony endocarp. Seeds often black, shining, smooth, and very hard.

Distribution centres: Middle and Eastern Asia, the Malaysian region, North and Central America, Australia, tropical and southern Africa, and Madagascar. Twenty-eight genera, about 75–140 species.

Subfamily Hamamelidoideae

Usually 1 (rarely to 3) ovule per carpel.

Subfamily Disanthoideae

Six ovules per carpel, flowers hypogynous (ovary positioned above other flower parts), bisexual.

Subfamily Symingtonioideae

Six ovules per carpel, flowers epigynous, polygamous (unisexual and bisexual flowers present on the same plant).

Subfamily Rhodoleioideae

Many ovules per carpel, flowers bisexual, forming showy, red pseudanthia.

Subfamily Liquidambaroideae

Many ovules per carpel, flowers unisexual, inconspicuous.

Family Platanaceae (sycamore or plane-tree family)

Large monoecious (male and female flowers on the same plant) trees. Stem bark exfoliating. Leaves usually palmately lobed, deciduous, with one sheathlike stipule on the leaf stalk. Candelabra-shaped hairs. Vessel elements with scalariform or with simple perforations. Inflorescences unisexual, in globose head-like clusters. Flowers small, calyx minute, petallike structures occasionally present. Staminodes present in female flowers. Stamens with broad, flat connective protrusion. Anthers open by 2 valves on each side. Carpels free. Stigma with unicellular nipple-like projections. Ovules 1 per carpel. Number of floral parts varies from 2 to 9. Nectaries lacking. Flowers wind pollinated. Fruits small, tufted nutlets. Distribution: North and Central America, eastern Mediterranean region, and eastern Asia. One genus, 2 to 9 species.

Family Myrothamnaceae

Bushy dioecious shrubs, extremely drought resistant. Leaves opposite, folded, apically toothed, persistent, long living, with paired stipules. Hairs totally absent. Inflorescences spicate. Flowers unisexual. Calyx and corolla absent. Stamens 3 to 8 with apical protrusion. Anthers open by lateral slits. Carpels 3 or 4, basally fused. Styles free. Stigmas large and warty. Ovules many per carpel, with a lateral placentation. Fruits capsules. Distribution: arid zones of Madagascar, tropical and southern Africa. One genus, 2 species.

Critical appraisal. Some authorities include additional families in the order Hamamelidales, among them Tetracentraceae, Cercidiphyllaceae, Eupteleaceae, Bruniaceae, Stachyuraceae, Daphniphyllaceae, and Buxaceae. All these families, however, occupy quite isolated positions. Furthermore, some of the subfamilies of the Hamamelidaceae, as presented here, have been raised to the status of separate families by some authors. These include Disanthaceae, Rhodoleiaceae, and Altingiaceae (= Liquidambaroideae).

BIBLIOGRAPHY. A.L. BOGLE, "Floral Morphology and Vascular Anatomy of the Hamamelidaceae: The Apetalous Genera of Hamamelidoideae," *J. Arnold Arbor.*, 51:310–366 (1970), a description of the floral structure, especially of the vascular anatomy of eight genera; P.K. ENDRESS, "Systematische Studie über die verwandtschaftlichen Beziehungen zwischen den Hamamelidaceen und Betulaceen," *Bot. Jb.*, 87: 431–525 (1967), morphology, anatomy, and ontogeny of certain members of the family Hamamelidaceae, especially of the inflorescences and flowers of *Corylopsis* (discussion of the phylogenetic role of the Hamamelidaceae included); "Die Infloreszenzen der apetalen Hamamelidaceen, ihre grundsätzliche morphologische und systematische Bedeutung," *Bot. Jb.*, 90:1–54 (1970), a description of the inflorescence structure and taxonomic revision of ten genera; W.R. ERNST, "The Genera of Hamamelidaceae and Platanaceae in the Southeastern United States," *J. Arnold Arbor.*, 44:193–210 (1963), a taxonomic survey of *Hamamelis, Fothergilla,* and *Platanus;* E.H. FULLING, "American Witch Hazel: History, Nomenclature and Modern Utilization," *Econ. Bot.*, 7:359–381 (1953), an historical survey of the utilization of *Hamamelis* in America; H. HARMS, "Hamamelidaceae," in A. ENGLER and K. PRANTL, *Die natürlichen Pflanzenfamilien*, 2nd ed., 18a:303–345, 487 (1930), a detailed taxonomic treatment of the whole family Hamamelidaceae; IRMGARD JAGERZURN, "Infloreszenz- und blütenmorphologische, sowie embryologische Untersuchungen an *Myrothamnus* Welw.," *Beitr. Biol. Pfl.*, 42:241–271 (1966), an investigation of the family Myrothamnaceae; H.F. KAMMEYER, "Die schönen Zaubernüsse (Hamamelisgewächse)," *Neue Brehm Büch.* 194 (1957), a popular survey of the ornamental genera of Hamamelidaceae.

(P.K.E.)

Hamburg

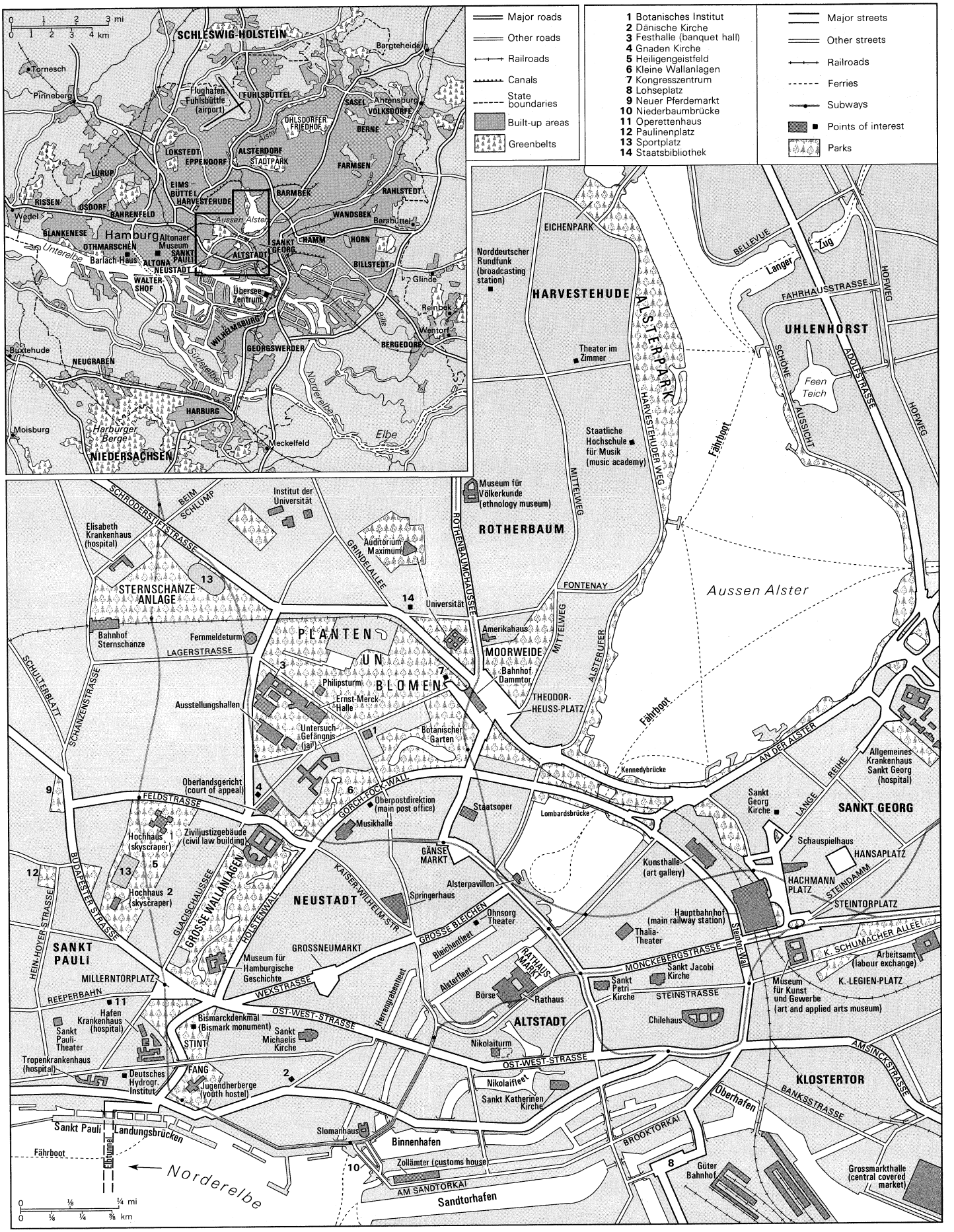
The Free and Hanseatic City (Freie und Hansestadt) of Hamburg is the second smallest of ten *Länder* ("states") of the Federal Republic of Germany, with a territory of only 291 square miles, or 753 square kilometres; but its 1,810,000 inhabitants (1971) make it the most populous municipality in West Germany. The official name, which covers both the *Land* and the actual town, reflects Hamburg's age-long tradition of particularism and self-government: Hamburg and Bremen, in fact, are the only German city-states to keep, in the 20th century, something of their medieval independence. The characteristic individuality of Hamburg has been deliberately maintained by its people so that, in many spheres of public and private life, the city's culture has not succumbed to the general trend of standardization.

Superficially, there is little in Hamburg to remind the visitor of this great city's millennial history. Only the five principal churches within the ancient city walls—dedicated to Sankt Jacobi, Sankt Petri, Sankt Katharine, Sankt Nikolai, and Sankt Michaelis—conserve traces of the past; and, even so, none of them is in its original condition. Fire has destroyed almost all the older residences and warehouses, and what was left untouched by conflagration has often been rebuilt for contemporary purposes by the citizens themselves, who are generally hard-headed businessmen for all their love of tradition. The layout of the old city centre, however, can still be detected in some of the ancient street names, and in the *Fleete* (canals such as the *Grachten* of Dutch cities), which connect the Alster River with the docks on the Elbe River. The best view of the city is to be enjoyed from the Lombardsbrücke (Lombard Bridge) across the Binnenalster (Inner Alster), whence the towers of the five churches can be seen rising high, but by the 1970s more and more glass-fronted skyscrapers were shooting up to change the horizon.

The largest port in Germany, Hamburg has prospered on trade, both inland and overseas. Merchants and shipowners, therefore, are at the top of its social scale and have set the standard of the city's life-style. With a cosmopolitan outlook, they are basically conservative and skeptical, mistrusting fashionable innovation however loudly publicized it may be. Concerned with practical realities, they are usually undogmatic and open-minded. Fairness, common sense, and understatement are among the ideas they have taken over from the English, long contact with whom has made Hamburg a remarkably Anglophile city.

History. Hamburg's history begins with the Hammaburg, a moated castle of modest size, built about AD 825 on a sandy promontory between the Alster and Elbe rivers, on the northeastern border of the Carolingian dynasty's empire. In 834, during the reign of the emperor Louis the Pious, the castle's baptistery became the seat of an archbishopric; and Archbishop Ansgar made the young city of Hamburg the base of his missions to the heathens of northern Europe. Vikings burned the city in 845, and the rebuilt Hamburg was burned down again eight times in the following 300 years. By the end of the 11th century, Hamburg's role as the spiritual metropolis of the north was fulfilled, and thenceforward commerce was to replace evangelism as the principal function of the city. Between 1120 and 1140 some trading businesses were installed there; and the foundation of Lübeck, on the Baltic, by Adolf II, count of Holstein, further promoted the economic development of Hamburg as Lübeck's port on the North Sea. In the autumn of 1188 a group of Hamburg entrepreneurs received from their feudal overlord, Adolph III of Schauenburg (Schaumburg), count of Holstein, a charter for the building of a new town, adjacent to the old one, with a harbour on the Alster River and with facilities for the use of the Elbe River as an outer roadstead; and, on May 7, 1189, the emperor Frederick I Barbarossa confirmed Count Adolph's dispositions in a charter granting special trading rights, toll exemptions, and navigation privileges to the nascent port.

The 13th century saw Hamburg growing steadily both



Central Hamburg and (inset) its metropolitan area.

Role in the Hanseatic League

in area and in economic importance thanks to the development of the Hanse (an association of merchants trading in a particular area) into a multilateral association of the north German merchant cities, the great Hanseatic League, in which Hamburg's role was second only to Lübeck's. A major entrepôt for the trade between Russia and Flanders, Hamburg proceeded to safeguard the trade routes by acquiring tracts of land along the branches of the Elbe in the immediate vicinity of the town and also on the estuary farther downstream (Ritzebüttel, nucleus of the later Cuxhaven, was acquired by Hamburg in 1393) and so came to control the use of the river and to be recognized as the protector, in the emperor's name, of navigation on its lower course. Some political complications arose with the death, in 1459, of the last Schauenburg count of Holstein, since his princely rights in Germany passed thereafter to the royal house of Denmark; but Hamburg scarcely recognized Danish suzerainty in anything but a formal and wholly ineffectual way.

Toward the end of the Middle Ages, the Hanseatic League gradually dissolved. Hamburg then went its own way and by 1550 had surpassed even Lübeck in economic importance. A stock exchange was founded in 1558 and the Bank of Hamburg in 1619; a convoy system for shipping was inaugurated in 1662, Hamburg's merchantmen being the first to be escorted on the high seas by German men-of-war; and about the same time marine insurance was first introduced into Germany. In 1678, furthermore, the first German opera house was established in Hamburg. There were two causes for this new ascendancy: firstly, the wars of religion in the Netherlands in the second half of the 16th century had prompted many Dutch merchants to emigrate to the Unterelbe (Lower Elbe) region, with the result that Hamburg was henceforth to be the focus of their already established international commerce; and, secondly, the city had been so efficiently fortified in the decade 1616–25 that it could pursue its business untroubled throughout the worst crises of the Thirty Years' War (conventionally 1618–48). By the end of the 17th century, Hamburg, with 70,000 inhabitants, was the largest city in Germany after Cologne.

The Treaty of Gottorp, concluded with the Danes on May 27, 1768, not only released Hamburg from theoretical subjection to the King of Denmark and so paved its way to being acknowledged, in 1770, as an "immediate" imperial city of Germany (that is, a city with no overlord other than the Emperor himself) but also ceded to Hamburg the islands, from Veddel to Finkenwerder, that lay between the city and the left banks of the Elbe River and that, a century later, were to be the site of new docks. Hamburg, however, was not long to enjoy its new advantage: the Napoleonic Wars overthrew the old order in Germany, and the little state was eventually annexed, in 1810, to Napoleon's French Empire.

After Napoleon's downfall (1814–15), Hamburg became a member state of the German Confederation, with the designation "Free and Hanseatic city of Hamburg" from 1819. Prosperity was quickly recovered, as Hamburg's trade was extended to newly opened territories in South and Central America, in West and East Africa, and in East Asia. Even the great fire of May 1842, which devastated one-fourth of the city centre, did not check the booming economy, and the harbour was converted into one accessible at any time, without regard to the state of the tides in the Elbe Estuary. Under the German Empire, founded in 1871, the political status of Hamburg was maintained, and development proceeded unchecked: the splendid Baroque houses of the densely populated Brook Islands were demolished in the 1880s to make room for the warehouses of the new free port; and, by the end of the 19th century, in the course of which the population grew from 130,000 to 700,000, Hamburg had expanded far beyond its previous limits, absorbing such former suburbs as Sankt Pauli and Sankt Georg and spreading its tentacles into the countryside, toward Eimsbüttel, Eppendorf, Harvestehude, and Barmbek.

Hamburg entered the 20th century determined to maintain and to strengthen its position as "Germany's gateway to the world": new docks and wharves were constructed on the left bank of the Elbe River. The outbreak of World War I in 1914 brought progress to a standstill, however. Hamburg's international trade collapsed, and its merchant fleet of 1,466 ships (3,171,246 gross registered tons) was virtually confined to port. After the war the victorious Allies demanded nearly all of Hamburg's ships by way of reparation from Germany.

The city during World Wars I and II

For many years after the war, Hamburg could undertake no further development because it had already exhausted all the potentialities of its territory. The Greater Hamburg Ordinance of January 26, 1937, changed this situation by allowing Hamburg to incorporate the neighbouring cities of Altona, Wandsbek, and Harburg, which until then had belonged to Prussia. The immediate prospect of expansion, with the development of these areas on a basis of large-scale planning, was shattered by the outbreak, in 1939, of World War II, during which repeated air raids demolished 55 percent of Hamburg's residential area and 60 percent of the harbour installations and killed 55,000 people. When the war ended in 1945, only the most strenuous efforts could supply the elementary needs for Hamburg's survival; but 25 years later reconstruction was almost complete. Two mayors gave conspicuous service to Hamburg in the postwar years: Max Brauer, in office from 1946 to 1953 and again from 1957 to 1960; and Herbert Weichmann, from 1965 to 1971.

The contemporary city. Hamburg stands at the northern extremity of the Lower Elbe Valley, which is between five and eight miles (eight and 12 kilometres) wide there. To the southeast of the old city, the Elbe divides itself into two branches, the Norderelbe and the Süderelbe; but these branches meet again opposite Altona, just west of the old city, to form the Unterelbe, which flows into the North Sea some 68 miles (110 kilometres) downstream from Hamburg. Tides on the Elbe produce an average difference of about 7.5 feet (2.3 metres) between high and low water in Hamburg's harbours. Two other rivers flow into the Elbe at Hamburg—the Alster from the north and the Bille from the east.

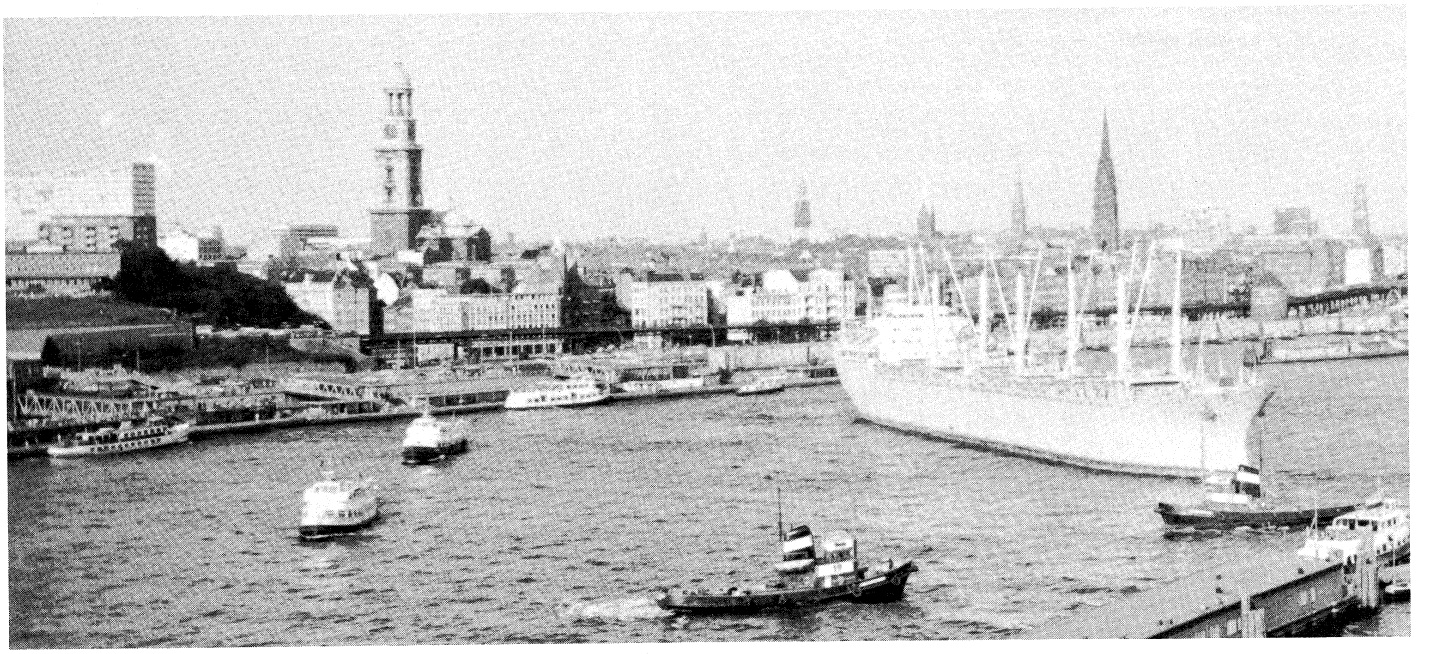
Hamburg's state territory covers 291 square miles (753 square kilometres) and is approximately 25 miles in diameter. Its border with Niedersachsen, to the south, is 49 miles long; and its border with Schleswig-Holstein, to the north, 79. Of the total area, 7.7 percent consists of running or standing water; 35 percent is agricultural land; 6.3 percent is forest; 41 percent is residential; 8.7 percent is devoted to industry; and 1.3 percent is occupied by railroads or by the airports.

Hamburg has mild winters, late springs, cool summers, high humidity, and frequent fog. The mean annual temperature is 47° F (8° C).

Population. The decline of Hamburg's population from 1,854,000 inhabitants in 1965 to 1,800,000 in 1971 led statisticians to forecast that by the mid-1980s there would be only about 1,735,000. Already by 1971 there were only 800,200 wage earners officially recorded, including 287,600 in manufacturing industry, 148,700 in commerce, and 38,600 in banking and insurance. These figures take no account of some 100,000 people employed in Hamburg but domiciled in Niedersachsen or in Schleswig-Holstein.

More than three-quarters of Hamburg's population is Protestant Christian, but there were about 140,000 Roman Catholics in 1971. Jews, of whom there had been 27,000 in 1933 (when Hitler took power), counted only 1,400 but had a new synagogue. There were two mosques for the 3,000 Muslims. Hamburg's non-German labour force amounted to 46,000 out of the 98,000 foreign residents.

Economy. There were some 2,000 Hamburg firms engaged in worldwide trade by the end of the 1960s. In 1970 they handled 32.6 percent of West Germany's total imports and 36.8 percent of her total exports. Chief among imports were vegetable oils and fats, tea, coffee, petroleum, tropical fruit, and uncured tobacco. Exports comprised machinery, electrotechnical products, processed petroleum fuel and lubricants, copper, and pharmaceutical products. Most of this foreign trade was with the



The harbour and city of Hamburg.
Hans Hartz

United States, Great Britain, The Netherlands, France, and Italy. Furthermore, Hamburg was dealing with more than 50 percent of the transit trade of West Germany as a whole. Of Hamburg's 110 credit institutions, over 40 are private banks.

Industrial importance to Germany

Having absorbed Altona, Harburg, and Wandsbek in 1937, Hamburg has become West Germany's major industrial city. All processing and manufacturing industries are represented there. Though few of the latter are in the hands of large-scale firms, Hamburg treats 80 percent of West Germany's copper supplies (the Norddeutsche Affinerie, on Veddel, is Europe's second largest copperworks), 40 percent of its vegetable oil, and 35 percent of its mineral oil, besides building 35 percent of its shipping and producing 25 percent of its cigarettes. The inauguration of a nuclear plant near Stade, in Niedersachsen, to provide power at a reasonable cost to the industries bordering the Unterelbe was obviously beneficial to Hamburg.

Government and institutions. According to the constitution of June 6, 1952, legislative authority is vested in the Bürgerschaft, or State Parliament, which comprises 120 members elected for a four-year term. The general election of March 22, 1970, returned only three political parties to this assembly: the Social Democrats, with 70 seats; the Christian Democrats, with 41; and the Free Democrats, or Liberals, with 9.

The State Parliament elects the government, the Senat, which is organized on a collegiate basis, the president, or "first mayor," (*erster Bürgermeister*) being elected by the Senat itself for one calendar year of office. The Senat as a whole represents the Free and Hanseatic City of Hamburg in its dealings with the other federal *Länder,* with the federal government, and with foreign states. Each senator is responsible for a particular department, but administrative problems of a local nature are delegated to the seven district offices and to the 15 local authorities.

The Rathaus (City Hall), where the Senat and the Bürgerschaft meet, in the centre of the city near the Binnenalster, was built late in the 19th century, in the Neo-Renaissance style.

Hamburg's coat of arms displays a three-towered castle, silver (argent), on a red (gules) field, its design being derived from the city's great seal of 1241. The state flag likewise shows a white castle on a red field. The state's anthem, "Stadt Hamburg an der Elbe Auen" ("City of Hamburg on the Shores of the Elbe"), was written by G.N. Bärmann in 1828 and set to music by Albert Methfessel.

Hamburg is second only to New York City in the international count of consulates: in 1972 there were 75 accredited foreign representatives in Hamburg.

Services. A focus of intercontinental as well as of western European and Scandinavian trade, Hamburg is the base of about 100 of Germany's 180 shipping concerns. Of its 200 miles of riverside, 40 have been built up as docks. The port dealt with 45,000,000 tons of goods in 1971, and, in an average month, 725 ships, representing 287 lines, sail from it for more than 1,000 destinations. About 19,200 oceangoing vessels docked at Hamburg in 1971, nearly 9,000 of them being liners that were calling regularly there. Hamburg's Übersee-Zentrum, with an area of 132,000 square yards (110,000 square metres), is the world's largest roofed warehouse: as many as 25,000 packages for overseas export may be handled there in a single day—sorted, stored, and finally dispatched to the appropriate vessels. There is also the Burchardkai container terminal, in the Waltershofer dock; enclosing a surface of 340,000 square metres, it is the largest accommodation of its kind in Europe. The harbour's labour force amounts to some 14,500 people, with a reputation for hard work and efficiency.

The port of Hamburg

The harbour proper has 400 miles of track linking the waterfront to the German railroad network. For motor transport, as many as 1,800 trucks enter the harbour zone in a day. To relieve the city centre of Hamburg from long-distance traffic, a 3,200-metre tunnel under the Elbe was undertaken, to be opened in 1973 as a part of the Stockholm—Lisbon highway.

The airport of Hamburg–Fuhlsbüttel, which dates from 1911, is one of the oldest in Europe. It has two runways from which even the largest jet-propelled aircraft could still take off in the early 1970s. Arrivals and departures of aircraft numbered 89,000 in 1970, with 3,130,000 passengers. As Fuhlsbüttel is reaching the limit of its capacity, another major airport is being built some 19 miles (30 kilometres) to the north of the city, between the little towns of Kaltenkirchen and Bad Bramstedt, to be one of the four German airports providing supersonic flights: it is hoped that it will be operative by the end of the 1970s.

Since World War II, Hamburg has succeeded Berlin as the chief city for much of the German press. It houses the head office of German Press Agency (DPA), which has branches in 70 foreign countries, and daily newspapers appearing in Hamburg, such as *Bild, Hamburger Abendblatt,* and *Die Welt,* have an aggregate circulation of 6,000,000. The leading weekly papers are *Bild am Sonntag, Welt am Sonntag,* and *Die Zeit.* Periodicals with an

aggregate circulation of 24,600,000, including *Stern* and *Der Spiegel*, also appear in Hamburg.

The Norddeutsche Rundfunk, Hamburg's broadcasting station, serves Niedersachsen and Schleswig-Holstein as well as its own city. Its television studios produce the daily news program of the Arbeitsgemeinschaft der Öffentlichrechtlichen Rundfunkanstalten der Bundesrepublik Deutschland (ARD, or Association of German Broadcasting Institutions).

The greatest economic centre of West Germany, Hamburg has since 1960 become the site of first-class trade fairs. The exhibition premises cover an area of 57,000 square yards on the fringe of the city. Especially worthy of mention are the German Boats Exhibition and Ship and Machine International.

Cultural life and recreation. The University of Hamburg, founded in 1919, ranks as the largest in West Germany, as its 15 departments have a total of 25,000 students. Suitably enough, in view of Hamburg's interests, the university has always given particular importance to teaching foreign languages, economics, geography, ethnology, and international law; and no less suitably, its 180 institutes and seminars include the Institute for Foreign Trade and Overseas Commerce, the German Hydrographic Institute, the Institute for Oceanography, and the Bernhard Nocht Institute for Seamen's and Tropical Diseases. Also attached to the university is the Deutsche Elektronen-Synchrotron, a research station ranking among the world's greatest particle accelerators.

Besides the university, Hamburg has a State Advanced School for Music and Interpretative Art, and a State Advanced School for the Sculptural Arts, both of which give a college-level education.

Among Hamburg's six museums, the Kunsthalle, founded in 1868 by Alfred Lichtwark, an outstanding patron of artists, is one of Europe's most remarkable galleries, in particular for its collection of 19th- and 20th-century work. The Museum für Kunst und Gewerbe (Museum of Art and Applied Arts), founded in 1877 by the jurist Justus Brinckmann, has one of the most significant collections of ancient artifacts in West Germany and especially famous examples of Japanese art and of Jugendstil (Art Nouveau). The Museum für Hamburgische Geschichte (Museum of Hamburg History), which has grown up from a collection of local antiquities started in 1839, exhibits architect's plans, artists' views, and models of the town and of the harbour, together with models illustrative of the history of navigation, in such a way as to present a most impressive conspectus of the state's past. The Museum of Ethnology and Prehistory, founded in 1878, has comprehensive collections in its own fields. The Altonaer Museum, opened in 1863, specializes in north German subjects, with special attention to Schleswig-Holstein, and houses Germany's largest collection of old ships' figureheads. The Helms-Museum, in the Harburg district, is a local museum for the part of Hamburg south of the Elbe but is soon to house all antiquities representing the prehistory and most ancient history of the whole territory. The Ernst-Barlach-Haus, in Jenisch Park, was founded in 1961–62 by another great patron of the arts, Hermann F. Reemtsma, so as to make his private collection accessible to the public. Hamburg's once famous Zoological Museum was destroyed by bombs in 1943 after a century of existence.

Music and drama in Hamburg

The Hamburg State Opera, which dates, as has been said, from 1678, won international celebrity in the 1960s and 1970s, thanks largely to its director, the Swiss composer Rolf Liebermann: its performances of classical and contemporary works bear comparison with those given by the great opera houses of Vienna, Milan, London, and New York. The Deutsche Schauspielhaus, a leading theatre, enjoyed a high reputation from 1955 to 1963, when Gustaf Gründgens directed and performed there. The Thalia-Theater, founded in 1843, with a multifaceted program that includes plenty of light entertainment, suits Hamburg's taste well enough. All these three establishments are generously subsidized by the state. There are also four other theatres, namely, the Hamburger Kammerspiele, the Theater im Zimmer, the Junge Theater, and the Altonaer Theater, which receive only meagre support from the state. Plays of a local character or in Plattdeutsch (Low German) are performed in the Ohnsorg-Theater and sometimes also in the Sankt Pauli-Theater, which dates from 1841 and is Hamburg's oldest playhouse.

The birthplace of Mendelssohn and of Brahms, Hamburg has a sustained tradition of musical activity. Three great orchestras—the Philharmonische Staatsorchester, the Symphonie-Orchester des Norddeutschen Rundfunks, and the Hamburger Symfoniker—familiarize the public with classical and contemporary compositions. There are also groups specializing in chamber music, in choral performances, and in church music; and orchestras, choirs, singers, and instrumentalists from other parts of Germany or from the outside world are also invited to Hamburg. Nearly all recitals are given in the Musikhalle, a building in the Neo-Baroque style founded in 1904–08 by the shipowner Carl Laeisz.

Sports have long been popular in Hamburg. The Hamburger Turnerschaft (1816) is Germany's most ancient athletic club. The first German rowing club, founded in Hamburg in 1836, took part in 1837 in Germany's first official rowing race—against the English Rowing Club formed by members of Hamburg's then numerous English colony. The Hamburger Rennklub, for horseracing, was founded in 1852; and the North German Derby, first run in 1869, became an annual event, as the German Derby, from 1889 onward. Hamburg's first public football matches were played in 1881–82, after disputes about the rules of the game with the local Anglo-American Football Club. Another noteworthy event is the annual German tennis championships.

BIBLIOGRAPHY. For an introduction to contemporary Hamburg, the English reader must consult guidebooks dealing with West Germany as a whole or with its northern territory specifically. The best historical presentation is still that by HEINRICH REINCKE, *Hamburg: Ein Abriss der Stadtgeschichte von den Anfängen bis zur Gegenwart* (1926), supplemented by ERICH VON LEHE, DIETRICH KAUSCHE, and HEINZ RAMM, *Heimatchronik der Freien und Hansestadt Hamburg* (1958).

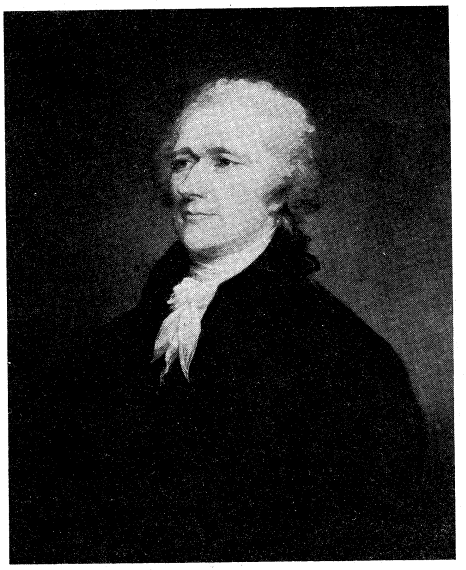
(He.Th.)

Hamilton, Alexander

Alexander Hamilton, publicist, politician, and first secretary of the treasury of the United States under George Washington, the first president, was the foremost champion of a strong central government for the newly united former colonies. As treasury secretary, he created the Bank of the United States, secured the nation's credit both at home and abroad, and established a national currency. A leader of the Federalist Party and chief author of *The Federalist Papers*, which propagated his political and financial views, his image of the United States as a prosperous industrial nation has proved singularly prescient.

Hamilton was born on the island of Nevis in the British West Indies, probably on January 11, 1755, and not in 1757 as he himself claimed. His father was James Hamilton, a drifting trader and the fourth son of Alexander Hamilton, the laird of Cambuskeith, Ayrshire, Scotland; his mother was Rachel Fawcett Lavine, the daughter of a French Huguenot physician and the wife of John Michael Lavine, a German or Danish merchant who had settled on the island of St. Croix in what were then the Danish West Indies. Rachel probably began living with James Hamilton in 1752, but Lavine did not divorce her until 1758. Although he came of good stock, Alexander Hamilton from his earliest years was sensitive to the fact that he was illegitimate. "My blood," he once said, "is as good as that of those who plume themselves upon their ancestry."

In 1765, after having brought his family to St. Croix, James Hamilton abandoned it. Destitute, Rachel set up a small shop, and at the age of 11 Alexander went to work, becoming a clerk in the countinghouse of two New York merchants who had recently established themselves at St. Croix. When Rachel died in 1768, Alexander be-



Alexander Hamilton, oil painting by John Trumbull (1756–1843). In the National Gallery of Art, Washington, D.C.
By courtesy of the National Gallery of Art, Washington, D.C. Andrew Mellon Collection

came a ward of his mother's relatives and continued to work in the countinghouse. In March 1772 young Hamilton's ability, industry, and engaging manners won him advancement from bookkeeper to manager, and several months later friends who were impressed by Hamilton's talents sent him to the North American mainland to further his education. For a year he studied at a preparatory school in Elizabethtown, New Jersey, and in the autumn of 1773 entered King's College (later Columbia) in New York. Intensely ambitious, he became a serious and successful student, but his studies were interrupted by the brewing revolt against Great Britain. Throwing himself behind the colonial cause, he publicly defended the Boston Tea Party, the name given to the destruction of several tea cargoes by the Boston colonists in defiance of the tea tax. In 1774–75, at the age of 17, he wrote three influential pamphlets, *A Full Vindication of the Measures of Congress*, *The Farmer Refuted*, and *Remarks on the Quebec Bill*. They upheld the agreements on the nonimportation, nonconsumption, and nonexportation of British products arrived at by the Continental Congress, the first federal body of the Thirteen Colonies, and attacked British policy in Quebec. Those anonymous publications, the first of which had been attributed to John Jay and John Adams, two of the ablest of American propagandists, gave the first solid evidence of Hamilton's precocity that astonished his contemporaries.

Service in Revolutionary War

In March 1776, through the influence of friends in the New York legislature, Hamilton was commissioned a captain in the provincial artillery. He organized his own company and at the Battle of Trenton, when he and his men prevented the British under Lord Cornwallis from crossing the Raritan River and attacking George Washington's main army, showed conspicuous bravery. Thereafter Hamilton's military brilliance was recognized. In February 1777 Washington invited him to become an aide-de-camp with the rank of lieutenant colonel. In his four years on Washington's staff he grew close to the general and was entrusted with his correspondence. He was sent on important military missions and, thanks to his fluent command of French, he became liaison officer between Washington and the French generals and admirals. During that time he formed some of the important friendships he was to retain for the rest of his life. He also attracted the attention of political leaders in New York. During the inactive periods of the winter campaigns he continued his studies, reading deeply in the classics and particularly in the literature of economics and public finance.

Eager to connect himself with wealth and influence, Hamilton married Elizabeth, the daughter of Gen. Philip Schuyler, the head of one of New York's most distinguished families, in December 1780. Meantime, having grown tired of the routine duties at headquarters and yearning for military glory, he pressed Washington for an active command in the field. Washington refused, and in February 1781, Hamilton impetuously seized upon a trivial quarrel to break with the General and leave his staff position. Fortunately, he had not forfeited the General's friendship, for in July Washington gave him command of a battalion. At the siege of Cornwallis' army at Yorktown in October, Hamilton gained a moment of glory in an assault on a British stronghold.

Early political activities. Meanwhile, in letters to a member of Congress and to Robert Morris, the superintendent of finance in 1780 and 1781, Hamilton had analyzed what he considered the financial and political weaknesses of the government. In November 1781, with the war virtually over, he moved to Albany where he studied law and was admitted to practice in July 1782. That same month he became receiver of continental taxes for the state of New York, a post he gave up a few months later after the New York legislature had elected him to the Continental Congress. Between July 1781 and July 1782 he wrote six essays for a newspaper, the *New York Packet*, under the pen name of "The Continentalist," in which he argued for a strong central government. In Congress from November 1782 to July 1783 he worked for the same end, being convinced that the Articles of Confederation, the nation's first constitution, were the source of the country's weakness and disunion (see UNITED STATES, HISTORY OF THE: *The Confederation and the framing of the Constitution*).

Election to the Continental Congress

In November 1783 Hamilton began to practice law in New York City. He defended unpopular Loyalists who had remained faithful to the British during the Revolution in suits brought against them under a state law of that year called the Trespass Act. Using the pseudonym "Phocion," he published two pamphlets in 1784 pleading for moderation and justice in the treatment of Loyalists and in 1786, partly as a result of his efforts, state acts disbarring Loyalist lawyers and disfranchising Loyalist voters were repealed. In that year he also won election to the lower house of the New York legislature, taking his seat in January 1787. In the meantime, the legislature had appointed him a delegate to the convention in Annapolis, Maryland, that met in September 1786 to consider the commercial plight of the Union. Hamilton suggested that the convention exceed its delegated powers and call for another meeting of representatives from all the states to discuss various problems confronting the nation. He drew up the draft of the address to the states from which emerged the Constitutional Convention that met in Philadelphia in May 1787. After persuading New York to send a delegation, Hamilton, through the influence of his father-in-law, General Schuyler, obtained a place for himself on the delegation.

Hamilton went to Philadelphia as an uncompromising nationalist who wished to replace the Articles of Confederation with a strong centralized government, but he did not take much part in the debates and his attendance was irregular. He served on two important committees, one on rules in the beginning of the convention and the other on style at the end of the convention. In one long speech on June 18 he presented his own idea of what the national government should be. His model was the England of George III. "I have no scruple in declaring . . . that the British government is the best in the world," he said, "and that I doubt much whether anything short of it will do in America." He suggested a government of three departments—legislative, executive, and judicial. The legislature would consist of an assembly or lower house elected for three years by free male citizens and a senate chosen indirectly by electors for life. The president, who also would hold office for life and was to be selected by a double set of electors, would have an absolute veto over the legislature. The central government would appoint the state governors who would have an absolute veto over state legislation. The judiciary would consist of a su-

preme court whose justices would have life tenure. Although the states were to be preserved, they would have virtually no power. The most significant feature of the plan, which was essentially monarchical, was that the national government would have unlimited sovereignty. Other than to reveal his own conservative ideas, Hamilton's plan had little impact on the convention; the delegates went ahead to frame a constitution that, while it gave strong power to a federal government, yet stood some chance of being accepted by the people. Since the other two delegates from New York, who were strong opponents of a Federalist constitution, had withdrawn from the convention, New York was not officially represented and Hamilton had no power to sign for his state. Nonetheless, even though he knew that his state wished to go no further than a revision of the Articles of Confederation, he signed the new constitution as an individual.

The Federalist

Opponents in New York quickly attacked the Constitution and Hamilton answered them in the newspapers under the signature Caesar. Since the Caesar letters did not win the influence he desired, Hamilton turned to another classical pseudonym, Publius, and to two collaborators, James Madison, the delegate from Virginia, and John Jay, the secretary of foreign affairs, to write *The Federalist*, a series of 85 essays in defense of the Constitution and republican government that appeared in the New York newspapers between October 1787 and May 1788. Hamilton wrote at least two-thirds of the essays, including some of the most important ones that interpreted the Constitution, explained the powers of the executive, the senate, and the judiciary, and expounded the theory of judicial review, the power of the Supreme Court to declare legislative acts unconstitutional and thus, void. Although written and published in haste, *The Federalist* was widely read, had a great influence on contemporaries, became one of the classics of political literature, and helped shape American political institutions. In January 1788 Hamilton was reappointed a delegate to the Continental Congress from New York. At New York's ratifying convention in Poughkeepsie, in June, where he was a delegate, he became the chief champion of the Constitution and, against strong antifederalist opposition, won approval for it.

Hamilton's financial program. President Washington launched the new federal government in April 1789 and in September appointed Hamilton the first secretary of the treasury. Congress quickly asked Hamilton to draw up a plan for the "adequate support of the public credit." This was all he needed. Envisaging himself as something of a prime minister in Washington's official family, Hamilton from this point on developed and carried out a bold and masterly program designed to build a strong union, one that would weave his political philosophy into the government. His immediate objectives were to establish the nation's credit at home and abroad and to strengthen the national government at the expense of the states. He outlined his program in four notable reports to Congress. In the first two, *Reports on the Public Credit*, which he submitted on January 14, 1790, and December 13, 1790, he urged the funding of the national debt at full value, the assumption in full by the federal government of debts incurred by the states during the Revolution, and a system of taxation to pay for the assumed debts. His motive was as much political as economic. Through payment by the central government of the states' debts he hoped to bind the men of wealth and influence, who had acquired most of the domestically held bonds, to the national government. But such powerful opposition arose to the funding and assumption scheme that Hamilton was able to push it through Congress only after he had made a bargain with Thomas Jefferson, who was then secretary of state, whereby he gained southern votes in Congress for it in exchange for his own support in locating the future national capital on the banks of the Potomac.

Hamilton's third report, the *Report on a National Bank*, which he submitted on December 14, 1790, advocated a national bank called the Bank of the United States and modelled after the Bank of England. With the bank he wished to solidify the partnership between the government and the business classes who would benefit most from it and further advance his program to strengthen the national government. After Congress passed the bank charter, Hamilton persuaded Washington to sign it into law. He advanced the argument that the Constitution was the source of implied as well as enumerated powers and that through implication the government had the right to charter a national bank as a proper means of regulating the currency. This doctrine of implied powers became the basis for interpreting and expanding the Constitution in later years. In the *Report on Manufactures*, the fourth, the longest, the most complex, and the most farsighted of his reports, submitted on December 5, 1791, he proposed to aid the growth of infant industries through various protective laws. Basic to it was his idea that the general welfare required the encouragement of manufacturers and that the federal government was obligated to direct the economy to that end. In writing his report, Hamilton had leaned heavily on *The Wealth of Nations*, written in 1776 by the Scottish political economist Adam Smith, but he revolted against Smith's laissez-faire idea that the State must keep hands off the economic processes, which meant that it could provide no bounties, tariffs, or other aid. The report had greater appeal to posterity than to Hamilton's contemporaries, for Congress did nothing with it.

Establishment of the Bank of the United States

Establishment of political parties. One of the results of the struggle over Hamilton's program and over issues of foreign policy was the emergence of national political parties. Like Washington, Hamilton had been one of the foremost among the Founding Fathers who had deplored parties, equating them with disorder and instability. He had hoped to establish a government of superior persons who would be above party. Yet he became the leader of the Federalist Party, a political organization in large part dedicated to the support of his policies. Hamilton placed himself at the head of that party because he needed organized political support and strong leadership in the executive branch to get his program through Congress. Washington usually supported Hamilton's policies and in effect became a Federalist. The political organization that challenged the Hamiltonians was the Republican Party created by James Madison, who was at this time a member of the House of Representatives, and Secretary of State Thomas Jefferson. In foreign affairs Hamilton and the Federalists favoured close ties with England, whereas Jefferson and his followers preferred to strengthen the old attachment to France. In attempting to carry out his program, Hamilton frequently interfered in Jefferson's domain of foreign affairs. Detesting the French Revolution and the equalitarian doctrines it spawned, he tried to thwart Jefferson's policies that might aid France or injure England and to induce Washington to follow his own ideas in foreign policy. Hamilton went so far as to warn British officials of Jefferson's attachment to France and to suggest that they bypass the Secretary of State and instead work through himself and the President in matters of foreign policy. This and other parts of Hamilton's program led to a feud with Jefferson from 1791 to 1793 in which the two men attempted to drive each other from the Cabinet. They even used party newspapers to attack each other and the policies each espoused, and each tried to turn Washington against the other.

Feud with Jefferson

When war broke out between France and England in February 1793, Washington asked his Cabinet's advice in formulating American policy toward the conflict. Hamilton wished to use the war as an excuse for jettisoning the French alliance of 1778 and steering the United States closer to England, whereas Jefferson insisted that the alliance was still binding. While not challenging the validity of the alliance, Washington essentially accepted Hamilton's advice and in April issued a proclamation of neutrality that Republicans said favoured England. In that month an emissary from republican France, Edmond-Charles Genet, called Citizen Genet, reached the United States. In trying to advance the cause of his own country he violated American neutrality by arming privateers in U.S. ports. Hamilton in June began a series of articles under the name "Pacificus" in defense of the

neutrality proclamation and in August another series under the title "No Jacobin" that condemned Genet's activities. He also took the lead in demanding Genet's recall, which, despite Jefferson's initial opposition, was accomplished early in 1794.

Jay treaty

At the same time, British violations of American neutrality, such as the seizure of American ships trading with the French West Indies, and other grievances led to strident popular demands for war against Britain, a clamour that Hamilton steadfastly opposed. He believed that war with England would be national suicide, for his program was anchored on trade with Britain and on the import duties that supported his funding system. To save his program, Hamilton persuaded the President to send John Jay to London to negotiate American grievances. Hamilton wrote Jay's instructions, manipulated the negotiations, and helped defend the unpopular treaty Jay brought back in 1795, notably in a series of essays he wrote for the New York newspapers beginning in July 1795 and continuing in 1796 under the signature "Camillus." As Hamilton planned it, the treaty kept the peace and saved his system.

The Republicans in Congress, meantime, had tried to destroy Hamilton. In January 1793 they had demanded a full accounting of treasury operations from the beginning until the end of 1792. Hamilton promptly offered the information. A month later William Branch Giles of Virginia introduced a number of resolutions in the House of Representatives charging Hamilton with dereliction of duty and irregularities in the management of the treasury. Hamilton's friends defended him so well that when the resolutions came to a vote, he was vindicated on every count. In 1794 his program was also challenged from another direction, by farmers in western Pennsylvania, who rose against an excise tax he had been instrumental in placing on distilled liquors, an important source of income for the frontiersmen. He had advanced the tax to raise money and to strengthen the power of the federal government at the expense of the states in an area of taxation that they had considered theirs exclusively. To Hamilton, who urged Washington to use every means to suppress the resistance, the Whiskey Rebellion presented an opportunity he had long desired, a chance for the first time to test the power of the federal government, as opposed to local defiance, by a showdown of arms. Congress authorized the president to call out the militia, and Hamilton in September took a leading part in the expedition sent to chastise the rebels, who melted away before the overwhelming federal force. Hamilton looked upon the suppression as a triumph for law and order that added immeasurably to the prestige of the federal government.

Out of the Cabinet. Lashed by criticism, tired and anxious to repair his private fortune, Hamilton left the Cabinet on January 31, 1795. He did not, however, sever his connection with the government; his influence, as an unofficial adviser, continued as strong as ever. Washington and his Cabinet consulted him on almost all matters of policy. When Washington decided to retire, he turned as usual to Hamilton, asking his opinion as to the best time to publish a farewell to the nation. With his eye on the coming presidential election, Hamilton advised withholding the announcement until a few months before the meeting of the presidential electors. Following that advice, Washington gave his *Farewell Address* to the nation in September 1796. Hamilton drafted most of the address, one of the great documents of American history, and some of his ideas were prominent in it. In the election, Federalist leaders passed over Hamilton's claims and nominated John Adams for the presidency and Thomas Pinckney for the vice presidency. Since Adams did not appear devoted to Hamiltonian principles, Hamilton tried to manipulate the electoral college so as to make Pinckney president. Adams won the election and Hamilton's intrigue succeeded only in sowing distrust within his own party. Nonetheless, Hamilton's influence in the government continued, for Adams retained Washington's Cabinet and its members consulted Hamilton on all matters of policy, gave him confidential information, and in effect urged his policies on the President.

Early in 1797 James T. Callender, a Republican hack, published a *History of the United States for the Year 1796* in which he accused Hamilton of corruption in connection with an affair Hamilton had had six years earlier with an attractive adventuress, Mrs. Maria Reynolds. Hamilton met the attack by writing a pamphlet in which he confessed the "irregular and indelicate amour" and printed the blackmailing letters that the woman's husband, a confidence man, had sent to him, but denied any corrupt dealings with him. Although Hamilton successfully defended his integrity as a public man, he subjected his private life to a bitter humiliation.

In retaliation against Jay's treaty, France meantime had broken relations with the United States. Although he stood for firmness, Hamilton agreed with the president's policy of trying to re-establish friendly relations. After the failure of a peace mission President Adams had sent to Paris in 1798, followed by the publication of certain dispatches insulting to American sovereignty, Hamilton wanted to place the country under arms immediately. He even believed that the French, who had embarked on an undeclared naval war against the United States, might attempt to invade the country. Still eager for military glory, Hamilton sought command of the new army, though Washington would be its titular head. Adams resisted Hamilton's desires, but in September 1798 Washington forced him to make Hamilton second in command of the army, the inspector general, with the rank of major general. Adams never forgave Hamilton for this humiliation. Hamilton wanted to lead his army into Spain's Louisiana and the Floridas and other points south, but never did. Through independent diplomacy, Adams kept the quarrel from spreading into a full-scale conflict and at the order of Congress disbanded the provisional army. With his dream of military glory faded, Hamilton resigned his commission in June 1800. Meantime Adams had purged his cabinet of those he regarded as "Hamilton's spies."

Attempt to command the army

In retaliation, Hamilton tried to prevent Adams' re-election to the presidency. In October 1800 he privately circulated a bitter personal attack on Adams, entitled *The Public Conduct and Character of John Adams, Esq., President of the United States.* Aaron Burr of New York, the Republican candidate for vice president and Hamilton's political enemy, obtained a copy and had it published. Hamilton was then compelled to acknowledge his authorship and to bring his quarrel with Adams into the open, a feud that revealed an irreparable schism in the Federalist Party. Thomas Jefferson and Aaron Burr won the election but because both had received the same number of electoral votes, the choice between them for president was cast into the House of Representatives. Hating Jefferson, the Federalists wanted to throw the election to Burr. Hamilton helped to persuade them to select Jefferson instead. By supporting his old Republican enemy, who won the presidency, Hamilton lost prestige within his own party and virtually ended his public career.

The Burr quarrel. Hamilton then concentrated on his law practice and his family. In 1801 he built a country house on Manhattan Island called The Grange, and helped found a Federalist newspaper, the *New York Evening Post.* Although Hamilton did not write for it under his own name, its policies reflected his ideas. Through the *Post* he hailed the purchase of Louisiana in 1803, even though New England Federalists had opposed it. Some of them talked of secession and in 1804 began to negotiate with Aaron Burr for his support. Almost all the Federalists but Hamilton favoured Burr's candidacy for the governorship of New York in that year. Hamilton urged the election of Burr's Republican opponent, who won by a close margin, but it is doubtful that Hamilton's influence decided the outcome. In any event, Hamilton and Burr had long been enemies, and Hamilton had several times thwarted Burr's ambitions. In June 1804, after the election, Burr demanded satisfaction for remarks Hamilton had allegedly made at a dinner party in April in which he said he held a "despicable opinion" of Burr. Although Hamilton held an aversion to the practice of dueling, as a man of honour he felt himself compelled to accept

Burr's challenge. If he declined, his reputation as a gentleman would be lost, as would whatever political influence he still retained. The two antagonists met early in the morning of July 11 on the heights of Weehawken in New Jersey where Hamilton's eldest son, Philip, a young man of 19, had died in a duel three years before. Burr's bullet found its mark and Hamilton fell. "This is a mortal wound, Doctor," he said and became unconscious. He died the following day. A poor man, Hamilton left his wife and seven children heavily in debt; friends, however, paid off the debts.

Death

Assessment. Hamilton was a small, slender man, about five feet seven inches tall, with a fair complexion, light hair, strong features, and deep-set eyes. He had an erect, dignified bearing, made striking by elegant clothes and fine manners. Some people found him stiff and formal, even vain and arrogant, but friends saw him as witty, warm, benevolent, and loyal. Although a fine orator, he was more persuasive as an essayist. He was both a man of action and of ideas, but all of his ideas involved action and were directed toward some specific goal in statecraft. Unlike Benjamin Franklin or Thomas Jefferson, he did not have a broad inquisitive mind nor was he speculative in his thinking in the philosophical sense of seeking intangible truths. He was ambitious, purposeful, a hard worker, and one of America's administrative geniuses. In foreign policy he was a realist, believing that self-interest should be the nation's polestar; questions of gratitude, benevolence, and moral principle, he held, were irrelevant.

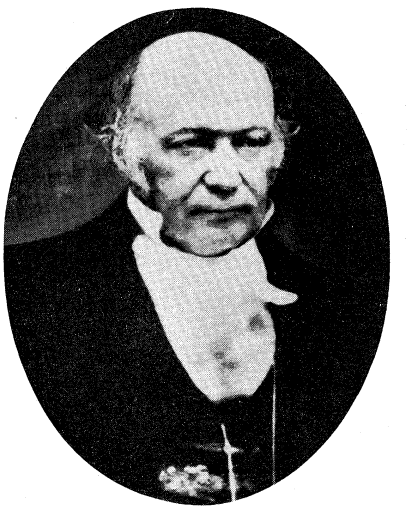
Most of all, Hamilton was one of America's first great nationalists. He believed in an indivisible nation where the people would give their loyalty not to any state but to the nation. Although a conservative, he did not fear change or experimentation. The conservatism that led him to denounce democracy as hostile to liberty stemmed from his fear that democracy tended to invade the rights of property, which he held sacred. His concern for property was a means to an end. He wished to make private property sacred because upon it he planned to build a strong central government, one capable of suppressing internal disorders and assuring tranquillity. His economic, political, military, and diplomatic schemes were all directed toward making the Union strong. Hamilton's most enduring monument was the Union, for much of it rested on his ideas.

BIBLIOGRAPHY. The most satisfactory complete biography, particularly strong on Hamilton's public career, is JOHN C. MILLER, *Alexander Hamilton: Portrait in Paradox* (1959). The most scholarly, but uncompleted, study is BROADUS MITCHELL, *Alexander Hamilton*, 2 vol. (1957–62). NATHAN SCHACHNER, *Alexander Hamilton* (1946), is a short, well-balanced, and readable study; even briefer is STUART G. BROWN, *Alexander Hamilton* (1967). For Hamilton's economic and political ideas, see LOUIS M. HACKER, *Alexander Hamilton in the American Tradition* (1957); CLINTON L. ROSSITER, *Alexander Hamilton and the Constitution* (1964); and GERALD STOURZH, *Alexander Hamilton and the Idea of Republican Government* (1970). BROADUS MITCHELL, *Alexander Hamilton: The Revolutionary Years* (1970), covers the military career. Three books analyze Hamilton's influence in shaping the foundations of foreign policy: JULIAN P. BOYD, *Number 7: Alexander Hamilton's Secret Attempts to Control American Foreign Policy* (1964); HELENE J. LOOZE, *Alexander Hamilton and the British Orientation of American Foreign Policy, 1783–1803* (1969); and GILBERT L. LYCAN, *Alexander Hamilton and American Foreign Policy: A Design for Greatness* (1970). Publication of a multivolume work designed to be definitive, *The Papers of Alexander Hamilton*, ed. by H.C. SYRETT and J.E. COOKE, began in 1961. The following are biographical accounts of varying quality: DAVID G. LOTH, *Alexander Hamilton: Portrait of a Prodigy* (1939); HENRY JONES FORD, *Alexander Hamilton* (1920); ALLAN M. HAMILTON, *The Intimate Life of Alexander Hamilton* (1910); FREDERICK S. OLIVER, *Alexander Hamilton: An Essay on American Union* (1906); WILLIAM GRAHAM SUMNER, *Alexander Hamilton* (1890); HENRY CABOT LODGE, *Alexander Hamilton* (1882, many later editions); JOHN T. MORSE, JR., *The Life of Alexander Hamilton*, 2 vol. (1876); and the works of Hamilton's son, JOHN C. HAMILTON, *The Life of Alexander Hamilton*, 2 vol. (1840–41), an unfinished biography that goes only to 1787; and his valuable but uncritical *History of the Republic of the United States of America, As Traced in the Writings of Alexander Hamilton and of His Contemporaries*, 3rd ed., 7 vol. (1868).

(A.De C.)

Hamilton, Sir William Rowan

William Rowan Hamilton, who excelled in modern languages and the classics, brought honour to his native Ireland with his achievements in mathematics. His unification of dynamics and optics have had a lasting influence on mathematical physics, even though the full significance of his work was not fully appreciated until after the rise of quantum mechanics. He also contributed to the development of modern algebra.



William Rowan Hamilton, 1862.
By courtesy of the Royal Irish Academy, Dublin

Early life. Hamilton was born at midnight, August 3/4, 1805, in Dublin, Ireland. Like his English contemporaries Thomas Babington Macaulay and John Stuart Mill, he showed unusual intellect as a child. Before the age of three his parents sent him to live with his father's brother, James, a learned clergyman and schoolmaster at a Church of England school at Trim, a small town near Dublin, where he remained until 1823, when he entered Trinity College, Dublin. Within a few months of his arrival at his uncle's he could read English easily and was advanced in arithmetic; at five he could translate Latin, Greek, and Hebrew and recite Homer, Milton, and Dryden. Before his 12th birthday he had compiled a grammar of Syriac, and by the age of 14 he had sufficient mastery of the Persian language to compose a welcome to the Persian ambassador on his visit to Dublin.

Hamilton became interested in mathematics after a meeting in 1820 with Zerah Colburn, an American who could calculate mentally with astonishing speed. Having read the *Eléments d'algèbre* of Alexis-Claude Clairaut and Isaac Newton's *Principia*, Hamilton had immersed himself in the five volumes of Pierre-Simon Laplace's *Traité de mécanique céleste* (1798–1827; *Celestial Mechanics*, 1966) by the time he was 16. His detection of a flaw in Laplace's reasoning brought him to the attention of John Brinkley, professor of astronomy at Trinity College. When Hamilton was 17, he sent Brinkley, then president of the Royal Irish Academy, an original memoir about geometrical optics. Brinkley, in forwarding the memoir to the Academy, is said to have remarked: "This young man, I do not say *will be*, but *is*, the first mathematician of his age."

College and early career. In 1823 Hamilton entered Trinity College, from which he obtained the highest honours in both classics and mathematics. Meanwhile, he continued his research in optics and in April 1827 submitted his "Theory of Systems of Rays" to the Academy. The paper transformed geometrical optics into a new mathematical science by establishing one uniform method for the solution of all problems in that field. Hamilton started from the principle, originated by the 17th-century

French mathematician Pierre de Fermat, that light takes the shortest possible time in going from one point to another, whether the path is straight or is bent by refraction. Hamilton's key idea was to consider the time (or a related quantity called the "action") as a function of the end points between which the light passes and to show that this quantity varied when the coordinates of the end points varied, according to a law that he called the law of varying action. He showed that the entire theory of systems of rays is reducible to the study of this characteristic function.

Rapid rise to prominence

Shortly after Hamilton submitted his paper and while still an undergraduate, Trinity College elected him to the post of Andrews Professor of Astronomy and Royal Astronomer of Ireland, to succeed Brinkley who had been made a bishop. Thus an undergraduate (not quite 22 years old) became ex officio an examiner of graduates who were candidates for the Bishop Law Prize in mathematics. The electors' object was to provide Hamilton with a research post free from heavy teaching duties. Accordingly, in October 1827 Hamilton took up residence next to Dunsink Observatory, five miles (eight kilometres) from Dublin, where he lived for the rest of his life. He proved to be an unsuccessful observer, but large audiences were attracted by the distinctly literary flavour of his lectures on astronomy. Throughout his life Hamilton was attracted to literature and considered the poet Wordsworth among his friends, although Wordsworth advised him to write mathematics rather than poetry.

Domestic life and honours. Six years after his move to Dunsink, Hamilton married Maria Bayley, daughter of a former rector in County Tipperary, who bore him two sons and a daughter. But his wife was less successful at running the household; as a result, Hamilton never had regular meals and came to rely excessively on alcohol. He would usually work all day in the dining room, and the cook would bring him a mutton chop from time to time. After his death scores of bones were found on plates sandwiched among his papers.

In 1835 Hamilton was the chief local organizer of the British Association for the Advancement of Science meeting in Dublin and at the closing dinner was knighted by the Lord Lieutenant. Two years later he became president of the Royal Irish Academy. In 1843 he was awarded a Civic List life pension of £200 a year by the British government.

During Hamilton's last illness, an attack of gout, he received with great satisfaction the news that his name had been placed at the head of the first list of Foreign Associates elected by the newly formed National Academy of the United States. He died on September 2, 1865, after months of suffering.

Life work and significance. In 1832 a supplement to Hamilton's theory of rays was published. In it he predicted that, as a result of the theory, a wholly unexpected phenomenon would be found in connection with the refraction of light in biaxial crystals, which produce interference figures consisting of two sets of concentric rings when light passes through them. It had been known for some time that certain crystals of this kind, such as topaz, give rise to two refracted rays for each incident ray. The theory of this double refraction had been worked out a few years earlier by Augustin Fresnel. Hamilton found by his general method that under certain conditions a single ray of incident light could actually produce an infinite number of refracted rays in a biaxial crystal and that they would form a cone. Hamilton's prediction of conical refraction, regarded in his lifetime as his most brilliant achievement in optics, was confirmed experimentally within two months by a colleague, Humphrey Lloyd.

"General Method in Dynamics"

Today his unification of optics and dynamics is regarded as far more important than his work on conical refraction. In 1835 his memoir, "On a General Method in Dynamics," was published. In it he applied his idea of the characteristic function to the motion of systems of bodies and expressed the equations of motion in a form that revealed the duality between the components of momentum of a dynamical system and the coordinates determining its position. Although Hamilton's canonical equations expressing this duality and his principle that reduces the whole of dynamics to a problem in the calculus of variations, have long been familiar to students of dynamics, the deep significance of the duality he discovered was not appreciated for nearly 100 years, until the rise of quantum mechanics.

Discovery of quaternions

In the same year Hamilton made his famous discovery of quaternions; these ordered sets of four ordinary numbers, satisfying special laws of equality, addition, and multiplication, are useful for studying quantities having magnitude and direction in three-dimensional space. This discovery was a landmark, since it freed algebra from the commutative postulate of multiplication—that the order or sequence of factors does not determine the result. His investigations in algebra had begun ten years before with a pioneer paper on algebraic couples of numbers in which the basic entity was not a single number but ordered pairs of numbers. Hamilton used this idea to develop a rigorous theory of complex numbers, involving the square root of -1. This paper was remarkable as a pioneer attempt to put algebra on an axiomatic basis like geometry. The geometry of complex numbers (*i.e.*, numbers of the form $a + bi$, in which i is the square root of -1) is that of the two-dimensional vectors in a plane. In attempting to develop an analogous technique for three-dimensional space, Hamilton was delayed for many years by a fundamental difficulty that could not be resolved so long as he restricted his attention to *triplets*. Suddenly, on October 16, 1843, the solution flashed into his mind as he was walking to Dublin along the Royal Canal: geometrical operations in three-dimensional space require not triplets but *quadruplets*. The reason is that, whereas the algebraic couple suffices in the plane because it is equivalent to a multiplier and an angle, in three dimensions the orientation of the plane itself is variable, and this gives rise to two more numbers. Hamilton was so excited by his discovery that as he passed Brougham Bridge he cut the fundamental formulas of quaternions on the stonework: $i^2 = j^2 = k^2 = ijk = -1$.

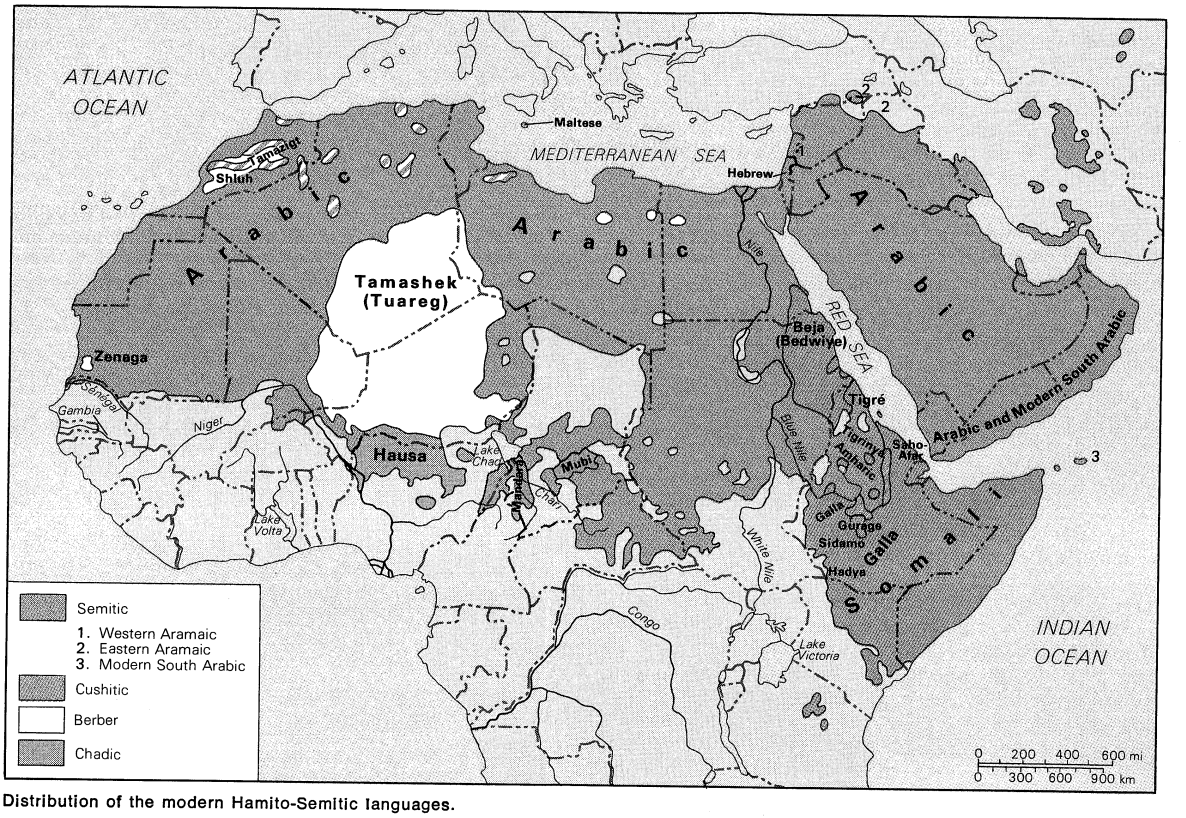
Hamilton's discovery was a break with tradition because it involved the surrender of the commutative law of multiplication—that b times a is the same as a times b. The remaining 22 years of his life were devoted to developing the algebra of quaternions and its applications. Following his death in Dublin on September 2, 1865, this work was published posthumously in 1866 as *The Elements of Quaternions*. Unfortunately, Hamilton believed that quaternions were ideally suited for the solution of problems in applied mathematics, but it was the simplified version of J. Willard Gibbs, known as vector analysis, that was eventually adopted by mathematical physicists. The value of Hamilton's discovery lay rather in pure mathematics, through its effect on the development of modern abstract algebra.

BIBLIOGRAPHY. A full-scale Victorian style biography, containing Hamilton's poems, correspondence, and miscellaneous writings is R.P. GRAVES, *Life of Sir William Rowan Hamilton*, 3 vol. (1882–89). *The Mathematical Papers of Sir William Rowan Hamilton* have been published in 3 volumes: vol. 1, *Geometrical Optics*, ed. by A.W. CONWAY and A.J. MCCONNELL (1931); vol. 2, *Dynamics*, ed. by A.W. CONWAY and J.L. SYNGE (1940); and vol. 3, *Algebra*, ed. by H. HALBERSTAM and R.E. INGRAM (1967). There are useful introductions by the editors in vol. 1 and vol. 3 and all three volumes have been beautifully produced.

(G.J.W.)

Hamito-Semitic Languages

The Hamito-Semitic languages, a family of genetically related languages, developed from a common parent language that presumably existed about the 6th–8th millennia BC and was perhaps located in the present-day Sahara. Also known as the Semito-Hamitic, Erythraean, Afro-Asiatic, and Afrasian language group, it is the main language family of northern Africa and southwestern Asia and includes such languages as Arabic, Hebrew, Amharic, and Hausa. The total number of speakers is estimated to be about 175,000,000.



Distribution of the modern Hamito-Semitic languages.

The term Hamito-Semitic, or Semito-Hamitic, was introduced by a German Egyptologist, Karl Richard Lepsius, in the 1860s. Although it has become traditional, it is an unfortunate label in suggesting that the family is divided into a group of Semitic and a group of Hamitic languages; in fact, the family has at least four other branches of the same order as the Semitic languages. The term Erythraean is inappropriate in implying that the family originated on both shores of the Red Sea, an assumption that cannot be proved; and Afro-Asiatic (proposed by a U.S. linguist, Joseph Greenberg, in 1950) may be too comprehensive insofar as it suggests that all the languages of Africa and Asia are included. Igor Diakonoff, a Soviet linguist, has suggested the term Afrasian, meaning "half African, half Asiatic," which corresponds to the area of the actual distribution of the languages of this family since at least the 5th millennium BC.

The five branches of Hamito-Semitic

The languages belonging to the Hamito-Semitic family can apparently be subdivided into branches representing dialects of the original parent language—namely, Semitic, Egyptian, Berber, Cushitic, and Chadic. Some linguists deny the genetic affinity of the Chadic languages with the other branches of Hamito-Semitic, while others (*e.g.*, Joseph Greenberg) accept it. Certain scholars have expressed doubts concerning the Hamito-Semitic character of some of the Chadic languages but not of others. Among the linguists who classify the Chadic languages as Hamito-Semitic there is some hesitation as to the degree and character of their affinity with the languages of the Cushitic branch, especially with West Cushitic. On the basis of the low percentage of vocabulary items held in common between the West Cushitic languages and the other Cushitic languages, some scholars classify West Cushitic as a separate branch of Hamito-Semitic, called Omotic. There is, however, a probability that the parent language common to Omotic and the Cushitic languages proper is not the Common Hamito-Semitic protolanguage but a later dialect (namely, Common Cushitic) and that Omotic (West Cushitic) is thus, nevertheless, a subgroup of Cushitic. Others connect Omotic with the Chadic group.

Some linguists have suggested that the Hamito-Semitic languages are related to the Indo-European languages; others have favoured the existence of a superfamily, including the Hamito-Semitic, Indo-European, Altaic, Finno-Ugric (Uralic), Kartvelian, and Dravidian languages; but most scholars regard such far-flung genetic ties as unproven and, indeed, hardly provable.

Because there has been a considerable difference of opinion as to the criteria to be applied when identifying a language as Hamito-Semitic, the basic principles of linguistic classification as applicable in this case should be stated. The only real criterion for classifying certain languages together as a family is the common origin of their most ancient vocabulary as well as of the word elements used to express grammatical relations. A common source language is revealed by a comparison of words expressing notions common to all human cultures (and therefore not as a rule likely to be borrowed from a group speaking another language) and also by a comparison of the inflectional forms (for tense, voice, case, or whatever). If, as a result of a step-by-step reconstruction of forms having existed at earlier periods, scholars arrive at an identical original phonological structure for each of the words or word elements compared in several different known languages, then such original forms can be ascribed to a common language, which, in the case of the languages here discussed, is conventionally termed Common Hamito-Semitic (or Proto-Hamito-Semitic). It also stands to reason that wherever one parent language has existed the daughter languages must to some degree reflect some of its grammatical characteristics.

Reconstruction of Common Hamito-Semitic

Despite the work of several scholars, only an approximate, provisional reconstruction of the parent language forms of Hamito-Semitic has so far been made. More has been done in comparing the language typologies.

COMMON HAMITO-SEMITIC CHARACTERISTICS

Certain typological features seem to have been common to all Hamito-Semitic languages at an early stage of their development. Among the phonological features are (1) a six-vowel system (*a, i, u, ā, ī, ū*—that is, short and long *a, i, u*), perhaps developed from an earlier two-vowel system (of **a*, and **ə* [pronounced as the *a* in "sofa"]; an asterisk before a sound or a word-form indicates that it is not attested but is reconstructed hypothetically); (2) pharyngeal fricative consonants, indicated by the symbols ʿ (voiced) and *ḥ* (voiceless) and produced in the

region of the pharynx; (3) the functioning of the glottal stop (articulated by closing the glottis, the space between the vocal cords) as a separate distinctive sound (phoneme)—this is conventionally indicated by ʾ; (4) the use of the semivowels *u̯* (*w*) and *i̯* (*y*) in the structural role of consonants; and (5) three types of consonants: voiceless, voiced, and "emphatic," the last type being phonetically realized either as voiceless consonants combined with a glottal stop, as pharyngealized voiceless or voiced consonants, or as consonants in which the air is drawn into the mouth (injective [preglottalized], or implosive), consonants in which the tongue is retracted from the usual position (velarized), or in which the tongue tip is curled upward toward the hard palate (retroflex or cerebral).

Common morphological features include (1) word bases for verbs and for nouns derived from verbs consisting of two elements that interweave with one another, a "root" consisting of consonants, and a "scheme" consisting of vowels (for examples see below); (2) a predominance of word roots consisting of three consonants over roots of two consonants; (3) a strongly developed system of infixation—*i.e.*, the insertion of elements within the root of a word to show grammatical changes and form new words with related meaning; and (4) a comparatively poorly developed system of prefixes and suffixes.

Morphological similarities among the languages

In the area of morphological typology, there are numerous similarities among the Hamito-Semitic languages, such as a system of declension of the noun and pronoun with at least three cases (nominative, genitive, accusative, with traces of a still earlier system including only the agentive [ergative], and unmarked [zero] cases, or agentive, genitive, and unmarked). There are three numbers in the noun, pronoun, and verb—singular, dual, and plural. An event considered from the point of view of the resulting state, as opposed to the point of view of the action itself, is expressed by a special predicative (zero) form of the noun that later developed into a new verbal "tense." In addition, there is a well-developed binary system of verbal aspects, indicating the mode of an action (*i.e.*, punctual contrasts with durative, or perfective [completed action] contrasts with imperfective [ongoing action]), but tenses and voices of the verb remained undeveloped until the later stages. Pronominal possession markers and object markers in the form of suffixes are another common Hamito-Semitic feature, as are the prefixing of certain actor markers to the verb and a two-gender system in the noun, pronoun, and verb, perhaps developed from a still earlier system of many genders. In syntax, the Hamito-Semitic languages show certain favoured types of attributive constructions, among other common characteristics.

The above inherited Hamito-Semitic characteristics are listed, for each linguistic level, in the approximate reverse order of their stability. Languages retaining all or most of these features can be classified as belonging to the Ancient Stage of Hamito-Semitic; those that retain no less than two-thirds of the ancient consonantal system and about one-half to two-thirds of the above-listed other features belong to the Middle Stage; those that have lost more than half of these characteristics belong to the New Stage. At the New Stage, however, there are usually enough of these features still preserved to identify the language as belonging to the Hamito-Semitic family, and most of the other features can, as a rule, be reliably reconstructed for one of the former stages of its development. Moreover, the original form of the word elements that express the typical Hamito-Semitic grammatical features is usually apparent in all languages of the family. All modern Hamito-Semitic languages except Literary Arabic and Hebrew belong to the New Stage.

The character of the relationship between the five branches of the Hamito-Semitic family—Semitic, Egyptian, Berber, Cushitic, and Chadic—can best be seen by comparing their systems of verbs and pronouns. There are several types of verbal systems in Hamito-Semitic, but all of them (with the exception of the Egyptian, which has developed in a quite different direction) can apparently be traced back to one single system. In this system the action (including intransitive action) is expressed by a verbal form proper, with a prefixed actor marker (singular: 1st person **ʾa-*, 2nd **ta-*, 3rd **ya-*) probably deriving from a separate personal pronoun in an oblique case; the state is expressed by a form of a noun used as a predicate, plus a personal pronoun in the direct case (this is called ştative). Hamito-Semitic apparently developed from a protolanguage with an ergative type of sentence construction (in which there is a special case denoting the agent of an action but no marker for the subject of a state and the direct object of an action) to a language of the nominative type (in which the subject both of an action and a state is always in the nominative case and the direct object is in the accusative case). At the same time, the predicate of state (the so-called stative) either developed into a perfective aspect (marking completion of the action of the verb) or a past tense of the verb, or it disappeared altogether. There are, however, enough traces of its existence in all branches of the family (*e.g.*, in Egyptian, in Kabyle of the Berber branch, in Sidamo of the Cushitic branch, in Mubi of the Chadic branch, and in all Semitic languages) to see that the form goes back to the parent language.

Hamito-Semitic verbal systems

As for the verbal forms that express action and have a prefixed actor marker, there is some discrepancy of opinion. Some scholars posit for the parent language only one form. It may be, however, that there were two forms for the transitive, a perfective and an imperfective type, and possibly only one form for the intransitive type.

In several languages of the New Stage, new verbal types have developed for all aspects and tenses, particularly in the languages of the Cushitic branch (the Northern, Eastern, and Central groups, in part; and the Southern and Western [Omotic] group, always), the Chadic branch (in most languages), and the Semitic branch (typically in Neo-Syriac). These verbal forms consisted originally of a noun (for the most part, derived from a verb) plus an auxiliary verb with a prefixed actor marker. Everywhere, as a rule, the perfective aspect (or the past tense) is formed from bases of the auxiliary verb with a reduced vowel scheme in the verbal base, while the imperfective aspect (or the present/future tense) is formed from bases with a full vowel scheme (*cf.* the Akkadian perfective form **yaprus* "he divided," with a reduced vowel scheme, and the imperfective form **yaparras* "he divides," with a full vowel scheme). (There are also forms based on the participle of the auxiliary verb; *e.g.*, Neo-Syriac *biktävövin* "I am writing" from **bi-ktābā-hāwē-ʾănā* "in-write-being-I.")

In that the Central Semitic verbal system (which has the imperfective with a reduced vowel scheme, as in Arabic, Hebrew, and Aramaic) is restricted to only two groups of languages inside only one branch of the entire family, it is improbable that it is this verbal system that is descended from the parent language.

Stem modification

A typical feature of the Hamito-Semitic verbal system is the existence of so-called stem modifications—*i.e.*, groups of systematically related verbal stems deriving from a single root, each having its own type of semantics —that variously characterizes the action or state from the point of view of its quality, quantity, frequency, causal relations, direction, and so on. In Hebrew, for example, *šāḇar* "he broke," *šibbēr* "he broke to pieces," *hišbīr* "he let (him) break out," and *nišbar* "he was broken, destroyed, stranded" all are from the root *šbr*.

The pronominal systems in the different branches of Hamito-Semitic are more or less alike. Some pronouns are virtually identical everywhere; *e.g.*, the possessive pronouns (2nd person masculine—"your": Semitic **-ka*, Egyptian *-k*, Berber *-k*, reconstructed Cushitic *-ka* or **-k^{w}a*, Chadic [Hausa] *-ka*). Suffixed pronouns expressing the object of the verb are very similar to the possessive.

Linguistic substrata

The diverging of the branches and the individual languages of the Hamito-Semitic family from the common ancestral language, although mainly explained by the internal development of the languages after loss of contact, also results to a great extent from the influence of different linguistic substrata. Thus, the ancient Hamito-Semitic language had in many cases probably spread to originally alien populations. This view is supported by

Table 1: Common Hamito-Semitic Vocabulary Items

	Semitic	Egyptian	Berber	Cushitic	Chadic
"bone"	*qōṣ* ("thorn," Hebrew)	*qs*	*i-ghəs*	**m-kkac*	**kasi* (Hausa)
"to die"	**mūt*	*m(w)t*	*əmmət*		*mutu* (Hausa)
"dog"	*kal-b-*			**kala-kara-*	**kala-kara-*
"eye"		*ỉr. t*		**yil*	**yil*
"heart"	**libb-*	*ỉb*		**libb-*	
"I"	**'anāku*	*ỉnk*	*n(ə)ki*	*(')ani	*an* (Sura)
	'anā	*anok* (Coptic)		**ta-(')ani*	*nī* (Hausa)
	'anī			*anu*	*n-ani* (Kana-kura)
				ana	
"jackal" ("wolf," "dog")		*wnš*	*ušš-ən*	**waẑẑ* *(from *wanẑ-*?) (Highland Eastern Cushitic)	
"man"	**mut-*	**mt*			*mito* (Jegu) *miji* (Hausa)
"name"	**šim-*		*i-səm*	**sim-*	**sim(m)-summ-sūn-*
"thou"	*'anta*	*ntk*		**'atta*	
	'atta	*ənṭok* (Coptic)			
"tongue"	**laš-ān-*	*ns* (pronounced *las*)	*i-ls*		**ha-ls(e)* (Hausa)
"tooth"	**šinn-*	*sn* ("harpoon")	*-sin*		**sinn-*(?)
"two"	**thin-*	*sn*	*sin*	**can-*("two equal parts")	
"water"	**mā'-may-*	*m(y)-w*	*a-ma-n*	*mā-n* ("sea") (Somali)	

*Reconstructed form.

the different racial types of the speakers. In some cases the substratum language (*i.e.*, that of the original population) can be identified—*e.g.*, Sumerian, Hurrian, and others for North Semitic; Nilo-Saharan and East Sudanese for Cushitic; East Sudanese and possibly some others for Chadic. The least substratum influence seems to have been experienced by the Berber branch.

SEMITIC LANGUAGES

Languages of the group. The Semitic languages can be subdivided into four groups: the Northern Peripheral, the Northern Central, the Southern Central, and the Southern Peripheral.

Northern Peripheral Semitic. The Northern Peripheral group, from the Ancient to Middle Stage, includes Akkadian with its dialects of Babylonian and Assyrian, spoken in Mesopotamia from about 3200 BC to the beginning of the Christian Era. Typical features are stative verb forms conjugated with suffixes and two verbal forms with a prefixed actor marker for the imperfective and perfective (with full and reduced vowel schemes, respectively; later a new "perfect" with an infixed *-ta-* in the stem developed). Originally there were five cases of the noun, plus an unmarked form for the nominal predicate and the noun without grammatical relations. Later three cases remained but were lost in the 1st millennium BC. Loss of the *gh*, *ḥ*, *ʿ*, and *h* sounds occurred from *c.* 2000 BC. The vowels were *a*, *e*, *i*, *u* (both long and short).

Northern Central Semitic. The Northern Central Semitic group includes the Canaanite, Ugaritic, and Amorite languages of the Ancient Stage, which were spoken in Palestine, Phoenicia, Syria, and Mesopotamia from the 3rd to the 2nd millennium BC. To the Middle Stage belongs Phoenician-Punic, spoken in Phoenicia, on islands of the Mediterranean, and in North Africa, from the 2nd millennium BC to the 1st millennium AD. Also to the Middle Stage belong Hebrew, Moabite, Ya'ūdī, and Old Aramaic. Hebrew, originally spoken in Palestine from the 13th century BC to the 2nd century AD, later spread all over the world as a written language. At present there are about 3,000,000 Hebrew speakers in Israel. Moabite and its kindred dialects in the Transjordan were alive in the 1st millennium BC but are now extinct. Ya'ūdī, spoken in northern Syria in the 9th century BC, is also extinct. Old Aramaic, from Syria and Mesopotamia, existed from the 14th century BC(?) through the 15th century AD. Its oldest written texts date from the 9th century BC. The dialects of Aramaic include Ancient Aramaic proper; Imperial Aramaic (the official language of Assyria and Achaemenid Persia, including also Biblical Aramaic, or Chaldean); Western Aramaic, with Palmyrenean, Nabataean, Palestinian, Galilean, and other varieties; Eastern Aramaic, including Syriac (Edessan, with subdialects), Babylonian Talmudic, and Mandaic. Most Aramaic dialects gave way to Arabic beginning with the 7th century AD. Aramaic dialects

The New Stage of Northern Central Semitic is represented by New West Aramaic, or Maʿlūla, in Syria, with a small number of speakers, and Neo-Syriac, or "Assyrian," in Iraq (al-Mawṣil), Turkey (Ṭūr-ʿAbdīn), Iran (Urmia), the Soviet Union, and the United States, with up to 1,000,000 speakers.

Typical features of the Northern Central Semitic group are the perfective aspect with suffix conjugation and the imperfective aspect with prefix conjugation and stems with a reduced vowel scheme. (An entirely new verbal system developed in Neo-Aramaic.) The group is also characterized by the article *ha-* (prefixed, in Hebrew and Punic) or *-ā* (suffixed, in Aramaic). The system of declension was lost from the Middle Stage on. In this group the number of vowel qualities increased beyond just *a*, *i*, *u*, while the number of consonants diminished considerably from the Middle Stage on. The sounds *p*, *t*, *k*, *b*, *d*, *g* became aspirated after vowels—*i.e.*, pronounced with an accompanying puff of breath (and are now pronounced as *f*, *t* or *s*, *kh*, *v*, *d*, *g* in Modern Hebrew).

Southern Central Semitic. The Southern Central group includes Classical, or Literary, Arabic belonging to the Ancient and Middle stages. Originally spoken in Arabia, Classical, or Literary, Arabic is now found from the Indian to the Atlantic Ocean and has been attested from the 5th century BC to the present time. From the New Stage come the modern Arabic dialects, some of them mutually unintelligible. They have about 100,000,000 speakers all over northern Africa, on the Arabian Peninsula, in Jordan, Palestine, Lebanon, Syria, and Iraq, and in some districts of Turkey, Iran, and the Soviet Union. Maltese, on the island of Malta, has developed into a separate language, spoken by about 300,000. A typical feature of the group is a verbal system that is very similar to that of the Northern Central group (with minor differences) but that developed tenses instead of aspects from the late Middle Stage (*e.g.*, in the Egyptian dialect, from prepositional constructions). There were three cases of the noun, but declension was lost at the late Middle and New Stage. Also characteristic of South Central Semitic are the article *al-* and a strongly developed system of internal inflection with the plural mostly of the *pluralis fractus* type ("broken plural," in which plurality is shown by means of internal vowel changes). The Proto-Semitic phonological system has been on the whole well preserved, but *š has become *s*, **ṯh* has become *ẓ*, and other similar changes.

Southern Peripheral Semitic. To the Southern Peripheral group of the Ancient to Middle Stage belong the South Arabic dialects, Sabaean (*cf.* Sheba), Minaean, Qatabānian, and Ḥaḍramawtian, spoken from the 1st millennium BC to the 1st millennium AD. The Middle Stage is represented by Geʿez (Geez or Gəʿəz), or Ethiopic, found in northern Ethiopia in the 1st millennium AD; and the New Stage by the South Arabic group, including Mahrī, Shaḥrī (Eḥkalī), Ḥarsūsī, and Baṭḥarī on the Arabian shore of the Indian Ocean, and Suquṭrī (possibly a dialect of Mahrī) on the island of Socotra, with the total number of speakers probably being around 50,000. Also of the New Stage is the Ethiopic group, consisting of three subgroups: North Ethiopic, Central Ethiopic, and the Gurage subgroup. North Ethiopic, nearest to Geʿez, includes Tigrinya (Tigrai) and Tigré, spoken in northern Ethiopia and Eritrea by 3,700,000 speakers. Within the Central Ethiopic group (7,800,000 speakers) are Amharic, the official language of Ethiopia, the near-extinct Argobba language, and the entirely extinct Gafat. Ethiopic group of languages

The Gurage cluster of languages is found south and east of Addis Ababa and has 550,000 speakers. In all, there are somewhat more than 12,000,000 speakers of Southern Peripheral Semitic languages.

Typical features of the group include traces of two types of verbal forms with prefixed actor marker (one type with full vowels and the other with reduced vowel schemes). In other respects the verbal system is as in South and North Central Semitic; considerable innovations, however, have developed at the New Stage, especially in the Ethiopic group (with a Cushitic substratum). Declension was lost from the Middle Stage. Phonetic development is as in Arabic, but more of the ancient consonants were lost. The Ethiopic group has lost most of the pharyngeal and laryngeal consonants.

Historical and cultural background. Glottochronological methods, which attempt to measure degrees of differences between related languages by comparing a list of basic vocabulary items, indicate that the first group to separate from the Common Semitic ancestral language was Akkadian (Northern Peripheral group, *c.* 3300 BC) and the second was the Southern Peripheral group (second half of 3rd millennium BC). The Northern Central group had contacts for a long time with the Southern Central languages, and linguistic division within the North Central group is dated at the beginning or middle of the 2nd millennium BC. The relative position of Arabic to the other Semitic languages is not quite clear, probably because of its uninterrupted contacts with Aramaic and other nomadic Semitic groups for many centuries.

The oldest of the attested Semitic languages, Akkadian, was the vehicle of a great ancient literature written in a logosyllabic cuneiform writing system of Sumerian origin. Records of other ancient Semitic languages exist in various forms. Amorite, another ancient Semitic language, is known from glosses and proper names; Ugaritic was written in a quasi-alphabetic cuneiform script unconnected with the Akkadian. The Canaanites of Phoenicia used a still undeciphered syllabic script, the Proto-Byblian, in the 2nd millennium BC, while those of Palestine and the Sinai Peninsula employed another undeciphered writing, the Sinaitic script, which may be alphabetic in nature. All the other Semites used and, for the most part, still use consonantal quasi-alphabets with no means or only imperfect means to distinguish the vowels. All such alphabets—of which the more important are the Hebrew, the Syriac, and the Arabic—are descended from the Phoenician linear quasi-alphabet of 22 signs,

Semitic writing forms

Table 2: The Arabic Alphabet and Numerals

consonants					equivalents		approximate pronunciation, classical Arabic
alone	initial	medial	final	name	EB preferred	alternatives	
ا	ا	ـا	ـا	alif	*		*
ب	بـ	ـبـ	ـب	bā'	b		*b*aby
ت	تـ	ـتـ	ـت	tā'	t		*t*ie†
ث	ثـ	ـثـ	ـث	thā'	th	th	*th*in
ج	جـ	ـجـ	ـج	jīm	j	dj	*j*ob
ح	حـ	ـحـ	ـح	ḥā'	h	ḥ	‡
خ	خـ	ـخـ	ـخ	khā'	kh	kh	Ger. Bu*ch*§
د	د	ـد	ـد	dāl	d	dh	*d*id†
ذ	ذ	ـذ	ـذ	dhāl	dh		*th*en
ر	ر	ـر	ـر	rā'	r		*r*ror (trilled)
ز	ز	ـز	ـز	zā'	z		*z*one
س	سـ	ـسـ	ـس	sīn	s		*s*and
ش	شـ	ـشـ	ـش	shīn	sh	sh	*sh*y
ص	صـ	ـصـ	ـص	ṣād	ṣ	ṣ	‖
ض	ضـ	ـضـ	ـض	ḍād	ḍ	ḍ	
ط	طـ	ـطـ	ـط	ṭā'	ṭ	ṭ	
ظ	ظـ	ـظـ	ـظ	ẓā'	ẓ	ẓ	
ع	عـ	ـعـ	ـع	ʿayn	ʿ		¶
غ	غـ	ـغـ	ـغ	ghayn	gh	gh	Fr. *r*ien
ف	فـ	ـفـ	ـف	fā'	f		*f*if*t*y
ق	قـ	ـقـ	ـق	qāf	q	ḳ	♀
ك	كـ	ـكـ	ـك	kāf	k	k	*k*in
ل	لـ	ـلـ	ـل	lām	l		*l*ily†
م	مـ	ـمـ	ـم	mīm	m		*m*aim
ن	نـ	ـنـ	ـن	nūn	n		*n*o†
ه	هـ	ـهـ	ـه	hā'	h		*h*at
و	و	ـو	ـو	wāw	w		*w*atch♁
ي	يـ	ـيـ	ـي	yā'	y		*y*et □
ء				hamzah	initial, omit; medial and final,'		*

vowels, diphthongs, and special diacritical marks		equivalents		approximate pronunciation, classical Arabic
letter	name	EB preferred	alternatives	
ـَ	fatḥah	a	e	*a*t
ـُ	ḍammah	u		f*oo*t
ـِ	kasrah	i		*i*f
ـَا	long fatḥah (alif)	ā		*a*dd, f*a*ther◇
ـُو	long ḍammah (wāw)	ū		f*oo*d
ـِي	long kasrah (yā')	ī		*ee*ve
ـَوْ	fatḥah wāw sukūn	aw		*ou*t
ـَيْ	fatḥah yā' sukūn	ay	ai, ei	*i*ce
ـَى	alif maqṣūrah	ā	á	*a*dd, f*a*ther◇
ـَة	tā' marbūṭah	-ah or -at	-a or -at	▲
ٱ	hamzat al-waṣl	restore a	'	+
ـَآ	alif maddah	ā	a	*a*dd, f*a*ther◇
ـِيّ	yā' shaddah followed by short vowel	īy-	iyy-	*ee*ve⊕
ـُوّ	wāw shaddah followed by short vowel	ūw-	uww-	f*oo*d⊕

numerals

Arabic	westernized Arabic	Arabic	westernized Arabic	Arabic	westernized Arabic
١	1	١٢	12	٢٣	23
٢	2	١٣	13	٢٤	24
٣	3	١٤	14	٢٥	25
٤	4	١٥	15	٢٦	26
٥	5	١٦	16	٢٧	27
٦	6	١٧	17	٢٨	28
٧	7	١٨	18	٢٩	29
٨	8	١٩	19	٣٠	30
٩	9	٢٠	20	١٠٠	100
١٠	10	٢١	21	١٠٠٠	1,000
١١	11	٢٢	22		

**Alif* has no consonantal value of its own; in transliteration, omit at the beginning of a word. *Hamzah* is pronounced as a glottal stop, as in Cockney or New York "bottle." As a vowel, *alif* is pronounced *a*dd. †Pronounced dentally. ‡A pharyngeal fricative. §A velar fricative. ‖It is impossible by English examples to indicate the difference from pronunciation of *ṣ, ḍ, ṭ,* and *ẓ*. The back of the tongue is raised and the pharynx is constricted (velarization), in addition to the regular articulation of these consonants. ¶A contraction of the throat (a pharyngealized vowel that is considered a consonant in Arabic). ♀A uvular stop; a *k* sound produced farther back in the throat than any English *k*. ♁As a consonant, *w*atch; as a vowel, f*oo*d. □As a consonant, *y*et; as a vowel, *ee*ve. ◇ Pronunciation varies from place to place and also depending on the accompanying consonant. ▲Pronounced as *t* in certain grammatical constructions; otherwise pronounced as silent *h*. +Initial vowel elision; that is, in pronunciation the *a* is simply omitted and the preceding vowel is pronounced with the following consonant. ⊕A long vowel sound followed by appropriate consonant sound. See also *footnotes* ♁ *and* □.

Table 3: The Hebrew Alphabet

consonants: printed*	consonants: written*	consonants: name	numerical value†	transliteration: EB preferred	transliteration: alternatives	approximate Israeli Sefardic pronunciation
א	א	alef	1	ʾ; omit at beginning of word		glottal stop or silent
ב	ב	bet	2	b		*b*oy
ו	ו	vet		v	bh	*v*end
ג	ג	gimmel	3	g		*g*irl
ד	ד	dalet	4	d		*d*ove
ה	ה	he	5	h; omit at end of word unless written with dot		*h*ow; silent at end of word
ו	ו	waw (vav)	6	w‡	v	*v*end
ז	ז	zayin	7	z		*z*ebra
ח	ח	ḥet	8	ḥ	h̲, ḥ, ch	Ger. Bu*ch*
ט	ט	ṭet	9	ṭ	t	*t*oy
י	י	yod	10	y§		*y*et
כ (ך)	כ (ך)	kaf	20	k		*k*ey
כ (ך)	כ (ך)	khaf		kh		Ger. Bu*ch*
ל	ל	lamed	30	l		*l*eg
מ (ם)	מ (ם)	mem	40	m		*m*ember
נ (ן)	נ (ן)	nun	50	n		*n*ow
ס	ס	samekh	60	s		*s*o
ע	ע	ʿayin	70	ʿ		glottal stop‖
פ (ף)	פ (ף)	pe	80	p		*p*aper
פ (ף)	פ (ף)	fe		f	ph	*f*an
צ (ץ)	צ (ץ)	tzade	90	tz	z̲, z, ẓ, ṣ, ts	pe*ts*
ק	ק	qof	100	q	k, k̲	*k*ey
ר	ר	resh	200	r		Fr. *r*ien: or trilled *r*
שׁ	שׁ	shin	300	sh	š	*sh*oe
שׂ	שׂ	sin		s	ṣ	*s*o
ת	ת	taw	400	t		*t*oy
ת	ת			t	th	*t*oy (Ashkenazi *s*o)

vowels: sign¶	vowels: name	transliteration: EB preferred	transliteration: alternatives	approximate Israeli Sefardic pronunciation
ַ	pattaḥ	a		
ֲ	ḥaṭaf pattaḥ	a	ă	*f*ather
ָ	qamatz gadol ♀	a	ā	
ֶ	segol	e		
ֱ	ḥaṭaf segol	e	ĕ	p*e*t
ֵ	tzere ḥaser	e	ē	
ֵי	tzere male	e	ê	th*ey*
ְ	shewa naʿ	e	ᵉ(above line) (silent)	

vowels: sign	vowels: name	transliteration: EB preferred	transliteration: alternatives	approximate Israeli Sefardic pronunciation
ִ	ḥiriq ḥaser	i		f*ee*t
ִי	ḥiriq male	i	î	
ֳ	ḥaṭaf qamatz	o	ŏ	
ָ	qamatz qaṭan	o		c*o*rd
ֹ	ḥolam ḥaser	o	ō	
וֹ	ḥolam male	o	ô	
וּ	shuruq	u	û	s*oo*n
ֻ	qubutz	u		

*Final forms in parentheses. †Hebrew has a "ciphered" numeral system in which the letters of the alphabet have numerical value and are used as numbers. For example, numbers 11 through 19 are written with the letter *yod* (10) plus the letter for 1, 2, etc. (except 15 which is written *ṭet* [9] plus *waw* [6], and 16 which is written *ṭet* [9] plus *zayin* [7]); the 20s are combinations of *kaf* (20) plus the letter for 1, 2, etc; 500 is written *taw* (400) plus *qof* (100). ‡Functions as both a consonant and a vowel. See *ḥolam male* and *shuruq* in "vowel signs" table above. §Functions as both a consonant and a vowel. See *tzere male* and *ḥiriq male* in "vowel signs" table above. ‖In formal pronunciation, pronounced as a pharyngeal voiced fricative sound. ¶The long horizontal line represents the consonant; vowel signs are placed above, below, or to the left of it, as shown. ♀In Ashkenazi pronounced as in "ball." *Note:* In transliteration, a letter is usually doubled to indicate a strong *dagesh* (*dagesh* is indicated in Hebrew by a point in the middle of a letter). *Pattaḥ* under final *he, ḥet,* and *ʿayin* is pronounced before the consonant under which it is written.

first attested at Byblos and externally similar to the Proto-Byblian script. (All the European alphabets are descendants of the Phoenician, and all the Asiatic alphabets are descendants of the Aramaic variants of the Phoenician.) From a South Arabic variant of the earliest Semitic alphabet the Ethiopians developed a syllabic writing still in use for the languages of Ethiopia. Maltese uses the Latin alphabet.

Two of the later Semitic languages, Hebrew and Arabic, have been the languages of great religions, Judaism and Islām. The religious significance of Hebrew explains why the language, although already partly replaced by an Aramaic vernacular in the everyday life of Palestine in the late 1st millennium BC and early 1st millennium AD, was still preserved as a literary language by the Jews who were expelled from Palestine by the Hellenistic kings and the Romans between the 3rd century BC and the 2nd century AD. It has been revived in a modernized form as a spoken and written language in Israel. Classical Arabic has been preserved as a literary language since the Arabic conquest of North Africa and the Near and Middle East in the 7th and 8th centuries AD. The language was used for literary purposes by Muslims of different nations all through the Middle Ages and is still used as a language of the school and the administration and as the spoken language of the educated in all Arabic countries, although the vernacular New Stage Arabic idioms are to a great extent mutually unintelligible. There is a great amount of literature—scholarly, religious, scientific, and fiction—both in Hebrew and in Arabic. Of the other Semitic languages, Syriac was and is the language of certain Eastern Christian sects and was the means by which the Greek tradition was passed on to the Arabs. Another Christian sect, that of the Monophysites of Ethiopia, used Ge'ez (Ethiopic) and still retains it in ecclesiastical use, but the literary and other secular remains are less important.

Linguistic characteristics. The Semitic branch of this language family is characterized by several general features.

Phonology. In phonology, the emphatic stop consonants *ṭ* and *q* (from **ḳ*) were retained but not **ṗ*. The affricates of the parent language (which are begun as stops and released as fricative sounds), if they ever existed, were lost or replaced by sibilant and interdental sounds (which scholars symbolize as *s*, *ṣ*, *z*; *th*, *ṯh*, *dh*); the lateral sounds and the interdentals were subsequently lost in most languages. The labialized velar sounds (except in the Ethiopic group) and all postvelar stops were lost. As for the glottal, pharyngeal, and laryngeal consonants, six of them (*g͟h*, *k͟h*, ʿ, *ḥ*, ʾ, *h*) are retained in Arabic and were retained in the other Semitic languages at the Ancient Stage. Hebrew and Aramaic retain ʿ, *ḥ*, ʾ, and *h* (but only *k͟h* and *h* in Modern Hebrew and in most New Aramaic dialects); later Ethiopic and Punic retained only ʾ and *h*, and Akkadian only *k͟h* and ʾ (but *a* became *e* in words that formerly included *g͟h*, ʿ, or *ḥ*). The original six-vowel system changed everywhere as early as the Middle Stage; Arabic preserved it the longest.

Common Semitic sounds

Word formation. Word formation is achieved by an intricate system of vowel infixation, sometimes accompanied by a few suffixes or prefixes. Each pattern of infixation ("scheme"), in combination with a consonantal root, plus the affixes, has its own peculiar type of meanings. The Arabic noun *ma-KTaB-*, for example, means "place of writing, school," and *KaTTāB-* means "writer, scribe"; *KāTiB-*, a participle, means "writer, [the one] writing"; *ya-KTuB-u*, the imperfective form, is "he writes"; *yu-KaTTiB-u*, another imperfective, is "he writes, he teaches to write"; and *KaTaBa*, the perfective, means "he wrote." (The capital letters indicate the sounds of the consonantal root.) The need to correlate these diverse patterns with the basic meaning of the root resulted in the absence of important positional changes in the Semitic consonant sounds as well as in the comparative scarcity of borrowed terms, especially of verbs. The primary nouns, those not derived from verbal forms, are not included in this system of patterns, except, by analogy, in Arabic and the other Southern languages.

Morphology. In regard to morphology, the masculine gender marker is zero (*i.e.*, it has no structural marker), but traces of a *-u* can be detected; the feminine gender marker is *-a* or *-ā* or more usually *-(a-)t-*, although *-t-* belonged originally to another series of gender markers (in which there were more than two genders). The declension of the noun and pronoun was retained in the Ancient Stage of Semitic, with nominative, genitive, accusative, dative-locative, and locative-adverbial cases. The dative-locative ending was lost in Arabic, and the locative-adverbial form appeared only in Akkadian. There are traces of an earlier suffixed definite article, *-m(a)* or *-n(a)*, retained in Arabic as the marker of the indefinite form of the noun. Later new definite articles developed. The dual number was expressed at the Ancient Stage but became lost in the later stages. The plural of the noun is formed in North Semitic by lengthening the singular form. This means of expressing the plural exists also in South Semitic, but here it is to a great extent replaced by the so-called *pluralis fractus* "broken plural" (*e.g.*, Arabic *kalb-* "dog," *kilāb-* "dogs").

Dual and plural forms

The West Semitic languages (except, in part, for Ethiopic and the South Arabic dialects) have lost the old imperfective form with a full vowel scheme and have replaced it by the form next in frequency, namely, the subjunctive mood (ya-qtul-u) with a reduced vowel scheme. The stative verb form, still preserved in Akkadian, developed into a new perfective (Arabic *qatala* "he killed," *mariḍa* "he was ill"), leaving the form **ya-qtul*, which was originally the perfective and jussive (a form expressing a wish or command), for the jussive only.

Syntax. Typical of the Semitic languages are attributive constructions: (1) a construction in which the governing noun appears in a shortened form before the governed noun in the genitive, as well as (2) a construction in which the two nouns, each in its complete case form, are connected by a pronoun (*e.g.*, Old Akkadian *t͟hu*, Aramaic *dh͟ī*).

Vocabulary. As mentioned above, the system of word formation in Semitic does not favour borrowings, especially verbal ones. There are, however, a number of nouns borrowed from Sumerian in Akkadian; from Akkadian, Iranian, and Greek in Aramaic; from Persian and Turkic in Arabic; and from the Agau and other Cushitic languages in the Semitic languages of Ethiopia.

EGYPTIAN

The Egyptian branch of the Hamito-Semitic family includes only one language (with local dialects), namely, Egyptian. It can be differentiated into several stages. The Ancient Stage, Old Egyptian, extended from before 3000 to *c.* 2200 BC; the transition from the Ancient to the Middle Stage, Middle Egyptian, lasted from *c.* 2200 to *c.* 1600 BC and, as a dead literary language, until *c.* 500 BC. The Middle Stage, Late Egyptian (also called Neo-Egyptian), is dated from *c.* 1550 to after *c.* 700 BC and the Demotic language between *c.* 700 BC and some time after AD 400. Finally, the New Stage, called Coptic, began in about the 2nd century AD and lasted at least to the 17th century and possibly, in some villages, until the 19th century. Thus, five literary dialects are differentiated. All these language periods refer to the written language, which often differed greatly from the spoken dialects. Coptic is still in ecclesiastical use (along with Arabic) among the Arabic-speaking Monophysite Christians of Egypt.

Phonology. The Egyptian hieroglyphic writing was not adapted for expressing vowels. By the Coptic period, when an alphabetic writing came into use, the Egyptian vowel system had undergone so radical a change that the original vowels can be reconstructed only very approximately. In the consonantal system the loss of the emphatics (except *p* from **ṗ* and *q* from **ḳ*) is characteristic, as are the changes of **-r* (at end of syllable) to *-ʾ*, **li-* and **lu-* to *i-*, **ki-* and **ku-* to *ṯ* (pronounced as *tch*), **gi-* and **gu-* to *ḏ* (pronounced *dj*). In some cases *ṯ* and *ḏ* apparently reflect the affricates of the parent language. In addition, the original lateral sounds were lost as well as the postvelar stops and labialized velars, and the system of

Characteristic Egyptian sound changes

spirants was simplified. Beginning with Middle Egyptian, *d*, *ḏ*, and *ṯ* developed gradually to *t*, and many final consonants (*e.g.*, *-t*, *-r*) were dropped.

Word formation, morphology, and syntax. Word formation in Egyptian was similar to the Semitic type, although probably less consistent. As for the inflection, there may have been only two cases of the noun in Old Egyptian. The actor case coincided with the genitive, and this may have been responsible for a drastic rearrangement of the entire verbal system. Of the original Hamito-Semitic verbal forms only the stative ("pseudo-participle") is preserved; its subject, when a pronoun, is in the ancient direct (zero) case. Verbal forms expressing action were replaced by attributive and prepositional constructions, with the person of the actor being expressed by a suffixed possessive pronoun or by a noun in the genitive. Stem modifications are less developed in Egyptian than in Semitic; a habitative form with reduplication (repetition) of the third consonant of the root exists along with the normal imperfective of the main stem. In the later periods a new complicated system of secondary verbal tenses developed. Masculine gender was marked by zero (the absence of any ending) or **-aw-*, feminine gender by *-at-*, plural masculine by **-ā-w-*, and plural feminine probably by **-ā-w-āt-*.

Typical of Egyptian syntax are a construction in which two nouns, each in its complete case form, are connected by a pronoun, called the *nota genitivi;* and a *status constructus*, in which the governing noun appears in a shortened form before the governed noun in the genitive.

Writing. The ancient Egyptian writing was a logosyllabic one, having symbols representing either complete words or syllables of words; identical signs were used for syllables with identical consonants but different vowels. According to the external form of the signs, the writing is classified as hieroglyphic when it is found on inscriptions on stone, metal, and other hard surfaces and as hieratic and later demotic when it is used for cursive writing on papyrus manuscripts. Typologically the three forms of writing are identical. Coptic was written in an alphabet based on Greek and partly on Demotic. There is a considerable literature in Egyptian and in Coptic (in the latter, mostly of a religious nature; see also HIEROGLYPHIC WRITING).

Hieroglyphic, hieratic, and demotic writing

BERBER LANGUAGES

The Berber (Berbero-Libyan) branch is represented by a multitude of New Stage Berber dialects distributed all over North Africa, from the Siwa Oasis in the Arab Republic of Egypt to Senegal (about 7,000,000 speakers). The more important dialect clusters are Tamashek (Tuareg), in the central Sahara and south of the Niger; Chaouia and Zouaouah (Kabyle), both in Algeria; Riff and Tamazigt, predominantly in Morocco; Tashelḥayt (Shleuh or Shilha), in Morocco and Mauritania; and Zenaga, in Senegal. Little is known of ancient Libyan, also called Numidian. It is attested by inscriptions found in Tunisia, Algeria, and elsewhere, dating from the times of the Roman Empire and written in a native consonantal quasi-alphabetic script still surviving in a modified form among the Tuaregs of Sahara. Whether the extinct language of the Guanches in the Canary Islands and of the Iberians of Spain belonged to the Berber branch or even to Hamito-Semitic is doubtful.

Phonology. In the phonologies of these languages the vowels **a*, **i*, **u* were lost or reduced to *ə*, and **ā*, **ī*, **ū* became *a*, *i*, *u*; **w* and **y* may appear both as consonants and as vowels, and the emphatics are represented by *ḍ*, *gh* (but in reduplication *ṭṭ*, *qq*), and *ẓ*. The system of spirants has been simplified but retains *š* (*sh*) and *ž* (*zh*) sounds. Interdentals, laterals, and affricates were lost.

Word formation and morphology. Except in the verb, there are only traces of the internal inflection type of word formation characteristic of the Semitic branch. Among grammatical features, a former article no longer retaining its determinative function (masculine **hā-*, plural *hī*, feminine **tā-*, plural **tī-*) is prefixed under certain conditions to the noun, displacing the prefixed markers of gender, *w-* and *t-*. These latter gender markers are at present used in a form of the noun as an attribute or as a subject of a verb when following the predicate in the sentence. The plural of the noun is masculine *-ən* and *-an* and feminine *-in*. A *pluralis fractus* "broken plural" has also developed (mostly an infixation of *-a-*). The perfective of the main verbal stem also has a habitative form (reduplication of the second consonant of the root, or prefixation of *-tt-* to the word base). Tamashek has several verbal tenses.

There is little or no intelligibility between the dialects, except for historically neighbouring ones. A great number of Arabic borrowings are evident in most dialects; there are also a number of borrowings from Punic, Latin, and from the languages south of Sahara.

CUSHITIC LANGUAGES

The Cushitic branch goes back to a reconstructed Common Cushitic parent language; this, according to the Soviet scholar A.B. Dolgopolsky, was the dialect of Common Hamito-Semitic that best preserved the original phonological system. Whether or not West Cushitic (Omotic) is a descendant of Common Cushitic is not clear. The Cushitic languages are all from the New Stage and have more than 13,000,000 speakers.

Languages of the group. The Cushitic languages, including the West Cushitic group, can be subdivided into five groups. The Northern group is represented by Beja, or Bedawiye, spoken mainly in The Sudan close to the Red Sea and also in Eritrea; it has about 30,000 speakers. Typical linguistic features include the scanty representation of affricates, velars changing partly to ʾ, and postvelar consonants changing to *h*. Two or, in some cases, three verbal forms with prefixed actor markers exist ("strong conjugation"), but many verbs are conjugated by suffixes (developed from an auxiliary verb with prefix conjugation). There are stem modifications similar to those of Berber in the strong conjugation, formed by suffixes in the other verbs (this is also typical of the other Cushitic languages). In addition, declension of the noun, with traceable relics of the ancient type similar to the Semitic, also can be seen.

The five Cushitic language groups

The Eastern group has several subgroups. The highland languages, spoken east of Addis Ababa, include Hadya (100,000 speakers), Kambata (300,000 speakers), Sidamo (850,000 speakers), Darasa (250,000 speakers), Burji, and some related languages. The total number of speakers of this subgroup is more than 2,000,000. The other subgroups include Saho-Afar in Eritrea, northeast Ethiopia, and the Territory of Afars and Issas, with 500,000 or more speakers; the Somali subgroup, with Somali, Bayso, Rendile, and other languages in Somalia, eastern Ethiopia, and eastern Kenya, having a total of about 3,000,000 speakers; the Galla subgroup, comprising Oromo (Galla, with several dialects) in western, central, and southern Ethiopia and northern and eastern Kenya; the subgroup of Arbore, Dathanaic (Geleba), and other languages, together having more than 7,000,000 speakers; Konso, Gidole, and related dialects, with about 70,000 speakers in western Ethiopia; Warazi (Warize) and related languages also in western Ethiopia, with about 50,000 speakers; and Mogogodo of northern Kenya. Typical features of the group are the presence of emphatic affricate sounds and the change of postvelar sounds to ʾ and *h;* in some languages the older **l* sound is represented as *j* or *r*, and **r* as *r*, *d*, or *n*. The number of verbs of the "strong conjugation" is very small in some languages and nonexistent in others. In addition, there are grammatical genders differing from the ancient type.

The Central, or Agau, group includes languages or dialects dispersed over Ethiopia. They include Bilin, Khamta, Awngi, and Kemant (Qimant), among others, and are spoken by about 100,000 people. The Quara dialect, spoken formerly by the Falashas, an Ethiopian Jewish ethnic group, is now extinct. Although the vocabulary of all Agau dialects is very similar, there is little mutual intelligibility as a result of the dissimilarity in the phonetic reflexes of the Proto-Cushitic sounds and the strong influence of Ethiopic and Amharic.

The Southern group, located in Tanzania, south of the

Equator, includes Iraqw and its related dialects, Asa, Ngomwia, and others. Characteristic of the group is the loss, for the most part, of emphatic consonants. The laterals, however, are partly preserved, as are the pharyngeal ʿ and a few of the affricates. Both *l and *r are reflected as l- and -r-. In spite of numerous innovations as a result of substratum influence, there are many similarities with Eastern Cushitic in grammar.

The Omotic languages

The Western group, also called the Omotic branch by some scholars, encompasses Ometo, a dialect cluster including the Walamo language, with about 1,000,000 speakers; Janjero (Yamma), Bworo (Shinasha, Gonga), Anfillo (Southern Mao), Benesho-She (Gimira), Ari-Banna, and others, all of which are languages with small numbers of speakers, perhaps about 100,000 in all; and Kafa (Kaficho)-Mocha, with 170,000 speakers. All of these languages are spoken along the western border of Ethiopia and in northern Kenya. Typical features include the change of *s to š, the preservation of most affricates, but the loss of laterals and of all postvelar, pharyngeal, and laryngeal fricatives. The sonants *l, *r, and *n are usually represented alike as n- and -r-. Some languages have tones that serve to differentiate word meaning. Also characteristic are drastic innovations in the pronoun and the verb. Traces of the genders are usually represented as masculine -ō (from *-aw) and feminine -ā and -ē (perhaps from *at).

There have been some attempts to create a written language for Oromo and especially for Somali on the basis of Ethiopic, Latin, or Arabic writing. An original Somali writing system was invented in the beginning of the 20th century, but at present Somali is written in the Arabic alphabet.

Linguistic characteristics. *Phonology.* All the Hamito-Semitic groups of consonants were preserved in Common Cushitic, and separate reflexes of each group can be traced in the different Cushitic languages. Because of an imperfect development of the system of word formation by vowel infixation, however, the stability of the consonantal root was not as necessary for correlation of forms as in Semitic. The reflexes of the sounds of the protolanguage in the individual Cushitic languages therefore depend to a great extent on positional circumstances; thus, a Proto-Cushitic *c, pronounced as *ts*, may have developed into a *d-* in an initial position and an *-s-* in an intervocalic or final position, and so forth. Emphatics are mostly preserved (ḍ, ç or ç, ḳ, etc.); *p is distinguished from *p by different reflexes (Omotic partly retains *p*). Affricates (and also *d*, š, *s*, etc.) represent what in Semitic are interdental consonants.

Morphology. Verbal conjugation by means of prefixed actor markers is preserved only in a part of the verbs or else in traces; in most of the verbs it is replaced by a new system of conjugation (originally a combination of verbal noun plus prefix conjugation of an auxiliary verb). Two genders (masculine *-w, feminine *-t) and traces of noun declension can be observed; partial and sometimes complete reduplication of stems is used as a means of expressing the plural, along with the means known from the other branches. The pronominal system (except in Omotic) is very close to that of Semitic. In vocabulary, there are many borrowings from Ethiopic, Amharic, Arabic, and Nilo-Saharan.

CHADIC LANGUAGES

Hausa

Languages of the group. Of the Chadic (Chado-Hamitic) branch the most important language by far is Hausa. Its approximately 25,000,000 speakers live in northern Nigeria, in the Republic of the Niger, in the northern part of Ghana, in Cameroon, and in parts of Togo, Dahomey, the Chad Republic, and the Central African Republic. Hausa is also spoken as a second language by many speakers of other African languages. All the other languages of the Chadic branch are spoken by smaller ethnic groups; for example, the Bura, with about 685,000 people; the Mandara, with about 382,000; the Angas, with about 280,000; the Bolewa and Karekare, numbering about 220,000; and the Kanakuru, with about 150,000. In regard to the classification of the Chadic languages, some units are at times classified as languages but should preferably be treated as dialects, and vice versa, and the information on many languages (dialects?) is very scanty. Thus, all attempts at classification must be regarded as provisional only. The more important of the approximately 150 "languages" follow.

1. Western group: Hausa, Gwandara, Ngizim, Bedde (Bade), and related languages; Warjawa, Afawa (Pa'a), and related languages; Gezawa, Seiyawa, Barawa of the Dass region; Bolewa, Karekare, Kanakuru, and related languages; Angas, Sura, Ankwe, Gerka, and related languages; Maha (?).
2. Ron group: Fyer, Bokkos, Daffo-Butura, Sha, and Kulere.
3. Kotoko group: Logone, Buduma, Afade, Gulfei, and related languages and dialects.
4. Musgu (Musgum).
5. Masa group: Masa (Banana), Bana, Kulung; Mussoi (?); Marba (?); and Dari (?).
6. Eastern group: Somrai and related languages; Gaberi and related languages and dialects; Sokoro and related languages; Modgel; Tuburi (Kera); Mubi; Dangla-Jegu, Jonkor, and related languages.

The affiliation of the following languages and dialects to the Hamito-Semitic family has been questioned by some scholars:

7. Tera, Jera; Hona, Ga'anda, and related languages; Bata, Gundu, and related languages; Margi, Bura, Chibak, and related languages; Higi (Hiji) and related languages; Laamang (Hidkala); Mandara (Wandala), Glavda, Yawotatakha, Sukur, and related languages.
8. Daba, Hina, Gauar, Musgoi (?); Matakam, Mofu, Gisiga, and related languages.
9. Gidder.

The total number of speakers of Chadic languages is probably about 30,000,000.

It is probable that the linguistic area of the Chadic branch formerly extended farther to the east, thus contacting the Omotic (Western Cushitic) group.

No Chadic language except Hausa has been reduced to writing; for Hausa, Arabic writing began to be used in the 16th century, and now a modification of the Latin alphabet is used.

Linguistic characteristics. In relation to the original phonological system of Proto-Hamito-Semitic, there are many missing sounds in the individual Chadic systems, but a comparison of the different Chadic groups shows that all the distinctive sounds of the parent language are represented in one way or another. Typical of all Chadic languages are tones serving to differentiate meaning in otherwise identical words.

The verb is for the most part a combination of an auxiliary verb (prefix conjugation) followed by a verbal noun (the reverse order is typical for Cushitic); the nominal part often has a vowel suffix pointing out certain qualities of the verb (transitiveness, intransitiveness, and so on). Verbal stem modifications exist, being expressed by such devices as reduplication and suffixes, but in most languages they are not strongly developed. Most of the languages have more or less clearly expressed genders but no declension. Plural is shown as in Cushitic and Berber. Chadic languages have innovations in the pronominal system. In vocabulary there are borrowings from English, Arabic (especially in Hausa), Fulani, and East Sudanic.

BIBLIOGRAPHY

General works: There is still no general survey of the field (including bibliography) to replace I.M. DIAKONOFF, *Semito-hamitskie iazyki* (1965; Eng. trans., *Semito-Hamitic Languages*, 1965), although the developments in the field after 1965 have been considerable (see the *Colloquium on Hamito-Semitic Comparative Linguistics, London, March 1970*). Among other general surveys are G.R. CASTELLINO, *The Akkadian Personal Pronouns and Verbal System in the Light of Semitic and Hamitic* (1962), and T.W. THACKER, *The Relationship of the Semitic and Egyptian Verbal Systems* (1954).

The theoretical problems of the Hamito-Semitic family have, until recently, been studied mostly on the basis of the Semitic branch alone. Especially important in this respect are I.J. GELB, *Sequential Reconstruction of Proto-Akkadian* (1969); JERZY KURYLOWICZ, *L'Apophonie en sémitique* (1961); FRI-

THIOF RUNDGREN, *Intensiv und Aspektkorrelation* (1959). Perhaps the most crucial problem of the Proto-Hamito-Semitic linguistic typology is the reconstruction of the verbal system, for which, besides the above-mentioned works, see also MARCEL COHEN, *Le Système verbal sémitique et l'expression du temps* (1924); O. ROESSLER, "Akkadisches und libysches Verbum," *Orientalia*, vol. 20 (1951): A. KLINGENHEBEN, "Die Präfix- und die Suffixkonjugation des Hamitosemitischen," *Mitteilungen des Instituts für Orientforschung*, vol. 2 (1957), and the critical, although certainly not final review of some later ideas in PELIO FRONZAROLI, "Ricostruzione interna del verbo semitico in alcuni studi recenti," *Accademia Toscana "La Colombaria,"* pp. 71–85 (1972). Another theoretical problem is broached in I.M. DIAKONOFF, "Problems of Root Structure in Proto-Semitic," *Archiv Orientální*, 38:453–480 (1970). The best general review of the Semitic branch of Hamito-Semitic is GOTTHELF BERGSTRAESSER, *Einführung in die semitischen Sprachen* (1928, reprinted 1963). See also GEORGIO LEVI DELLA VIDA (ed.), *Semitic Linguistics: Present and Future* (1961); SABATINO MOSCATI, *An Introduction to the Comparative Grammar of the Semitic Languages* (1964); I.M. DIAKONOFF, *Jazyki drevnej Perednej Azii* (1967), on languages of the Ancient Near East, including Semitic and a general survey of Common Hamito-Semitic. Recent developments are evaluated in *Current Trends in Linguistics*, vol. 6: *Linguistics in South West Asia and North Africa*, with a comprehensive bibliography.

For the individual Semitic languages and the other branches of Hamito-Semitic, the following works may be consulted:

Akkadian: WOLFRAM VON SODEN, *Grundriss der akkadischen Grammatik*, 2nd ed. (1969), and *Akkadisches Handwörterbuch* (1959–); ERICA REINER, *A Linguistic Analysis of Akkadian* (1966); *The Assyrian Dictionary*, published by the Oriental Institute, the University of Chicago (1956–).

Northern Central Semitic: C.H. GORDON, *Ugaritic Textbook*, 2nd ed. (1965); JOSEPH AISTLEITNER, *Wörterbuch der ugaritischen Sprache* (1963); I.J. GELB, "La lingua degli Amoriti," *Rendiconti d. Accademia Nazionale dei Lincei*, 13:143–164 (1958); GIOVANNI GARBINI, *Il Semitico di Nord-Ovest* (1960); Z.S. HARRIS, *Development of the Canaanite Dialects* (1939); JOHANNES FRIEDRICH, *Phönizisch-punische Grammatik* (1951); GEORG BEER and RUDOLF MEYER, *Hebräische Grammatik*, 2 vol. (1952–55); WILHELM GESENIUS and GOTTHELF BERGSTRAESSER, *Hebräische Grammatik*, 29th ed., 2 vol. (1918–29); HANS BAUER and PONTUS LEANDER, *Historische Grammatik der Hebräischen Sprache des Alten Testamentes* (1922); H.B. ROSEN, *A Textbook of Israeli Hebrew*, 2nd ed. (1966); LUDWIG KOEHLER and WALTER BAUMGARTNER, *Lexicon in Veteris Testamenti Libros* (1958); E. BEN YEHUDA, *Thesaurus totius Hebraitatis* (1908–58); FRANZ ROSENTHAL, *Die aramaistische Forschung seit Th. Nöldeke's Veröffentlichungen* (1939) and *A Grammar of Biblical Aramaic* (1961); PONTUS LEANDER, *Laut- und Formenlehre des Ägyptisch-Aramäischen* (1928, reprinted 1966); JEAN CANTINEAU, *Le Nabatéen*, 2 vol. (1930–32); HARRIS BIRKELAND, *The Language of Jesus* (1954); E.Y. KUTSCHER, *Studies in Galilean Aramaic* (1952); THEODOR NOELDEKE, *Kurzgefasste syrische Grammatik* (1880, reprinted 1966); CARL BROCKELMANN, *Syrische Grammatik*, 8th ed. (1960); RUDOLF MACUCH, *Handbook of Classical and Modern Mandaic* (1965); KONSTANTIN CERETELI, "Abriss der vergleichenden Phonetik der modernen assyrischen Dialekte," in FRANZ ALTHEIM, *Geschichte der Hunnen*, vol. 3, pp. 218–266 (1961); IRENE GARBELL, *The Jewish Neo-Aramaic Dialect of Persian Azerbaijan* (1965); CHARLES F. JEAN and JACOB HOFTIJZER, *Dictionnaire des inscriptions sémitiques de l'Ouest* (1960–65); MARCUS JASTROW, *A Dictionary of the Targumim, The Talmud Babli and Jerushalmi, and the Midrashic Literature*, 2 vol. (1950); ROBERT PAYNE SMITH, *Thesaurus Syriacus*, 2 vol. (1868–1901); E.S. DROWER and RUDOLF MACUCH, *A Mandaic Dictionary* (1963).

Southern Central Semitic, or Arabic: CHIAM RABIN, *Ancient West Arabian* (1951); HENRI FLEISCH, *L'Arabe classique* (1956); JEAN CANTINEAU, *Cours de phonétique arabe* (1960) *and La Dialectologie arabe* (1955); A. SUTCLIFFE, *Grammar of the Maltese Language* (1936); E.W. LANE, *An Arabic-English Lexicon*, 8 vol. (1863–93).

Southern Peripheral Semitic: A.F.L. BEESTON, *A Descriptive Grammar of Epigraphic South Arabian* (1962); MARIA HOEFNER, *Altsüdarabische Grammatik* (1943); EWALD WAGNER, *Syntax der Mehri-Sprache* (1953); WOLF LESLAU *Lexique soqoṭri* (1938); M. BITTNER, "Charakteristik der Sprache der Insel Soqotra," *Anzeiger der Wiener Akademie der Wissenschaften, Ph.-hist. Kl.*, vol. 55 (1918)—the same author has published a number of studies on the important languages Mahrī and Shaḥrī ("Shkhauri") in the *Sitzungsberichte* of the *Wiener Akademie* between 1909 and 1917; CHRISTIAN DILLMANN, *Ethiopic Grammar*, 2nd ed. (1907) and *Lexicon linguae Aethiopicae* (1865); WOLF LESLAU, *Étude descriptive et comparative du Gafat* (1956) and *Etymological Dictionary of Harari* (1963); EDWARD ULLENDORFF, *The Semitic Languages of Ethiopia: A Comparative Phonology* (1955).

Egyptian: ELMAR EDEL, *Altägyptische Grammatik* (1955–64); A.H. GARDINER, *Egyptian Grammar*, 3rd ed. (1957); WALTER TILL, *Koptische Grammatik (Saïdischer Dialekt)* (1955); ADOLF ERMAN and HERMANN GRAPOW (eds.), *Wörterbuch der ägyptischen Sprache*, 6 vol. (1926–31, reprinted 1955); WOLJA ERICHSEN, *Demotisches Glossar* (1954).

Berber: ANDRE BASSET, *Handbook of African Languages*, vol. 1, *La Langue berbère* (1952), and *Articles de dialectologie berbère* (1959). An extensive classified bibliography is provided by JOSEPH R. APPLEGATE as a sequel to his article "The Berber Languages" in *Current Trends in Linguistics*, vol. 6 (1970).

Cushitic: M.L. BENDER, "The Languages of Ethiopia: A New Lexicostatistic Classification and Some Problems of Diffusion," *Anthropological Linguistics*, vol. 13 (1971); on the individual branches and languages of Cushitic, see C.R. BELL, *The Somali Language* (1953), a manual of the Isāq dialect; M.M. MORENO, *Il somalo della Somalia* (1955), devoted to the Benadir, Darod, and Digil dialects; ENRICO CERULLI, *Studi etiopici*, 4 vol. (1936–51), contains grammars and vocabularies of Sidamo, Janjero, some Ometo dialects, and Kafa (Kaficho). For a good survey of the individual branches and languages of Cushitic, see F.R. Palmer, "Cushitic," in *Current Trends in Linguistics*, vol. 6, pp. 571–585.

Chadic: On the general problems of the Chadic branch see PAUL NEWMAN and ROXANA MA, "Comparative Chadic: Phonology and Lexicon," *Journal of African Languages*, vol. 5 (1966); D. WESTERMANN and M. BRYAN, "Languages of West Africa," *Handbook of African Languages*, vol. 2 (1952); JOSEPH H. GREENBERG, "Studies in African Linguistic Classification, IV, Hamito-Semitic," *SWest. J. Anthrop.*, vol. 6 (1950). Among the studies of individual Chadic languages those devoted to Hausa are the most numerous of all; the following may be especially useful: C.T. HODGE and IBRAHIM UMARU, *Hausa: Basic Course* (1963); CHARLES H. KRAFT, *A Study of Hausa Syntax*, 3 vol. (1963); R.C. ABRAHAM, *The Language of the Hausa People* (1959). Among the grammatical studies of the other Chadic languages are HERRMANN JUNGRAITHMAYR, *Die Ron-Sprachen* (1970); CARL HOFFMAN, *A Grammar of the Margi Language* (1963); and H.D. FOULKES, *Angass Manual* (1915). Of the many vocabularies of Hausa the most important is G.P.A. BARGERY (comp.), *A Hausa-English Dictionary and English-Hausa Vocabulary* (1934).

(I.M.D.)

Hammurabi

Hammurabi (more correctly Hammurapi) was the sixth ruler of the 1st dynasty of Babylon, which was—as were most other royal houses in Mesopotamia at this time—of nomadic stock. The dynasty belonged to the Amnanum, an Amorite tribe. Hammurabi ruled Babylon from 1792 to 1750 BC, carrying on with the attempts of his predecessors to gain control of the vital waters of the Euphrates River. In so doing he unified Mesopotamia and elevated its northern region to a position of prime historical importance. He was a just ruler who promulgated a famous set of laws for his people. His laws, however, have less historical significance than was once thought.

Like all the kings of his dynasty except his father and grandfather, Hammurabi bore an Amorite name. Only scanty information exists about his immediate family: his father, Sin-muballit; his sister, Iltani; and his firstborn son and successor, Samsuiluna, are known by name.

Early reign

When Hammurabi succeeded Sin-muballit in 1792 BC he was still young, but as was customary in Mesopotamian royal courts of the time he had probably already been entrusted with some official duties in the administration of the realm. In that same year Rim-Sin of Larsa, who ruled over the entire south of Babylonia, conquered Isin, which served as a buffer between Babylon and Larsa. Rim-Sin later became Hammurabi's chief rival.

The reconstruction of Hammurabi's rule is based mainly on his date formulas (years were named for a significant act the king had performed in the previous year or at the beginning of the year thus named). These show him engaged in the traditional activities of an ancient Mesopotamian king: building and restoring temples, city walls, and public buildings, digging canals, dedicating cult objects to the deities in the cities and towns of his realm,



Hammurabi, limestone relief. In the British Museum.
By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

and fighting wars. His official inscriptions commemorating his building activities corroborate this but add no significant historical information.

The size, location, and military strength of the realm left to Hammurabi made it one of the major powers in Babylonia. That Hammurabi was not strong enough to change the balance of power by his own will is well expressed in a diplomatic report: "There is no king who is powerful for himself: with Hammurabi, 'the man of Babylon,' go ten or 15 kings, so with Rim-Sin, 'the man of Larsa'; with Ibalpiel, 'the man of Eshnunna,' . . . go 20 kings."

Hammurabi inherited one major direction for his political activity: to succeed in controlling the Euphrates waters—important in an area that depended exclusively on irrigation agriculture. Such a policy naturally led to conflicts with the kingdom of Larsa, situated in a disadvantageous downstream position. This policy, begun by Hammurabi's great-grandfather but most forcefully and partially successfully pursued by his father, Hammurabi himself took up in 1787 BC, near the beginning of his reign, when he conquered the cities Uruk (Erech) and Isin, held by Rim-Sin, and clashed again with Rim-Sin the year after. But according to Hammurabi's date formulas and contemporary diplomatic correspondence, these operations led no further because Hammurabi shifted the direction of his military operations in 1784 toward the northwest and the east. For almost 20 years thereafter no significant warlike activity is reported. These years were characterized by changing coalitions among the main kingdoms—Mari, Ashur, Eshnunna, Babylon, and Larsa. Hammurabi used this time of uneasy stalemate to fortify several cities on his northern borders (1776–1768 BC).

Warfare against rivals

The last 14 years of Hammurabi's reign were overshadowed by continuous warfare. In 1764 Hammurabi dealt with a coalition of Ashur, Eshnunna, and Elam—the main powers east of the Tigris—whose position threatened to block his access to the metal-producing areas of Iran. It can be assumed, however, that Hammurabi took the initiative in moving against Rim-Sin of Larsa in 1763 BC. Few particulars are reported about this latter war, but it seems that Hammurabi successfully employed a stratagem that apparently had been used before by Sin-muballit against Rim-Sin: damming up the water of a main watercourse and then either releasing it suddenly to create a devastating flood or simply withholding it—the main resource of life—from the enemy's people (that Hammurabi used this device to bring about Rim-Sin's defeat can be assumed from the fact that in 1760 he rebuilt a canal—the western branch of the Euphrates—to facilitate the resettlement of the uprooted population that lived along its course before this war). The final siege of Larsa, Rim-Sin's last stronghold, lasted for several months. It was the last step to Hammurabi's victory.

In 1762 BC Hammurabi again engaged in hostilities with the eastern powers. It is unknown whether this was a protective move on his part or a reaction on theirs to the change in the balance of power. The motives that led Hammurabi in 1761 BC against his longtime ally, Zimrilim, king of Mari, 250 miles upstream from Babylon on the Euphrates remain enigmatic. Two explanations are likely: either it was again a fight over water rights or an attempt by Hammurabi to gain control over Mari's excellent location at the crossroads of the ancient Near East's overland trade.

Two years later Hammurabi had to direct his armies eastward for a third time (1757–1755 BC). The final destruction of Eshnunna during this campaign—again achieved by damming up the waters—most likely proved to be a pyrrhic victory because it removed a buffer zone between Babylonia proper and the peoples of the east (among them probably the Kassites, who were to take over in Babylonia 160 years hence). During his last two years, Hammurabi thus had to concentrate on the construction of defense fortifications. By this time he was a sick man; and he died in 1750 BC, with the burden of government already being carried by his son, Samsuiluna.

Significance of Hammurabi's reign

Changes affecting nearly all spheres of life took place during Hammurabi's reign. They were aimed at the consolidation of conditions resulting from the transformation of a small city-state into a large territorial state. His letters show that he personally engaged in the details of implementing these changes and in the daily routine of the administration of his realm. This personal style is characteristic for Hammurabi and also for other contemporary rulers. Hammurabi's laws—not a law code in the modern sense—must also be considered as an expression of his concern to be a just ruler—an ideal pursued by Mesopotamian kings at all times.

That Hammurabi failed to set up an effective bureaucratic system may be attributed to his personal style in the governance of his realm and the fact that he was fully engaged in wars during the last part of his reign. The lack of effective administration might have been one reason for the fast deterioration after his death of what he had achieved in military terms.

When Hammurabi conquered southern Babylonia he did not follow the century-old tradition of having himself deified during his lifetime. There is reason to believe that this was his personal decision, probably based on a different view of the nature of kingship, setting a precedent for the concept of kingship until Hellenistic times.

Hammurabi's eminence in Mesopotamian history has long been exaggerated. It was first based on the discovery of his laws but subsequent discoveries of older, though less voluminous collections of laws have led to a less enthusiastic view. Moreover, the frequently noted resemblance between Hammurabi's laws and the Mosaic laws is now seen in terms of common heritage rather than as proof for direct dependency.

Hammurabi is also credited with bringing Mesopotamia again under a single rule. Although there existed certain trends toward such unification—particularly expressed in the themes depicted on contemporary seals and in the apodoses of omens evoking a past, when such kings as Sargon of Akkad and Shulgi ruled Mesopotamia from the Persian Gulf to the Mediterranean Sea—it is doubtful that unification was the only motive for Hammurabi's conquests. The lasting achievement of Hammurabi's rule was that the theatre of Mesopotamian history, which had been in the south from the beginning of the 3rd millennium BC, was shifted to the north, where it remained for more than 1,000 years.

BIBLIOGRAPHY. CYRIL J. GADD, "Hammurabi and the End of His Dynasty," *Cambridge Ancient History*, rev. ed., vol. 2, ch. 5 (1965), well-balanced description of Hammurabi's reign and time, good bibliography; JEAN BOTTERO, ELENA CASSIN, and JEAN VERCOUTTER (eds.), *Die Altorientalischen Reiche* (1965; Eng. trans., *The Near East: The Early Civilizations*, 1967), emphasis is more on general trends than on single events, more up to date than previous work; JOHANNES RENGER, "Zur Lokalisierung von Karkar," *Archiv für Orientforschung*, 23:73–78 (1970), discusses the conflict over water rights during the reigns of Hammurabi and his father.

(J.M.R.)

Handball and Fives

Handball and fives are similar games played in walled courts or against a single wall, with a small rubber ball that is struck with hand or fist against the wall. The object is to cause the ball to rebound with variations of power or speed and at such an angle that the opposition cannot return it.

HANDBALL

There are three versions of handball: four-wall, three-wall, and one-wall. Each may be played by two (singles) or four (doubles). (There is an unofficial three-player practice game called "cut-throat" handball.) Although handball fundamentals can be explained and learned quickly, developing the skill, strategy, and stamina required to play well presents a continuing challenge.

Background. One of the oldest games played with a ball, handball has been traced back to the *thermae*, or baths, of Rome. Played later as a bare-handed game called *pelota* in Spain and France, it was the forerunner of modern jai alai (*pelota vasca* or *pelota basque*), a three-wall game in which players hurl, catch, and return a small, hard ball using a wicker scoop strapped to one arm. Handball was adapted in the British Isles during the 16th century and called fives.

Irish origins

The immediate forerunner of the modern game was developed in Ireland, where handball was played about 1,000 years ago. In the 1850s, Irish town and county championships were regularly played, using a hard leather-covered ball on courts that were about 80 feet (24 metres) long and 40 feet (12 metres) wide. Irish emigrants took the game to the United States in the 1880s, and Phil Casey built the first U.S. walled court in Brooklyn in 1886. Kicking was permitted, and some players developed unusual skill in returning low balls with their feet.

The first international match was played in 1887 between the Irish champion, John Lawlor, and the U.S. champion, Casey. Begun in Ireland on August 4, the match ended on Casey's U.S. court on November 29. Casey won, 11 games to six, and then retained his title against all challengers until his retirement in 1900.

During the 1890s play began with a soft ball—usually a tennis ball with the outer covering removed. This ball was used in smaller four-wall courts in the New York City area. Then young players began using this larger ball outdoors against the solid wall of a building. The use of the soft ball spread to other cities, mostly for four-wall play. Dissatisfaction with the large, slow ball led to development of a smaller gas-filled ball that proved to be more acceptable to players and stimulated new interest in the game.

The four-wall soft-ball game was taken up avidly in such midwestern U.S. cities as Detroit, Cleveland, Milwaukee, and Chicago. When the Milwaukee fire department converted from horse-drawn to motorized equipment, its stables were converted into handball courts. The sport helped keep fire fighters fit, and a number of the firemen developed into ranking competitive players.

The one-wall game

At about the same time, the game underwent changes and courts were reduced in size. A one-wall game was developed in New York about 1913 and played on the beaches against bathhouses and bulkheads, with the hard sand forming the floor of the court. Within a few years, the one-wall game was played by both men and women throughout the eastern United States. It was taken indoors by YMCA's and clubs and, where space in large cities was at a premium, was even installed on roofs. As a forerunner of the three-wall game some one-wall courts were equipped on each side of the front wall with a hinged triangular "wing," which swung out and locked into position on the forward portion of the two side lines, forming a partial three-wall court and making difficult shots into the front corner possible.

Irish handball. The Irish game of handball was originally played on a hard clay floor, with one wall of stone at the front of the court, against which the ball was struck. The ball, the four-wall court, and the system of scoring were developed during the 18th century, when the game was a popular pastime in many parts of Ireland. The Irish ball has a cork or wood centre covered by woolen thread and thin pieces of cork or rubber. The outer cover is of thin sheepskin. This ball is called an alley cracker or hard ball. Some tournaments or championships are played with a soft, rubber ball, but the Irish ball is the recognized championship medium. The standard ball is 1⅞ inches (4.7 centimetres) in diameter and weighs 1½ to 1¾ ounces (43–50 grams).

Rules have changed little since the 1880s, when the first Irish championships were played. For many years challenge and championship games for stakes as high as £1,000 were not uncommon. Then the Gaelic Athletic Association took over supervision of the game, betting was abolished in 1924, and Irish handball became an amateur game. Regular county, provincial, and all-Ireland championships were established. There are many courts throughout Ireland and at most schools and colleges, although a soft, rubber ball is now generally used in schools.

Organized competition. National handball championships in the United States have been conducted by the Amateur Athletic Union (AAU), since 1919; the Young Men's Christian Association (YMCA), since 1925; the National Jewish Welfare Board (JWB); and the United States Handball Association (USHA). Collegiate tournaments sponsored by the USHA and junior tournaments involving younger players contribute to the increasing number of handball participants.

During a consultation held in Chicago in 1959 that involved representatives from the AAU, JWB, USHA, and YMCA, an agreement was reached recommending one set of uniform playing rules to be used by all four groups. Three years later it was decided to conduct an All-American Handball Championship on a quadrennial basis, during each Olympic year. Sponsored by three of the four national organizations the first such event was conducted successfully in St. Louis in the spring of 1964. Unfortunately, no subsequent similar events were conducted.

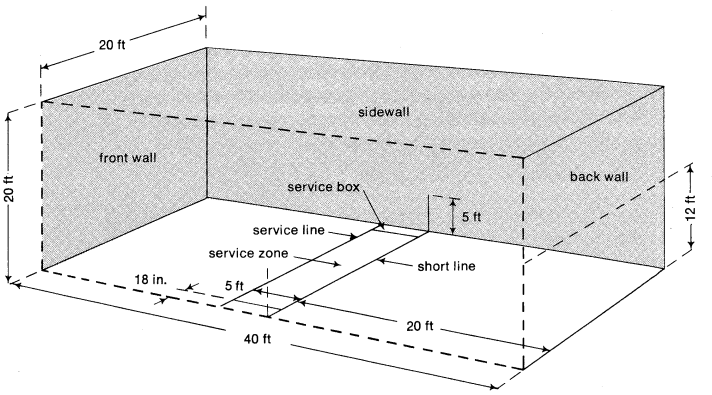
Handball is now also played in Canada and Ireland and to some extent in Mexico, Argentina, Australia, New Zealand, and France. International competition has been stimulated in recent years through invitational events conducted by Canada and the United States. One, announced as a world championship, held during the 1964 New York World's Fair, brought together players from Ireland, Mexico, Australia, Canada, and the United States. The U.S. team won both the singles and doubles events over Canadian players. Another international event held in the fall of 1967, in Toronto, involved players from Ireland, Australia, Canada, and the United States. A third international event planned by Ireland for 1970 was postponed because "standard" courts could not be provided and some top players were not able to participate.

International status

The desire for international recognition of handball continues to build. Efforts have been made to standardize playing facilities, equipment, and rules, and to bring about a recognized international organization to plan and govern handball competition. A proposal was made to have handball included as a demonstration sport during the 1976 Olympic Games.

The court. Standard four-wall courts are 40 feet (12 metres) long, 20 feet wide, and 20 feet high, with a back wall 12 feet high. A short line, parallel to the front wall, divides the court in half; the service line is parallel to and five feet in front of the short line. Between these two lines on each side of the court is a service box, formed by a line parallel to and 18 inches (46 centimetres) from each sidewall. The serving zone is the space between the outer edges of the short and service lines. A vertical line marked on each sidewall, five feet behind the short line, indicates the front edge of the back court receiving zone.

In one-wall handball the wall (front) is 20 feet wide and 16 feet high, with a playing zone 34 feet long (the back edge is the long line) by 20 feet wide. The sidelines extend three feet farther from the wall than the long line. The



Four-wall court according to handball unified playing rules.

short line runs parallel to and 16 feet from the wall. Service markers (lines) six inches long, parallel to and midway between the long and short lines, extend inward from the sidelines. The imaginary joining of these markers forms the service line. The service area, 20 feet by nine feet, includes the short lines, service lines, and sidelines. The 20 feet by 18 feet receiving zone behind the short line includes the long and sidelines. One-wall courts are popular because they permit more spectators to watch games and are less expensive to build.

Three-wall courts also permit more spectators to view matches but are less common and are not standardized. They may have a front and two sidewalls, the back being open, or a front wall, back wall, and one sidewall as in a jai alai court. Court dimensions and markings are similar to four-wall courts, with a back court (long) boundary line or a sideline added for the jai alai type court.

Development of courts with one or more glass walls allowed more spectators to watch four-wall matches and enjoy skillful play off the back wall.

Principles of play. The ball is made of black rubber; it is 1⅞ inches in diameter and 2.3 ounces in weight. Only one hand may be used in striking the ball; no other part of the body may be used. When attempting to return the ball, the player cannot strike it more than once. Gloves made of soft material or leather must be worn to prevent moisture from affecting the ball.

Serving the ball

To start the game a server stands anywhere within the serving zone, drops the ball to the floor, and strikes it on the first bounce with one hand, causing it to hit the front wall on the fly. In the four-wall game, the rebounding ball must land on the floor back of the short line, either before or after striking one of the sidewalls. If it does not cross this line, it is a short ball, which is a fault. Two successive faults retire the side. In the one-wall game, if the ball lands beyond the long line, it is a long ball, also a fault; if it goes outside the sidelines, it is a handout—that is, the side (hand) serving loses service but the score remains the same. A fault is called if, while serving, the server steps beyond the serving zone or leaves the zone before the rebound passes the short line. In four-wall doubles, the server's partner must remain within the service box until the rebound passes the short line; in one-wall doubles, he must be outside the sidelines straddling the service line. Breaches of these regulations are faults. If a server's partner enters the playing zone before the served ball passes him, it is a fault.

The return

The receiving side may return the serve (rebound) either on the volley (fly) or on the first bounce; the return may hit the sidewalls and must hit the front wall. In four-wall handball the receiver or receivers must be at least five feet back of the short line until the server strikes the ball; in the one-wall game the receiver must remain back of the service line until the rebound passes the short line. The rally continues, sides alternating in hitting the ball until one side misses. If a receiver misses, the server scores a point; if the serving side misses, the receiving side wins only the right to serve. The server continues to serve as long as he scores. Game is 21 points.

In doubles, one player on the starting side serves. If he faults twice, service passes to the other team. Thereafter, both players on each team must serve before the serve goes to the opposition.

Players cannot block each other from playing the ball. If they do, or if the ball hits an opponent before hitting the front wall, the ball is dead and must be served again.

Three-wall regulations are the same as those for four-wall, except that a ball in play striking outside the long line (a long ball) is a point if struck last by the receiver or handout if struck last by the server.

Any shot beyond an opponent's reach is called a placement. The most effective placement is a kill, in which the ball rebounds at a height so low that it is impossible to return it. An ace is a legal serve that eludes the receiver. The greater variety of angle shots—for example, sidewall to front wall, ceiling to front, side to back—makes the four-wall game the most demanding form of handball. All versions require good physical condition, speed, control, and stamina, and many athletes play handball to condition themselves for other sports.

Strokes

The three strokes used in handball are the underhand, the basic service stroke; the overhand for high bounding balls or fly balls; and the sidearm, the best stroke for the kill shot. It is important to move into position in the court to be ready to make the return stroke. A player who can strike the ball with either hand can save energy and be a double threat to his opponent. (H.T.F.)

FIVES

The predominantly British games of fives had their origin in the casual play of boys hitting a ball against any convenient wall. They differ from handball games insofar as they are played with a small, hard, composition ball (leather-covered until recently) and because in their present sophisticated form they are a product of the British public schools, as reformed in the 19th century for the sons of the aristocracy and the well-to-do. The word fives derives from the same source as "a bunch of fives" meaning the closed fist or the five fingers held together.

Eton fives. The Eton fives court is a remarkably close copy of the court used by generations of Eton College boys on the steps of the school chapel. It is enclosed on three sides and open at the back. A shallow step divides the court into an upper and lower part. On the left-hand sidewall, by the step, there is a small buttress, known as the "pepper-box." With the step it makes a small, square cavity on the lower floor, called the "Dead Man's Hole." The upper court has an angled ledge, the lower part of which (four feet six inches from the ground) runs across the front wall and is called the "line." There is a vertical line marked on the front wall, three feet eight inches from the right-hand sidewall, that has to do with returning service. There is a second and straight ledge below the top ledge in the upper court, which is an additional hazard but has no further significance.

The game is played between four players, two on each side, who wear leather gloves to protect their hands. There is no recognized singles game. When the game starts, the server alone stands in the upper court. To begin the game he throws the ball so that it hits the front wall above the line, then the right-hand side wall, and falls to the lower court.

The "first cut"

The opponent who returns the service is said to make the "first cut," and he need not do so until he gets a service to his liking. Serving is only a method of putting the ball in play. The ball must be hit with a single blow of the hand or wrist and must not touch any other part of the striker's person.

After the first cut the ball is played alternately by a player of each side, providing the ball is played not later than after the first bound on the floor and is returned above the line, the rally continues.

A game is won by the side that first obtains 12 points. When the server's side wins a rally, a point is won; if the server's side loses a rally, the server's partner serves and no point is scored; if the server's side loses another rally, the service changes to the opponents. A point can only be

won if the winners of the rally are the serving side. A match is usually the best of five games.

Governed by the Eton Fives Association, the main competitions are for an amateur championship and a Public Schools competition.

For the uninitiated spectator the game is bewildering. The ball rebounds in many unexpected ways off the hazards, and the players seem often to be in each other's way. It involves mental as well as physical exercise, there being a variety of ways to outwit an opponent.

Rugby fives. The Rugby fives court has four plain composition walls and a hard composition floor. The front wall, which is 15 feet high, has a board running across it, the top being two feet six inches from the floor. The height of the sidewalls is 15 feet for the first 12 feet from the front wall, thence sloping down to six feet, the height of the back wall. Courts are enclosed, the spectators at the back wall. All courts built since 1930 have been of standard dimensions, but older courts are still in existence.

The game is played like Eton fives, except that the player who makes the first return is called the server. He can either throw the ball up for himself, and usually does, or he can require his opponent, called the receiver, to do so. The ball must be hit only with the hand or forearm. The receiver is "up," and only a side that is "up" can score points. Matches are usually the best of three games, the first team or player to get 15 points winning the game. Club and school games are won on total points. Both singles and doubles are played.

Competition, regulated by the Rugby Fives Association, includes several regional championships, a universities championship, amateur singles and doubles championships, and schools championships.

The emphasis in Rugby fives is physical. The games can be exhausting because of the speed of the ball, the low board, and the use made of the back wall. It is the most popular of the three games.

Winchester fives. Winchester fives is a game confined to a few schools, there being no association or championships and few courts. The court is similar to the Rugby one, but a change of direction of the left-hand wall makes the court slightly narrower at the back than at the front. This changes the positioning of players and Rugby fives players are at a disadvantage. Despite these differences, Winchester players have little difficulty in adapting to the Rugby game, and several schools that play the Winchester game are affiliated with the Rugby Fives Association. (J.Ar.)

BIBLIOGRAPHY. *Ace* (bimonthly), a United States Handball Association publication containing news, articles, announcements, and results of major regional and national handball and racquetball tournaments in the U.S. and Canada; H.T. FRIERMOOD (ed.), *Handball: Official, Unified Playing Rules* (1968), playing rules as adopted and recommended by three U.S. national organizations, codified for easy reference, dealing with the four-wall, one-wall, and three-wall handball games; *Handball Court Specifications: One-wall, Three-wall, and Four-wall Courts* (1968), a USHA booklet giving practical suggestions and guidelines on planning and constructing conventional and glass courts; C.L. MAND, *Handball Fundamentals* (1968), a basic work containing information on skills and techniques, basic strategy, testing techniques, equipment, safety practices, and a glossary; C.J. O'CONNELL, *Handball Illustrated* (1964), the fundamentals of footwork, serving, receiving, angle and kill shots, game technique, and singles and doubles play, and *Suggestions and Ideas on Handball Court Construction* (1969), a discussion of standard dimensions, materials, lighting, ventilation, spectator galleries, glass walls, special problems, and costs of construction; B.E. PHILLIPS, *Handball: Its Play and Management* (1957), the fundamentals of one-wall and four-wall handball skills, including use of the hand, proper stance, body control, footwork, and use of underhand, overhand and sidearm strokes; well illustrated.

(H.T.F.)

Handel, George Frideric

One of the greatest composers of the late Baroque era, George Frideric Handel, German by birth and an Englishman by adoption, mastered the techniques of German, French, and Italian musical styles and adapted them to a new environment to create works of such power that he became a familiar institution in Britain and a key figure in the national tradition of Germany. Ultimately his work became an essential factor in the popularization of European music throughout the world.

By courtesy of the National Portrait Gallery, London



Handel, oil painting after Thomas Hudson, 1756. In the National Portrait Gallery, London.

Handel was born in the city of Halle-an-der-Saale in Saxony on February 23, 1685. His father, Georg Händel (the usual German form of the name), was a successful surgeon; his mother was the daughter of a Lutheran clergyman.

Education in Halle. Handel showed a marked gift for music. Since at that time music was eagerly cultivated in all sections of German society and was considered a useful social accomplishment, his talent was encouraged by his family. A romantic legend stating that Handel's father disliked music is without foundation; actually, he was a friend of several well-known musicians. Handel had the opportunity to hear excellent church and civic music in Halle; also at nearby Weissenfels, where his father was court surgeon to the duke of Saxe-Weissenfels, he was brought into contact with the music played at court. He became a pupil in Halle of the composer F.W. Zachow, organist of the Liebfrauenkirche and director of the town choir, and from him he learned the principles of keyboard performance and of composition; others instructed him in oboe and violin. Oboe sonatas composed by him at that time still survive. In 1696 he was taken to Berlin, where his talents were noticed by the elector Frederick III (later King Frederick I of Prussia) and the electress Sophia Charlotte.

Musical training

His father died when Handel was 11, but his general education had been provided for, and on February 10, 1702, he enrolled as a law student at Halle University. Directed by forward-looking professors, this institution rapidly became a strong influence in the liberal, humanitarian, scientific movement of the 18th century known as the *Aufklärung* (Enlightenment). The stimulating character of this intellectual community played a notable role in Handel's later development.

Between 1700 and 1702 Barthold Heinrich Brockes, who was to become an important figure in the literary and musical life of Hamburg, was a student at Halle, where Handel probably attended concerts organized by him. Later in life Handel set Brockes' poems, as well as a Passion text of his, to music. The composer Georg Philipp Telemann passed through Halle on his way to

Leipzig in 1701, and he and Handel became close friends.

In order to enlarge his musical experience and to guarantee some part of his living expenses, Handel became organist of the Reformed (Calvinist) Cathedral in Halle when he entered the university. The appointment was for one year, and at the end of that time he decided to go north to Hamburg, where greater opportunities awaited him.

Young composer in Hamburg and Italy. In Hamburg, Handel made friends with Johann Mattheson, musician and man of letters, who became secretary to the British resident there. Before settling down in Hamburg, however, Handel travelled with Mattheson to Lübeck in August 1703 to inquire about the chances for either to succeed the famous Lübeck organist Dietrich Buxtehude. Unsuccessful in their attempts, the two young men returned to Hamburg where Handel joined the violin section of the opera orchestra. He also undertook to instruct the children of the British resident in music.

Encouraged by the director of the Hamburg opera, Handel took over some of the duties of harpsichordist, and at the beginning of 1705 he presided over the premiere of his first opera, *Almira*. This was sufficiently successful for him to be invited to contribute further works to the repertoire. The music for these works—*Nero* (1705) and *Florindo* and *Daphne* (1706)—is, however, lost.

First opera

At this time Italy was considered to be the musical centre of Europe, and ambitious musicians who could raise the necessary money went there to study. Handel, who had already made the acquaintance of several eminent Italians, was unusual in that not only did he manage to support himself for three years in Italy but was able to leave that country richer both in funds and reputation. His "grand tour" was a great success, indicating command both of musical expression and social graces. He resided for periods of varying length in Florence, Rome, Naples, and Venice. In Florence, he met the composer Alessandro Scarlatti, and it was in this city, in 1707, that his first Italian opera, *Rodrigo*, was written. The production of *Rodrigo* resulted in enthusiastic acclaim for Handel from the principal patrons of art in Rome. Handel moved to Rome, where, responsive to environment, stimulated by the competition afforded by the presence of such masters as the violin composer and player Arcangelo Corelli, and eager to realize his ambitions, he composed prolifically. His works included many Italian solo cantatas (vocal compositions) and duets, the extended cantata *Il trionfo del tempo e del disinganno* (1708), Latin church music, and the oratorio *La Resurrezione*.

From Rome, threatened by hostilities during the War of the Spanish Succession, Handel moved on to Naples. Here he composed the serenata—a dramatic cantata—*Aci, Galatea e Polifemo* in June 1708, an interesting set of French songs, a *Cantata Spagnuola* (*No se enmenderá jamás*) that apparently stemmed from a flirtation with a "Donna Laura" from Spain or Portugal, and began to compose the opera *Agrippina*. The librettist was Cardinal Vincenzo Grimani, the Venetian viceroy of Naples. *Agrippina*, first performed in Venice on December 26, 1709, enjoyed sensational success, the performance being frequently interrupted by cries of "Viva il caro Sassone!" ("Long live our beloved Saxon!").

Before his arrival in Italy Handel enjoyed a national reputation; his mastery of the Italian opera style made him an international figure. Through his friend Agostino Steffani—a papal ambassador at large and, until lately, musical director in Hanover—he met Prince Ernest Augustus of Hanover, who offered him Steffani's place at court. Handel accepted and remained director of music at Hanover officially until 1716, but mostly he held the post *in absentia*. He arrived in Hanover at the beginning of 1710 and left in the autumn for London. Since the aged electress Sophia of Hanover was heir-presumptive to the English throne, a special relationship already existed between the two countries, so that Handel's departure for England did not seriously imperil his position.

Musical director at Hanover court

Successful career in England. Early in 1711 his sparkling opera *Rinaldo* was performed in London and aroused the normally phlegmatic English to such enthusiasm that Handel sensed the possibility of continuing popularity and prosperity. A year later, with electoral permission for a limited stay, he was back in London for the production of *Il pastor fido* and *Teseo*. In 1713 he won his way into royal favour by his *Ode for the Queen's Birthday* and the *Utrecht Te Deum and Jubilate* in celebration of the Peace of Utrecht that ended the War of the Spanish Succession, and he was granted an annual allowance of £200 by Queen Anne.

Recognized and respected as man and musician by prominent members of both the English aristocracy and the intelligentsia, Handel was in no hurry to return to Hanover. Since, on the death of the Queen in 1714, the elector George Louis—whose mother Sophia had died in the previous June—became King George I of England, Handel had no need to do so. On October 26 the King and his family heard music by Handel in the Chapel Royal, one fact (among others) that destroys the legend that the King was annoyed by Handel's lengthy absences from Hanover and became reconciled to him only in 1717 by his *Water Music*. By that time Handel had again visited Hanover in the retinue of the king-elector. While in Germany he also went to Hamburg, where *Rinaldo* had been performed in 1715 and where his setting of Brockes' Passion (*c.* 1716) was sung several times from 1717 on; to Dresden, where he spent three months and engaged some singers for the London opera season; to Halle to see his family; and to Ansbach, once the home of his friend Caroline, princess of Wales, where he persuaded a local tradesman and old acquaintance, Johann Christoph Schmidt (who in England changed his name to John Christopher Smith), to become his secretary. Two years later Handel was again in Germany, primarily to gather singers for the Royal Academy of Music, newly established for the purpose of producing operas, but also to see his family and friends again. A German cantata attributed to Handel, dated 1719, may belong to this visit; in any case it serves as a reminder that despite increasing commitment to English affairs Handel still felt himself to be German.

In 1718 Handel became director of music to the Duke of Chandos, whose palace at Cannons lay a little north of London. There Handel composed 12 *Chandos Anthems*, a setting of the poet John Gay's *Acis and Galatea* (a different work from *Aci, Galatea, e Polifemo*), and *Haman and Mordecai*. This last—the text by several hands (probably including Alexander Pope, the English satirical poet) and after the French playwright Jean Racine's *Esther*—was to be the effective starting point for English oratorio—a large musical composition for solo voices, chorus, and orchestra, without acting or scenery, usually dramatizing a story from the Bible.

Secure in his profession Handel now bought the house in Brook Street, which was to be his home for the rest of his life. In 1726 he became a British subject, which enabled him to be appointed a composer of the Chapel Royal. In this capacity he wrote much music, including the *Coronation Anthems for George II* in 1727, and the *Funeral Anthem for Queen Caroline* ten years later.

British citizenship

From 1720 until 1728 the operas at the King's Theatre were staged by the Royal Academy of Music, and Handel, as a director of the academy, was responsible for composing the music of most of them. Although the songs from them were very popular, the operas themselves enjoyed varying degrees of success. From 1728, after the sensation caused by the satirical *Beggar's Opera*, the future of opera—which meant opera in the Italian style—appeared increasingly uncertain. Handel, however, was loath to relinquish interest in what he took to be not only the highest form of composition but also the most lucrative. In 1729 he once more visited Europe in search of singers, and in order to visit his mother whom, now in failing health, he was not to see again. During this visit, in June, Johann Sebastian Bach, sending his son Wilhelm Friedemann as emissary, attempted to meet a master whom he admired and from whose works he had made copies. Handel and Bach, however, were never able to meet.

In spite of the vagaries of some of his Italian star singers and the uncertainties of public taste, Handel went on composing operas until 1741, by which time he had written about 40 such works. Opera went into decline in London for a variety of reasons, one of them being the impatience of the English with a form of entertainment in an unintelligible language sung by artists of whose morals they disapproved, and as it did oratorio became increasingly popular. Oratorio, comparable in structure to opera, was already firmly established in southern Europe. Its adoption by the English was more or less fortuitous. It was primarily due to revivals in 1732 of *Acis and Galatea* and *Haman and Mordecai* (renamed *Esther*), but also to a general love of choral music, to the importunities of those who wished Handel to use English lyrics, to the interest of the royal family, and to the moral value set on it by the middle class. General approval was generated by the use of biblical words and by the fact that Handel's music afforded a combination of pleasure and piety. Handel, however, did not apply the form exclusively to biblical subjects, although these preponderated. He and his collaborators chose excellent plots and with the extra dimension allowed by considerable chorus participation he made it a fine medium for dramatic expression.

Growth of oratorio in England

As musical director, impresario, composer, virtuoso performer on harpsichord and organ, and fashionable teacher, Handel expended his energy recklessly and in 1737, having suffered what appears to have been a mild stroke, was very ill. After recuperating at Aix-la-Chapelle, he was saddened by the death of the Queen. The sombre anthem, incorporating a German chorale melody, composed for her funeral was publicly performed on December 17 and privately for the royal family in the Chapel Royal a week later. Soon Handel was as active as ever. He produced one of his most charming and humourful operas, *Serse*, published the Opus 4 organ concerti, composed *Saul* and *Israel in Egypt* as well as shorter works, and helped establish the Fund for the Support of Decayed Musicians (now the Royal Society of Musicians). In May 1738, the general public was able for the first time to see the statue of Handel by the French sculptor Louis-François Roubillac, erected in the composer's honour in Vauxhall Gardens where his music was frequently sung and played.

The extent to which Handel had steered away from Italian opera is indicated by his use of classical English texts in 1739 and 1740. These were the English poet John Dryden's "Ode for St. Cecilia's Day" and John Milton's "L'Allegro" and "Il Penseroso." In 1740 Handel published the concerti grossi of Opus 6 and a second set of organ concerti. But during that year, in response to the decline of opera, Handel may seriously have considered a change of course but appears to have been discouraged. In September a Hamburg newspaper reported that Handel had lately played the organ in the Dutch town of Haarlem and that he was en route to Berlin. If this account was accurate it would indicate the possibility of overtures having been made to him by Frederick the Great of Prussia, who was then reorganizing his own musical establishment. Whatever the veracity of these reports, Handel returned to England where he held a "farewell concert" on April 8, 1741.

At this juncture an invitation from the duke of Devonshire, the lord lieutenant of Ireland, to furnish a work for Dublin charities pulled Handel out of his state of dejection. Characteristically, he now moved to the opposite extreme and worked so intensively that his promised oratorio, *Messiah*, was written in its first form between August 22 and September 14. A month later its scarcely less inspired successor *Samson* was also ready for the copyists. In November, accompanied by some of his performers, Handel travelled to Dublin by way of Chester, where, with the aid of the cathedral musicians, he tried out *Messiah* in private. On December 23 he began a series of subscription concerts in Dublin and on April 13, 1742, *Messiah* was given its first performance. It created a deep impression and was repeated on June 3. Handel remained in Ireland through the summer, during which time he rearranged the suite of pieces comprised in *Forest Music*.

Messiah

The works of the next three years included the oratorios *Joseph and his Brethren* (first performed in 1744) and *Belshazzar* (1745), the secular oratorios *Semele* (1743) and *Hercules* (1745), and the *Dettingen Te Deum* (1743) celebrating the English victory over the French at the Battle of Dettingen during the War of the Austrian Succession. In 1745, however, there was further concern about Handel's health, and a contemporary wrote that he appeared "dejected, wan, and dark [as he sat by] not playing the harpsichord" at a performance of *Alexander's Feast*. Yet the decade that followed *Messiah* was one of great significance in Handel's career. During this time he made oratorio and large-scale choral works the most popular musical form in England, a fact of some social importance. In the mid-1740s many voices were being raised against "permissiveness," and the feeling that the great oratorios were some sort of bulwark against moral laxity grew stronger. This feeling of righteousness grew when, after 1749, *Messiah* became especially associated with the Foundling Hospital. A general benefactor of the hospital, and a governor since 1750, Handel superintended *Messiah* performances in its Chapel until 1754. In his testament Handel gave a score and a set of the parts of *Messiah* to the institution.

Handel's last years. Handel's sense of charity and his general concern for other people raised him to a high place in the affections of the British. His music was even then recognized as a reflection of the national personality, and his capacity for realizing the common mood was nowhere better shown than in the *Music for the Royal Fireworks* with which he symbolized the end of the Silesian Wars, a part of the War of the Austrian Succession, in 1749. Twelve thousand people are said to have attended a rehearsal of this music in Vauxhall Gardens.

Music for the Royal Fireworks

In the summer of 1750 Handel, having visited Germany for the last time, was seriously injured in a coach accident in Holland. Recovered from the mishap, he returned home to compose *Jephtha*, the last of his oratorios. While working on this—as he noted on the manuscript on February 13, 1751—his eyes began to trouble him, and he underwent eye surgery. Although he never became completely blind, there is no doubt that he was gravely handicapped in his work in his final years.

Jephtha was performed at Covent Garden Theatre (where the oratorios were usually performed), London, on February 26, 1752. From then on the Lenten oratorio seasons were eagerly anticipated. Handel attended the performances, and he sometimes played concerti on the organ. He had indeed become an institution to the English people. His friendships, however, extended far beyond his own profession and England's borders.

Assisted by his friend and secretary John Christopher Smith, Handel kept his interests in musical activities alive until the end, and on March 30, 1759, he was present at the Foundling Hospital *Messiah* performance. He was not well enough to attend a second performance a week later. On April 14 he died, and on April 20 he was buried in Poets' Corner in Westminster Abbey in the presence of a congregation of 3,000 people who had come to mourn the passing not only of a great composer but of one held in universal esteem as a man.

Death

By naturalization Handel was British, and he thoroughly and happily adapted himself to the English way of life and was deeply involved—not least through his many benefactions—in public affairs. He was a keen student of English literature and set the language to music with understanding of its structure and inner significance. He never married but was a lifelong friend of Mrs. Mary Delany, who first made his acquaintance as a small girl in 1710. Handel, however, never forgot that by birth he was German. As a boy he had often walked along the leafy road from Halle to his grandfather's parish. In his testament he recalled those days in generous bequests to his relatives in Germany. Although from time to time he suffered financial embarrassment (but never bankruptcy), he possessed a keen business sense and through prudent investment accumulated a considerable fortune.

Growth of Handel's reputation after his death. Handel, accorded the status of a classic in England during his

own lifetime, is unique among musicians in never having suffered diminution of reputation either as man or artist. As a young man he was aware of, and to some extent supplied, the demands of aristocratic patronage. In England, however, by adapting himself to a different climate of opinion and taste he came to serve and indeed to express the needs of a wider public. More than anyone else he democratized music, and in this respect his popular oratorios, his songs, and his best loved instrumental works have a significance transcending their purely musical importance. Handel's music became an indispensable part of England's national culture. In 1784 the first of many Handel Commemorations took place in Westminster Abbey and exerted a profound influence on the already existing cult of Handel. Amateur singers all over Britain were stimulated by the 1784 Commemoration to the extent that musical festivals, largely based on the oratorios, were held in countless urban and even rural centres. Some instrumentalists who had played in the first Commemoration emigrated to the United States and, together with organists who had preceded them, implanted an interest in Handel, which led to the foundation of the Handel and Haydn Society of Boston in 1815.

In Germany, meanwhile, interest in Handel was growing apace. In September 1771, the English conductor and composer Michael Arne directed *Alexander's Feast* in Hamburg as well as the first performance of *Messiah* in Germany in April 1772. These ventures in Hamburg, the city where Handel's music had been most cultivated in Germany, led to performances of the choral works under Carl Philipp Emanuel Bach—the first, and most important, of *Messiah*, taking place in December 1775. Through performances of his works and the critical attention paid to Handel, he became re-established in Germany as a national composer.

In the present century, the fact that Handel wrote much more music than had become generally familiar, together with a further secularization of values, has led to re-examination of works other than the oratorios. The production of his opera *Rodelinda* in Göttingen in 1920 had important consequences, leading to productions of the operas elsewhere and to the establishment of an annual Handel Festival in Göttingen. Since World War II, efforts to secure proper recognition for Handel in Halle, the city of his birth (where interest in Handel had remained sporadic), led to the formation of a new Händel-Gesellschaft (Society). The headquarters of this society is in the house in which Handel was born, now maintained as a museum and a centre for lectures and recitals. Halle has also promoted (since 1952) an annual festival of Handel's work, which has proved notable especially for the production of the operas. The most important undertaking of the Händel-Gesellschaft, however, is the inception of a new complete edition of his music. Under the supervision of an editorial board representing scholars from various countries, a number of volumes have already appeared. The vast bulk of Handel's autographs, which passed to the English Royal House and was in 1957 presented by Queen Elizabeth II to the nation, are kept together in the British Museum.

MAJOR WORKS

Theatre music

OPERAS: About 40, including *Almira* (1705), *Agrippina* (1709), *Rinaldo* (1711), *Radamisto* (1720), *Ottone* (1723), *Giulio Cesare* (1724), *Tamerlano* (1724), *Rodelinda* (1725), *Scipione* (1726), *Admeto* (1727), *Partenope* (1730), *Sosarme* (1732), *Orlando* (1733), *Ariodante* (1735), *Alcina* (1735), *Serse* (1738), *Deidamia* (1741).

Choral and other vocal works

ORATORIOS AND MAJOR SECULAR WORKS: About 20, including *Acis and Galatea* (*c.* 1720), *Esther* (1732), *Alexander's Feast* (1736), *Saul* (1739), *Israel in Egypt* (1739), *L'allegro, il penseroso ed il moderato* (1740), *Messiah* (1742), *Samson* (1743), *Semele* (1744), *Hercules* (1745), *Belshazzar* (1745), *Judas Maccabaeus* (1747), *Susanna* (1749), *Solomon* (1749), *Theodora* (1750), *Jephtha* (1752).

CHURCH WORKS: *Dixit Dominus* (1707), *Utrecht Te Deum and Jubilate* (1713), *Brockes' Passion* (*c.* 1716); 11 *Chandos Anthems* (*c.* 1717–20); four *Coronation Anthems for George II* (1727); various anthems and hymns.

OTHER VOCAL MUSIC: 22 Italian duets; 72 Italian cantatas; *Nine German Songs;* English songs; *French Cantatas* and various other cantatas.

Instrumental music

CONCERTOS: *Six Concerti Grossi* (known as *The Oboe Concertos*), op. 3; *Twelve Grand Concertos*, op. 6; 21 organ concertos; concertos for various instruments.

MISCELLANEOUS ORCHESTRAL WORKS: *Water Music* (1717); *Fireworks Music* (1749); various overtures.

CHAMBER MUSIC: *Six Sonatas for Two Oboes and Continuo* (*c.* 1696); *Six Sonatas for Two Violins, Oboes or German Flutes and Continuo*, op. 2 (*c.* 1731); *Seven Sonatas or Trios for Two Violins or German Flutes*, op. 5 (*c.* 1739); various solos, sonatas, and trios.

HARPSICHORD MUSIC: *Suites de pièces* (*c.* 1720); *Suites de pièces* (*c.* 1733); 12 fugues; *Fantasia in C Major* (*c.* 1732); various sonatas, suites, and short pieces. The so-called *Harmonious Blacksmith* variations are in No. 5 of the first set of *Suites de pièces.* The *Chaconne and Variations* is No. 9 of the second set.

BIBLIOGRAPHY. O.E. DEUTSCH, *Handel: A Documentary Biography* (1955), the life and career of Handel exposed through contemporary documents; P.M. YOUNG, *Handel*, rev. ed. (1965), a comprehensive, popular study, and *A History of British Music* (1967), includes a description of Handel in relation to British music as a whole and a discussion of the importance of the Handelian tradition; G.E.H. ABRAHAM (ed.), *Handel: A Symposium* (1954), a detailed examination of the works by various specialists, prefaced by a short study of Handel's personality; P.H. LANG, *George Frederic Handel* (1966), a monumental study.

(P.M.Y.)

Hand Tools

Hand tools, within the original meaning of the term, are manual instruments, traditionally, operated by the muscular energy of the user. The present array of hand tools have as common ancestors the sharpened stones that were the keys to early mankind's survival and rise. Rudely fractured stones, first found and later "made" by hunters who needed a general-purpose tool, were a "knife" of sorts that could also be used to hack, to pound, and to grub. In the course of a vast interval of time, a variety of single-purpose tools came into being. With the twin inventions of agriculture and animal domestication, roughly 10,000 years ago, the many demands of a settled way of life led to a higher degree of tool specialization; the identities of the ax, adz, chisel, and saw were clearly established over 4,000 years ago.

Hand tools may conveniently be categorized as those used by craftsmen in the performance of a manual operation, such as chopping, chiselling, sawing, filing, or forging, that directly shapes a piece of material into a desired form. (The development of agricultural implements is detailed in AGRICULTURE, HISTORY OF.) The common denominator of these tools is removal of material from a workpiece, usually by some form of cutting. The presence of a cutting edge is therefore characteristic of most tools, and the principal concern of toolmakers has been the pursuit and creation of improved cutting edges. Tool effectiveness was enhanced enormously by hafting—*i.e.*, the fitting of a handle to a piece of sharp stone otherwise held in the hand. An invention of almost unparalleled consequences, hafting endowed the tool with better control, more energy, or both.

The common denominator of most hand tools

After a cutting tool has functioned, other implements—*e.g.*, the carpenter's hammer for nailing—may be necessary for the completion of a task. Complementary tools, often needed as auxiliaries, include such things as tongs or a vise for holding. The advanced craftsman also needs layout instruments that outline the desired shape or establish final orientation by measuring or marking: the rule, divider, square, and others. Today, power tools perform many of the old manual operations; these too are hand tools in the modern sense.

The tools of different crafts may often be similar to each other, though proportioned to the size of the task: the heavy chisel and stout mallet of the stonemason and the dentist's delicate chisel and tiny mallet are analogous. Refinement of a tool and adaptation to a particular job are dependent upon available materials and the acquisi-

tion of experience. Special-purpose tools developed as needs for them appeared and the toolmaker's ingenuity met the challenge with new designs and the exploitation of new materials.

The date of the earliest tools is incredibly remote. Tools found in northern Kenya in 1969 have been given an age of about 2,600,000 years, and their state of development suggests that even older tools may remain to be discovered.

The history of tools is easily divided into two parts. The first phase of the story spans over 2,000,000 years, a period that saw the genesis of a few basic tools in the Stone Ages and their slow evolution to modern form in the early centuries of the Age of Metals, as first copper, then bronze, and lastly iron came under man's control as tool materials. Because it was so plentiful, iron was the ultimate metal to supplant stone and other materials. The stage was set, perhaps 4,000 years ago, for the forms of the basic tools to be defined; the ax, chisel, file, and saw of today are only the latest versions of ancient forebears.

The second part of a tool's history lies within the span of approximately the last 2,000 years and includes improvements in detail and innovations that extended the range of the tool. A number of the principal tools of this short period are treated separately, with occasional references to earlier times for the sake of clarity and completeness.

This article is divided into the following sections:

I. Early history of hand tools
- Geological and archaeological aspects
- Stone as a tool material
- Paleolithic tools
- Neolithic tools
- Early metals and smelting
- Iron and steel tools

II. Later development of hand tools
- Hammers and hammer-like tools
- Ax and adz
- Knife
- Drilling and boring tools
- Saw
- File
- Chisel
- Plane
- Workbench and vise
- Tongs, pincers, and pliers
- Wrench and screwdriver
- Plumb line, level, and square
- Compass, divider, and caliper
- Chalk line

I. Early history of hand tools

GEOLOGICAL AND ARCHAEOLOGICAL ASPECTS

The oldest known tools date from 2,600,000 years ago; geologically, this is close to the end of the Pliocene Epoch, which had extended over 4,500,000 years and was the last of five epochs constituting the Tertiary Period (the 65,000,000 years of which had seen the rise of mammals). The Pliocene was succeeded by the Pleistocene Epoch that began about 2,500,000 years ago and was terminated only recently, perhaps 10,000 years ago, with the recession of the last glaciers, when it was supplanted by the geologists' Holocene (Recent) Epoch. Pleistocene and Stone Age are in rough correspondence, for, until the first use of metal, about 5,000 years ago, stone was the principal material of tools and implements.

At first, man was a casual tool user, using a convenient stick or stone to achieve a purpose and discarding it. But man may have shared this characteristic with some other animals, and part of his differentiation from them may have begun with the deliberate making of tools to a plan and for a purpose. A cutting instrument was most essential, for, of all carnivorous animals, man is the only one not equipped with tearing claws or canine teeth long enough to pierce and rend skin: man needs a sharp tool to get through the skin to the meat. Naturally fractured pieces of stone with a sharp edge that could cut after a fashion were the first tools; they were followed by intentionally chipped stones. Archaeologically, it is the finding of primitive cutting tools intentionally made that is taken to indicate and confirm man's early presence at a site. Once understood, fire helped shape wooden implements before adequate stone tools were available for the purpose. Fire is also the basis of metallurgy. When in historic time the powers of water and wind were applied to the daily tasks of grinding grain and raising water, the way to industrialization was opened.

The idea of relating the history of man to the material from which he made his tools dates from 1836, when Christian Jürgensen Thomsen, a Danish museum director, was faced with the task of exhibiting an undocumented collection of clearly ancient tools and implements. Thomsen used three categories of materials, stone, bronze, and iron, to represent what he felt had been the ordered succession of man's technological development. The idea has since been formalized in the designation of a Stone Age, Bronze Age, and Iron Age.

Thomsen and the three-age system

The three-age system does not apply to the Americas, many Pacific Islands, or Australia, in which places no Bronze Age existed before the natives were introduced to the Iron Age or rather the products of it by European explorers. The Stone Age is still quite real in some remote regions of Australia and South America and existed in the New World at the time of Columbus' first visit. Despite these qualifications, the Stone–Bronze–Iron sequence is of value as a concept in the early history of tools.

Nor does the three-age system imply equal length for the periods; the Stone Age was of vast duration, having occupied practically all of the Pleistocene Epoch. Copper and bronze came on the scene over 5,000 years ago; iron followed in the next millennium or so and as an age includes the present.

The apparently abrupt transition from stone to bronze tends to mask the critical discovery of native metals and their utilitarian use and fails to indicate the significant discoveries of melting and casting. From bronze one can infer the really crucial discovery of smelting, the process by which most of the common metals can be recovered from their ores. Smelted copper necessarily preceded bronze; bronze is a mixture of copper and tin, the first alloy. Iron came later when technique, experience, and equipment were able to provide higher temperatures and cope with problems involved with its use.

STONE AS A TOOL MATERIAL

The Stone Age subdivides into two contrasting periods, the Old Stone Age, a long era of stagnation, and the New Stone Age, a brief period of swift progress.

During the Paleolithic Period, or Old Stone Age, flashes of keen empirical perception, few and far between, set standards for hundreds of millennia. The Paleolithic that endured until about 10,000 years ago was characterized by tools of chipped stone, cutting tools with rough and pock-marked surfaces and generally serrated cutting edges. The later Paleolithic was also an era of wood, horn (antler), and bone. These three materials, all softer than stone but nevertheless intractable, could not be worked successfully without the aid of a harder stone tool such as a serrated, or sawlike, blade and the graver, or burin, a small scraper with either pointed or narrow, chisel-like end. Bone was a particularly useful material, for its toughness made feasible barbed fishhooks, eyed needles, and small leatherworking awls.

The term Neolithic Period, or New Stone Age, defines the second period, at the beginning of which ground and usually polished stone tools, notably axes, came into widespread use after the adoption of a new technique of stoneworking. The beginning of the Neolithic, the retreat of the last glaciers, and the invention of food crops, involving agriculture and animal domestication, were more or less contemporary events. The period terminated with the discovery of metals, which showed the way to superior tools.

The revolutionary art that created the definitive ground and polished tool of Neolithic man was essentially a finishing operation that slicked a chipped tool by rubbing it on or with an abrasive stone to remove the scars of the chipping process that had produced the tool in the

rough. Not only was the edge keener than ever before but the smooth sides of the edge promoted deeper penetration and, hence, greater effectiveness, with the added advantage of easier tool extraction from a deep and wedging cut.

Stones for tool-making

As a tool material, the term stone covers a wide variety of rocks, ranging from the dense and grainless flint and obsidian to coarse-grained granite and quartzite. Each kind of stone has certain unique properties that are further influenced by heat or cold, wet or dry. Stone of any kind is difficult to manipulate, as is readily discovered by knocking two stones together with the intention of producing no more than a sharp but jagged crest. Some archaeologists have studied stoneworking techniques and become adept at producing good imitations of artifacts; further insight has been gained by observing the methods of modern stoneworking peoples in Africa and Australia. It has been noted, for example, that the Australian Aborigines reject as unsuitable a great many of the flints they have worked on, sometimes in the ratio of 300 rejects to one accepted tool. This high discard rate may help explain the thousands upon thousands of stone artifacts that have been found.

Flint, homogeneous and isotropic (equal properties in all directions), is the stone of first choice for toolmaking. Reasonably well distributed over much of the world, it is an impure quartz or silica, usually opaque and commonly of gray or smoky-brown colour. It is harder than most steels, having no cleavage planes, but instead the conchoidal, or shell-like, fracture of a brittle material that leaves a sharp edge when small or large flakes are detached from a piece (glass, which may be considered an artificial quartz, also exhibits the conchoidal fracture). Obsidian, a volcanic glass and also a silicate, but of rather limited distribution, is usually black or at least very dark and, because of its conchoidal fracture, was used like flint. Most edged stone tools, however, are of flint. Flint was once an object of trade, and flint mines were in Neolithic time what iron mines were at a later age.

Three principal types of tools appeared in the long Paleolithic Period with substantial variations occurring within each type. The types are distinguished principally by workmanship but vary also in size and appearance and are known as core, flake, and blade tools. The core tools are the largest; the earliest and most primitive were made by working on a fist-sized piece of stone (core) with a similar stone (hammerstone) and knocking off several large flakes on one side to produce a jagged but sharp crest. This was a general-purpose implement for the roughest work such as hacking, pounding, or cutting. The angle of the cutting edge was rather large because of the sphericity of the stone. With time, thinner, sharper, and more versatile core tools were developed.

Although large flakes with sharp edges of small angle were a by-product of core-tool manufacture and well suited for slitting and scraping, they were not flake tools in the proper sense. True flake tools derived from an advanced technique practiced more than 2,000,000 years later that sought the flake and discarded the core from which it had been detached; flake tools were made deliberately to serve a certain function and were not the casual spin-off of another operation. Finally, there were the blade tools, longish slivers of stone with keen unserrated edges, directly useful as knives or as stock from which other pieces might be skillfully broken to serve numerous purposes. While flake and blade tools were developing, core tools were refined by overall chipping to create thinner and more efficient forms.

Archaeologists have noted three different techniques for working stone to successive stages of refinement in the Paleolithic. The first and always basic method employed the hammerstone to fashion either a large and rude core tool such as the chopper, whose form persisted for perhaps 2,000,000 years or for roughing out (blocking in) large tool blanks that would be brought to final form by removing small flakes. The hammerstone technique produced short and deep flake scars. A variation employed the anvil stone, a large stationary stone against which the workpiece was swung to batter off some large flakes.

Soft-hammer technique

The second method was the soft-hammer, or baton, technique, based upon a discovery of perhaps 500,000 years ago that hard stone (flint in particular) could be chipped by striking it with a softer material. The baton was a light "hammer," an almost foot-long piece of bone, antler, or, even, wood, whose gentler blows detached only quite small flakes that left smooth, shallow scars. Such small flakes, when removed from the large scars left by the hammerstone, reduced the coarse and jagged edge to many small serrations, giving a straighter and more uniform cutting edge whose angle was also more acute than formerly and, hence, sharper.

Pressure flaking was the third technique. In this, a short, pointed instrument of bone, antler, or wood was used to pry (not strike) off tiny flakes to leave the smallest scars. As the least violent and most advanced of the methods of working stone, it gave the craftsman the ultimate in control for the removal of materials in the shaping of his implement.

To judge from the few remaining hand-tool making societies, it is likely that every early man was adept at making new tools quickly and easily and on the spot, as fast as the old ones were blunted or broken. The earliest, simple tools, made by taking a stone of size convenient to the hand and giving it a sharp crest by a few well-placed blows, were evidently discarded after use, for their widespread dispersal suggests that they were made at the place of use and abandoned after serving their purpose. Tens of thousands of prehistoric stone tools survive, compared with only very few bits and pieces of the skeletal remains of the makers. Stone of course is imperishable whereas bone is not, and one individual might have made several hundred tools.

The possibilities in the design of stone tools are limited by the inflexibility and brittleness of the material. The design effort was constrained to the sizing of the tool to the intended task and the development of sharper, longer, and more usefully shaped cutting edges that always required backing to support them. In use, bending and twisting of long knifelike tools had to be avoided lest the action destroy them; this would also have been true of chisels and gouges. Similarly, even the much later heavier tools, such as ax and adz, required care in use.

The effectiveness of stone tools has been demonstrated from time to time by both archaeologists and modern workmen unaccustomed to such tools. An experienced operator using a stone knife can skin a small animal about as quickly and deftly as he could using steel. When the stone tool is subjected to substantial forces, however, the worker must use caution, intelligence, and control. Care is required to avoid twisting or prying with a stone blade (knife or ax); if thin it may snap or, if thick, collect local nicks.

Quality of Stone Age tools

Ancient objects such as stone implements should not be judged by contemporary standards. The use of such adjectives as clumsy, crude, inefficient, and other depreciations is inappropriate. Given Stone Age materials and tools and within the framework of experience then possible, modern man can probably do no better than his forebears; indeed, his product might be worse because of his impatience and frustrations with the unfamiliar tools and the lack of a holding fixture. An ancient artifact and its workmanship are a problem solution, and it may be assumed that the implement was the best possible for the time and circumstances and may even represent an unprecedented achievement.

PALEOLITHIC TOOLS

Early tools are classified by their industry, or type of workmanship. Such tool traditions are identified by a name derived from the site at which the type first drew archaeological attention. For example, the primitive chopping tools that persisted for nearly 2,000,000 years, first identified in Olduvai Gorge, east of Lake Victoria, Tanzania, constitute the so-called Oldowan industry regardless of the part of the world in which implements of similar workmanship happen to be found.

The sequence of traditions shows growth and development; it does not imply abrupt transitions at certain times nor the disappearance of an old industry with the advent of another. A new technique simply meant that something better or different could be accomplished, from the refinement of the cutting edge or the upgrading of old tool forms to the manufacture of a completely new tool. Innovation sometimes was possible only by drawing upon previously unworkable materials.

It becomes apparent from an overview of the products of the successive toolmaking industries that much effort went into cutting edges in the longitudinal direction of the pieces of flint: knifelike instruments predominate and, thus, define the nature of the fundamental need, namely, that of a cutting tool, one that can slit and sever.

With the passage of time and the acquisition of skills, the average size of the tool decreased; there was more cutting edge per pound of material, an important factor when flint had to be imported to a region. This trend was reversed in the Neolithic Period when the heavy woodsman's ax and adz became essential elements to forest clearance for agriculture and timber utilization in settlements, for the world was in transition from an economy based on gathering and hunting food to a way of life founded on food raising.

Archaeologists have named the early tools by guessing at their presumed use, often in the light of other known facts about the culture in which the tradition appeared. As the tools move closer to the present and specialized forms are seen in the creation of a wider variety of products, the descriptive name is on firmer ground. A name, however, can designate only a principal or design use. Even a modern screwdriver is typically used for many tasks other than driving screws.

Eoliths. The first act of the man–tool drama is hazy. There are what have been called eoliths, "tools from the dawn of the Stone Age." Such stones with sharp fractures, found in great quantities in the layers of the geological epochs before the Pleistocene, were once assumed to be tokens of man's presence in the preceding Pliocene and even prior Miocene. These rocks, fractured by glacier pressure or wave action or temperature change, are no longer taken as indices of man, although primitive man undoubtedly used them as ready-made objects before he started deliberately to fracture similar rock in the late Pliocene. There are detail criteria by which human-flaked and nature-flaked stones can be distinguished almost unerringly. Human origin is also evidenced by association with detached flakes and the stones that served as hammers.

The oldest known tool

The tools found in 1969 at the Koobi Fora site, near Lake Rudolph in northern Kenya, consisted of five choppers, a number of flakes, and a couple of battered stones. The tools lay on the surface; three feet below them were found flakes in tuff datable to about 2,600,000 years ago. The oldest previously known tools had been from Olduvai Gorge, Tanzania. These Oldowan tools, as well as the jaw and teeth of a man who may have been the toolmaker, were found in the 1950s fortuitously lying under volcanic rock (tuff) having a potassium–argon date of about 1,800,000 years, a lower Pleistocene age.

All these tools are of a single type, a general-purpose implement that during the next 2,000,000 years hardly changed in form. It is variously known as pebble tool, pebble chopper, chopping tool, or simply as chopper (see Figure 1A). Waterworn and hence rounded, up to about the size of a fist, the pebble, preferably flattish rather than spherical, was given a few violent but skillfully applied blows by a hammerstone or pounder. Several large flakes or chips were knocked off the tool stone to create on it a sharp and roughly serrated crest or ridge, yielding an implement that was edged at one end and could be gripped at the opposite end. Rudimentary yet versatile, the chopper could be used to hack, mash, cut, grub roots, scrape, and break bones for their marrow.

Although the large, sharp-edged flakes struck from the pebble were themselves useful for light cutting and scraping, it was not until perhaps 40,000 years ago that there was a development of flake-tool industries in which preshaped flakes were purposefully detached from a core that was then discarded. But the Oldowan chopper together with the struck-off flakes—the earliest primitive, generalized tools—between them solved the problems of how to get through the skin of a slain animal, dismember it, and divide the meat.

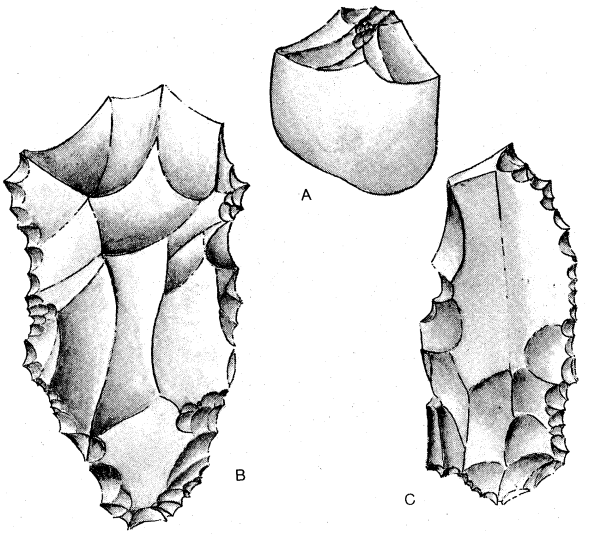


Figure 1: *Paleolithic tools.*
(A) Oldowan chopper, the oldest manufactured tool.
(B) Acheulean hand ax or fist hatchet, a superior chopper.
(C) Scraper, a flake tool.

The Acheulean and Mousterian industries. As the Pleistocene moved along, man slowly developed the primitive chopper into a better instrument. About half a million years ago a superior implement finally appeared after nearly 2,000,000 years of effort. The industry or style is known as the Acheulean, and the typical implement was the hand ax (sometimes called fist hatchet) of flint in the Western part of the world (Figure 1B). Over the ages the plump chopper and its bluntly angled crest had been streamlined by starting with a longer piece of stone and flaking the entire surface to produce an almond-shaped (amygdaloid) implement eight to ten inches (20 to 25 centimetres) long. This stone, much thinner than the chopper, was also sharper and more effective because the cutting edges were formed from the intersection of two curved and flaked surfaces (bifacial working).

The Acheulean hand ax

This Acheulean hand ax was the product of evolution; certain of the intermediate stages, clearly leading to the typical and standardized form, have been identified as Chellean and Abbevillean. Despite the term ax, the tool was not hafted but was simply held in the hand. One end was tapered, the other rounded. The tapered end might be rather pointed or have a small straight edge. The tool was sharp for most of its periphery and seems to have been primarily a hunter's knife but probably very useful, too, for other purposes such as chopping, scraping, grubbing, and even piercing. Sharp, thin and symmetrical, light and elegant, it was quite different from the heavy chopper with its rather blunt edge.

Another biface, the Acheulean cleaver, assumed prominence about 250,000 years later. A variant of the hand ax, it had a wide cutting edge across the end instead of a point, better suited than the hand ax for either hunting purposes or hacking wood.

Although the hand ax is the first and principal index of the Acheulean, the industry had a broader scope, dealing also heavily in smaller flake tools suited to lighter tasks. Of particular importance was the discovery of soft-hammer flaking, a technique described previously.

The Neanderthal tool kit

Neanderthal man, an excellent hunter and toolmaker, appeared on the scene about 110,000 years ago just ahead of the last glaciation, well within the Acheulean. His tool kit was impressive for the wide variety of hand axes, borers, knives, and choppers that it contained. The

kit was novel for the scrapers and the heavily serrated blades having a sawlike appearance, implements that were essential to the working of wood, bone, and horn into tools and weapons. The Neanderthalers regularly used fire, and it is presumed that they could make it, although the direct evidence is missing. Fire was useful in tool manufacture, for charring the end of a stick not only helps shape the point by making it easier to scrape but also hardens it, as for a spear point. This fire hardening was probably the first man-made modification of a natural property. Thoroughly wet wood, bent to shape and brought to dryness over the heat of a fire, retains its bent form, a most useful property.

Flake tools. Another distinct phase of toolmaking is the Mousterian–Levalloisian, sometimes assigned to the upper part of the Acheulean, or it may be said to overlap. This phase is marked by the major innovation called the prepared-core technique, a flake-tool industry. In this, a core was craftily trimmed in such a manner that a skillfully applied last blow would detach a large, preshaped flake directly usable as an implement; the core was discarded. Such a flake tool has one flat surface and is known as a unifacial tool because a single bevel forms the working edge. There are two principal kinds of flakes, the points and scrapers. The former are roughly triangular with two trimmed or sharp edges meeting in a point, the base or butt of the triangle being thick and blunt. The side scrapers have a sharp edge in the long direction of the flake with an opposite, thicker butt section (Figure 1C). The scraper could function as a knife, although it is speculated that it was used for working wood and skins, a supposition leading to the idea that skins were being used for clothing.

Late Paleolithic toolmaking. Phase four of Paleolithic toolmaking was introduced perhaps 40,000 years ago by Aurignacian culture, pregnant forerunner of the last and most brilliant achievements of the Old Stone Age. Extraordinary inventiveness is characteristic of this Aurignacian tradition and its several short-term successors. They can be lumped into a unit of development over the next 25,000 years.

Fully modern man—whose first representative is the Cro-Magnon—emerged within this period, perhaps 35,000 years ago, during a time of action, of development and elaboration of stone technology, which, by providing a variety of specialized tools, mostly of the flake and blade types, would at last bring materials other than stone into extensive use. It was also a time of great plains in northern and eastern Europe that carried such a heavy reindeer population, in addition to wild horses and mammoths, that it has been called the Reindeer Age. This meant a hunting economy providing food and great quantities of bone, horn, skin, sinews, and, while the mammoth lasted, ivory; with it grew new technologies exploiting the unique properties of materials hitherto unworkable because of their hardness. This technological diversification was possible only by reason of stone tools and new techniques, specialization and complexity fitting them to the fresh tasks. The most significant tool was the burin, or graver, a stout, narrow-bladed flint able to cut (scrape is more like it) narrow grooves in bone; two parallel grooves, for example, would allow a sliver of bone to be detached as stock for a needle, pin, awl, or other small object. Larger pieces of bone were worked into hooks with one or more barbs or points. Sections of antler were carved into splitting wedges to work out long pieces of bone to form the dartlike projectiles of the spear-thrower. Sandstone polishers were added to the tool kit to sharpen and shape tips and needles and other articles.

The stone burin

A spectacular item that developed by the end of the Paleolithic was the spear-thrower, the first projectile weapon. This was a hand-held stick, of wood or antler, notched at one end. Functioning as an extension of the arm, it added considerable kinetic energy to a short spear or javelin tipped with flint or bone; the toolmaking hunter no longer had to creep close to his game but could attack from a distance. The tipped projectile represented still another innovation, for it was the first hafted implement, the first evidence of fitting of a "handle" to a tool point, which heightened the utility of both small and large cutting edges.

Hafting, or the fitting of a handle to a cutting edge, was a momentous and far-reaching invention of about 35,000 years ago. By hindsight an obvious idea, it was, nevertheless, over 2,000,000 years in the making. Hafting was a critical step toward the creation of new tools and improved models of the old. A hafted tool is also called a composite or compound tool, for it consists of two or more pieces connected to form a single implement. In its simplest form, the haft may be no more than a grass or leaf bundle whose limited function is to protect the hand when a fractured stone is used as a knife. At the other extreme, where the higher function of making the previously impossible a practicality is realized, there is woodchopping: a man using a hand-held axhead can cut only small trees, whereas with a hafted ax he can fell a tree of almost any size. Mechanically, a handle, whatever its form, is a force-transmitting intermediary between the source of the force and the toolhead. The handle, viewed as an extension of the arms, provides an increased radius of swing. This moves the toolhead faster to give it more kinetic energy for a harder and more telling blow than the arms alone can provide.

The prepared-core technique that provided preshaped flakes was refined and extended to provide preshaped blades, long and slender pieces of flint of trapezoidal cross section, each corner having a straight cutting edge without the serrations of a chipped tool. This is known as the blade-tool industry, a final complement to the core- and flake-tool technologies. Such blades made thin and splendid knives of great variety; many of these knives were backed; that is to say, the back of the blade was blunted for safer handling. Thin blades were further reduced to smaller pieces, often having a geometric form such as triangular, square, or trapezoidal, called microliths. These small bits of sharp flint were cemented (using resin) into a groove in a piece of wood to form a tool with a cutting edge longer than it was feasible to produce in a single piece of brittle flint: a spear with a long cutting edge or the farmer's sickle of later date.

The second major mechanical invention of the Upper Paleolithic was the bow, a device even more effective than the spear-thrower for increasing the distance between the hunter and the hunted. It is difficult to date precisely, for the only conclusive evidence is found in cave paintings. Mere finds of stone points without bows prove nothing because such tips were used on the projectiles of spear-throwers. The earliest representations of the bow come from North Africa, 30,000–15,000 BC. Once the bow had been devised, it spread with astonishing rapidity, its effectiveness making it the weapon *par excellence*. When a bow is pulled, it stores the gradually expended energy of the archer's muscles; this energy is suddenly released to give the projectile a "muzzle velocity" far higher than that possible from a spear-thrower and of superior accuracy. It was a principal weapon through the 15th-century AD and was ousted then only by gunpowder.

The first bows

NEOLITHIC TOOLS

The Neolithic Period, or New Stone Age, the age of the ground tool, is defined by the advent, around 7000 BC, of ground and polished celts (ax and adz heads) as well as similarly treated chisels and gouges, often made of such stones as jadeite, diorite, or schist, all tougher than flint. A ground tool is one that was chipped to rough shape in the old manner and then rubbed on or with a coarse abrasive stone to remove the chip scars either from the entire surface or around the working edge (see Figure 2A). Polishing was a last step, a final grind with fine abrasive. That such a tool is pleasing to the eye is incidental; the real worth of the smoothing lies in the even cutting edge, superior strength, and better handling of the tool. The new ax would sink deeper for a given blow while delivering a clean and broad cut; its smooth bit, more shock resistant than the former flaked edge, had less tendency to wedge in a cut.

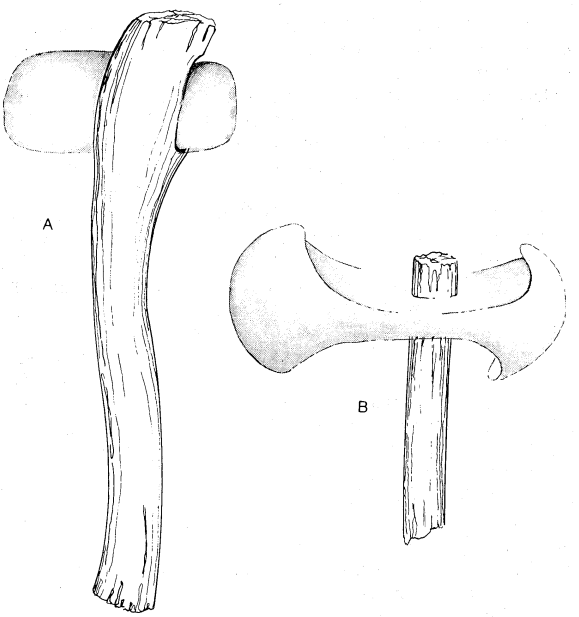


Figure 2: *Neolithic tools.*
(A) Stone ax, a ground tool, made by chipping to rough shape, then rubbing to give a nonserrated edge. (B) Stone ax made by pecking, a technique involving light, rapid blows followed by grinding.

Although the polished stone tool is the index to the Neolithic Period, it may be noted that the ice sheets were receding and climatic conditions were assisting the conversion of hunter into herdsman. The new, relatively sedentary life spawned further inventions, such as pottery. From a tool standpoint, the potters' kiln and art were necessary steps to the metals, for a modification of the kiln probably provided the high temperatures and equipment needed for metalworking, first for melting native metals and later for the ultimate smelting process that gave rise to a wealth of metals, several of which proved superior tool materials.

The polished Neolithic ax, a heavy implement, was in sharp contrast to the delicate small-stone work of the last stages of the Paleolithic and was, as has been noted earlier, a reversal of the traditions the products of which had yielded ever more lineal feet of cutting edge per pound of stone. The ax and its companion adz met the needs of forest-land clearance as agriculture developed. An efficient tree-cutting tool was indispensable for the slash-and-burn agriculture then devised and practiced the world over for millenia. Trees were either cut down or killed by ringing with an ax; the debris was burned over, with the ashes conferring a slight enrichment of the stump-filled field. The earth was next scarified with sticks or stone-headed hoes resembling the adz to prepare it for the seeding among the stumps. Without manuring or other treatment, the land was exhausted after a few years, necessitating a repetition of the clearing process elsewhere. The consequence was a shifting settlement pattern, with a good ax needed not only for felling the trees but also for working the timber for settlement purposes.

The ax in land clearance

Wood began its broad role in the life of man with the ground and polished tools of the Neolithic. Home and fire, furniture and utensils, cradle and coffin were products of the ax, adz, and chisel that could fashion wood intricately and with precision. This kit of tools turned wood into an almost universal building material, for a whole host of new things was now possible, as dugout canoes of oak, paddles, framing for hide-covered boats, sledges, skis, wooden platters, ladles, and other household gear. Mortise and tenon joints were invented for the structural framing of substantial habitations. Some of the gabled houses were up to 100 feet long and 20 wide and are believed to have served as both granaries and living quarters for perhaps 20 people comprising several families.

In a revealing experiment, 4,000-year-old polished stone axes, furnished by the Danish National Museum and carrying the sharpness left after their last use 4,000 years ago, were fitted with ash handles modelled after that of a Neolithic hafted ax preserved in a bog, giving the ax an overall length of nearly 25 inches (a modern steel felling ax has a 36-inch handle). When these were used in a Danish forest, it was soon found that the violent action of the modern technique of swinging a steel ax and putting shoulder and weight behind the blade to give long and powerful blows was disastrous, either ruining the edge or breaking the blade. Proper handling meant short, quick strokes that chipped at the tree, the body action being constrained to mainly elbow and wrist motion. After getting into form, the men found it possible to fell oak trees of more than one-foot diameter in half an hour or a two-foot-diameter pine in less than 20 minutes. Six hundred square yards (one-eighth acre) of silver-birch forest were cleared by three men in four hours. One axhead cut down more than 100 trees on its original (old) sharpening. It was concluded that Neolithic men and their ground flint axes had no great difficulties in making large clearings in the forest for the purposes of cultivation. It may be remarked that it was less trouble to clear the forest than to break the ages-old and tough sod of the plains.

The Neolithic farmer of northern Europe and his program of deforestation for agriculture were completely dependent upon polished axes, creating a heavy demand for good stone that depleted local sources. The response to this was flint mining in well-endowed locations in England, Belgium, Holland, France, Denmark, Sweden, Poland, Portugal, Sicily, and Egypt. Actually, it was often more than just mining; it would be proper to speak of ax factories, for the flints were shaped to rough form by chipping at the pithead and then traded, grinding and polishing being done by the ultimate consumer.

Flint mining

An idea of the magnitude of such a mining enterprise is offered by the well-explored workings known as Grimes Graves about 80 miles northeast of London. The site covers about 34 acres and includes both opencast workings and 40-feet-deep shafts with radiating galleries that exploited the flint deposit laid down as a floor under chalk beds. Excavation was probably by wooden shovel (product of polished ax and chisel) or possibly the shoulder blade of oxen. It is estimated that perhaps 50,000 picks made of red-deer antler were used during the 600-year-long activity begun about 2300 BC.

A last innovation of the Neolithic was the augmentation of the two older techniques of working stone, chipping (or flaking) and grinding, by a third way, the pecking, or crumbling, method (Figure 2B). In this procedure, a point of the stone being worked is bruised by a hard hammerstone, the struck points crumbling to powder under relatively light but rapidly delivered blows. This technique allowed the manufacture of tools from numerous varieties of appropriate but nonflaking stone and the production of hollow ware, such as querns for grinding grain, mortars, and bowls. It could also be applied to flakable stone, and such a stone, after having been roughed out by flaking, was pecked to level the ridges between flake scars before grinding and polishing.

Stone tools maintained themselves during the Metals Age, yielding only slowly to the new material, which was expensive and the product of specialists' skills. The copper and bronze tools and weapons for hunting, warfare, husbandry, and domestic use that constitute impressive displays in museums were rare luxuries. Even the much more abundant iron that overtook and replaced copper and bronze articles was only sparingly available for many centuries.

EARLY METALS AND SMELTING

The discovery that certain heavy "stones" did not respond to hammerblows by flaking or fracturing but were instead soft and remained intact as their shapes changed marked the end of the almost infinitely long Stone Age. Of the pure, or native, metals, gold and silver seem to have attracted attention at an early date, but both were too soft for tools. The first metals of value for toolmak-

ing were natural copper and meteoric iron. As tough and potentially versatile, if scarce, materials, they were suited for new purposes, as well as many of the old. They also introduced a new problem, corrosion.

Metalworking. Copper occurs in the native state in many parts of the world, sometimes in nuggets or lumps of convenient size. It is malleable; that is, it can be shaped by hammering while cold. This also hardens it by giving it a closer texture and allows it to carry a sharp edge, the hammered edge being capable of further improvement on an abrasive stone. After a certain amount of hammering (cold-working), copper becomes brittle, a condition that can be removed as often as necessary by heating the material and plunging it into cold water (quenching): the whole softening operation is known an annealing. Repeated annealings are necessary if much hammering is required for shaping. Nuggets were hammered into sheets, divided into strips and then pieces to be worked into arrowheads, knives, awls, choppers, and the like. Copper was also shaped by beating pieces of the soft metal into appropriately shaped stone cavities (molds).

Naturally pure, or virgin, iron (as differentiated from iron ore) is so rare as to be of no significance in the history of toolmaking, despite one spectacular occurrence in Greenland, where metallic iron and basalt happen to lie over coal beds; it is thought that the underlying iron ore, while erupting through the coal, was reduced to iron. Eskimos have long used this exceptional source of iron for knife and small-implement manufacture.

Use of meteoric iron

Meteoric iron, widely distributed but not in heavy deposits, was a highly prized material more difficult to fabricate than the softer copper. Its celestial origin was recognized by the ancients, for the Egyptians called it black copper from heaven, and the Sumerians denoted it by two characters representing heaven and fire.

This meteoric iron was beaten into tools in much the same way as copper, although it could not be forced into a mold in the manner of the softer metal. Like copper, iron hardens under the hammer and will then take a superior edge. Iron can be annealed, but the process is quite different from that of copper because, with iron, slow cooling from a high temperature is necessary. Meteoric iron is practically carbonless and, hence, cannot be hardened in the manner of steel; a high nickel content of about 8 percent makes it relatively corrosion resistant.

Much rarer than copper, meteoric iron was often used for jewelry, attested to by burial finds of necklaces of iron and gold beads, iron rings along with gold rings, and ornaments in sheet form. Small meteorites were the most convenient sources, but larger bodies were hacked at with copper and stone tools to yield tool-size pieces for knives, spear points, arrowpoints, axheads, and other implements.

Casting. A great step forward was made with the discovery that gold, silver, and copper could be melted and cast with many advantages. In casting, a liquid metal is poured into a cavity or a mold to give it the shape of the mold on congealing; casting shapes the metal to essentially final form once a proper cavity has been prepared. Some touch-up work may be needed; for an edged copper tool, such as an ax or knife, hammering the cutting side gives a keen edge.

Casting meant that the size of the tool was no longer dependent on the size of a chunk of available copper. Old tools could be added to a melt instead of being thrown out. This reuse of old metal accounts in part for the scarcity of virgin-copper implements.

To make viable procedures of melting and casting required a round of innovations, all minor but vital to success.

Pottery making, already well established, provided the knowledge of heat-based processes. Clay vessels are essential to working with fluid metal, for, in all but the most primitive operations, it is necessary to convey the melt from furnace to mold. Aside from providing crucibles, pottery making taught restructuring the fire with a deep bed of prepared charcoal to provide a heat superior to that of a simple campfire. Tongs of some sort had to be devised to carry the hot crucible; it is surmised that green branches were bent around the pot and replaced as needed.

Molds

As for the molds themselves, a number of forms were developed. The most primitive was simply an impression of a stone tool in clay or sand to give a cavity of the desired form. A more durable mold resulted when the cavity was worked into stone. Cavities of uniform depth allowed flat but profiled pieces to be cast. For example, some ax-blade castings were roughly T-shaped, the arms of the T being afterward bent around to clasp a handle of some sort, with the bottom of the T becoming the cutting edge. A one-piece mold, prepared for a dagger, could have a groove for most of the length of the cavity to provide a stiffening rib on one side. With experience, closed but longitudinally split and, hence, two-piece molds were devised, each side now having a groove down the middle to furnish a strengthening rib on both sides of the blade. Split molds for copper left much to be desired because pure copper is a poor metal for casting. It contracts a good deal on cooling and has a tendency to absorb gases and thereby become porous, blistered, and weak. Also, molten copper exposed to atmospheric oxygen contains embrittling cuprous oxide.

Smelting. Perhaps 1,000 years after man learned about melting virgin copper, he found that still another stone, a brittle one directly useless for tools, would produce liquid copper if sufficiently heated while in contact with charcoal. This step was epoch-making; it was the discovery of smelting, or the separation of a metal from a chemical compound called ore. Smelting (as differentiated from melting) was the first metallurgical operation and is still the principal method of gaining metals from their ores. Copper was the first tool metal to be won in this fashion; it was another 1,000 years before iron could be reduced from its ores.

It has been hypothesized that metallic copper and iron were discovered from the accidental reduction of their ores to metal in a campfire or other large blaze. An ordinary open wood fire has a maximum temperature of about 600° C (1,100° F), enough to produce a red heat in metals, but the necessary temperature for producing copper from an ore such as malachite is 700° to 800° C (1,300° to 1,500° F) under a proper combination of circumstances. With a big fire in a good wind and with pieces of copper ore in direct contact with much charcoal, small copper beads may be sweated out to be recovered when the ashes are cold; even if primitive man had recognized the beads as copper, he would not have had a useful amount of metal. It is more likely that copper ore was inadvertently placed in the reducing atmosphere of an overheated pottery kiln and that the strange and malleable residue was recognized as being similar to the very utilitarian copper.

As mined, raw ore is a mechanical mixture of ore proper (heavy) and earthy matter, or gangue (light); the two may be largely separated by crushing the raw ore and washing away the lighter gangue. The ore proper is a chemical compound of oxides, sulfides, carbonates, hydrates, silicates, and such impurities as arsenic and other elements in small amounts. Smelting frees the metal from the various combinations with which it is bound into the compound form. A preparatory step is to heat the washed ore (roasting or dressing) not only to dry it but also to burn off sulfides and organic matter. Early practice involved heating the ore in intimate contact with charcoal to provide the essential reducing atmosphere, producing a metallic sponge made up of metal and slag.

Some of the earliest copper smelting was terminated at the spongy stage. For chemical as well as practical reasons the iron of tools, wrought iron, continued to be worked out of the spongy mass until the Middle Ages. With copper smelting, the metallic sponge was soon allowed to remain in the furnace and subjected to draft-induced high temperature. The metal became liquid and seeped down to the hearth, as did the slag, which, being lighter than the metal, floated over it, permitting recovery of the copper.

At some time during the copper period, a new kind of "copper" happened to be made by smelting together two

separate ores, one bearing copper, the other tin. The resulting metal was recognized as being far more useful than copper alone, and the short period of copper tools came to an end. The new metal, a copper–tin alloy of mostly copper, was bronze. It was produced in the fluid state at a temperature less than that needed for copper, could be formed economically by casting, and could be hammer hardened more than copper. The tin noticeably increased the liquidity of the melt, checked the absorption of oxygen and other gases, and suppressed the formation of cuprous oxide, all features that facilitated the casting operation. A two-piece, or split, mold, impracticable for copper, works very well with bronze. Furthermore, bronze expands just a bit before solidifying and thus picks up the detail of a mold before it contracts in cooling.

Dawn of the Bronze Age

The earliest bronzes were of uneven composition. Later, the tin content was under good control, 10 percent was a common amount, with a little less for hammered goods, a little more for ornamental castings. The edges of hammered bronze tools of this composition are more than twice as hard as those obtained from copper.

The Bronze Age of tools and implements began about 3000 BC. In the course of 1,000 years the much more abundant iron supplanted bronze for tools, but bronze continued to be used in the arts.

All the early metals were expensive commodities in antiquity and were monopolized by kings, priests, and officials. Most metal was diverted to weapons manufacture for professional soldiers. Industrial use was severely limited. The metal chisel was used on the stone for buildings of state or fashioning furniture for the wealthy; the common man living in a mud or reed hut had no reason to own such a tool.

Generally speaking, molds were of baked clay, although soft stone was sometimes carved; metal molds are known from about 1000 BC. Sectional molds of three and four pieces, permitting more complex castings, are known from about 2600 BC. The earliest metal tools and implements were simply copies of existing stone models. It was only slowly that the plasticity of the new medium and especially the possibilities inherent to casting were appreciated. The stone dagger, for example, was necessarily short because of its extreme brittleness. With copper and then bronze, it became longer and was adapted to slashing as well as stabbing. Casting allowed forms impossible of execution in flaked stone, such as deeply concave surfaces. Holes could be cast in, rather than worked out of, the solid.

Sometimes the process was reversed. There were, for example, pottery imitations of bronze vessels for the poorer classes, with such necessary adjustments as a heavier lip for the pottery jug. The lines of bronze daggers have been noted in stone daggers of a later date, despite the difficulty of imitating a metal object in stone. Bronze axheads were copied in stone, even to the shaft hole, difficult to produce and impractical for a stone tool; it is possible that some of the stone replicas of bronze daggers and axes were for ceremonial rather than utilitarian purposes.

Pottery imitations of bronze

Malleable metal has several advantages over a brittle material, such as stone or bone or antler, the first of which is that it can be severely deformed without breaking and, if badly bent, can probably be returned to service after straightening. It is shock resistant and chipproof, welcome qualities for ax, adz, and chisel, and the edges can be kept keen by hammering or abrasion; sharpness is, however, inferior to that of good stone. In particular, metal allows the fashioning of many small items, articles of a size awkward to make of bone or horn, such as pins, fishhooks, and awls. Copper pins were stronger, tidier, and more attractive than the fish bones and thorns they replaced for securing clothing; even in the 3rd century BC there were shapes resembling the modern safety pin. Tweezers were invented, but whether for depilatory or surgical purpose is unknown; there are artifacts presumed to be scalpels. Plates, nails, and rivets also came early.

The most common tools were awls and pointed instruments suitable only for wood and leather. Woodworking was facilitated by the invention of the toothed copper saw, made of smelted metal and cast to shape. Edged tools—ax, adz, and chisel—at first similar to stone models—asserted themselves and, although not nearly as sharp as the tools they replaced, had the advantage of toughness and could be easily resharpened. In particular, the chisel made it possible to use cut stone for construction purposes, principally in temples and monuments. Abrasive sand under metal "saw blades" allowed stone to be cut neatly, just as the sand under tubes (made from rolled-up strips) that were turned provided a boring device for larger holes.

IRON AND STEEL TOOLS

Iron technology was derived from the known art of reducing copper and bronze. The principal requirement was a furnace capable of maintaining a reducing atmosphere—*i.e.*, one in which high temperature could be maintained from a good draft of air. The furnace had to be of sufficient height to allow the iron to drop from the smelting zone, forming a slaggy lump usually called a bloom.

After aluminum, iron is the most abundant metal, constituting about 5 percent of the earth's crust. Copper is in short supply, having a presence of, but 0.01 percent, iron is thus 500 times more plentiful. Iron ore suitable for simple smelting was widely distributed in the form of surface deposits that could be scraped up without elaborate mining procedures.

The limitations imposed by the dearth of metals in the Bronze Age were now lifted; new tools and implements became possible, and their numbers could increase until even the poorer classes would have access to metal tools. The iron of antiquity was wrought iron, a malleable and weldable material. Brittle cast iron, versatile and widely used in modern industry, was unknown to the ancients, which is just as well, for it would have been of no value for their edged tools and implements. Tools require toughness—shock or impact resistance—an inherent property of wrought iron that is subject to enhancement by forging. The earliest history of smelted iron is obscure, with the first scanty evidence of man-made iron dating from about 2500 BC in the Near East. A thousand years later, the abundance of ores had brought the displacement of copper and bronze by iron in the Hittite Empire.

During most of its history, iron was not recovered in a molten state but reduced to a spongy aggregate of iron and slag formed at a temperature well below the melting point of pure iron (1,535° C, or 2,795° F). This plastic metallic sponge was consolidated by hammering to squeeze out slag and weld the iron particles into a compact and ductile mass; thus it was called wrought iron, essentially pure iron with remnants of unexpelled slag coating the iron particles. Wrought iron contains so little carbon that it does not harden usefully when cooled rapidly (quenched). When iron containing 0.4 to 1.25 percent carbon is heated to 950° C, or 1,740° F, and then plunged into water or oil, it is hardened.

Iron importance in Near East

By about 1200 BC, when iron had become important in the Near East, man had learned how to create a steel surface, or case, on wrought iron, a case that could be hardened by heating and quenching. This case had been produced by the prolonged heating of wrought iron packed in a deep bed of glowing charcoal; the principle is that the surface of red-hot but carbonless iron readily absorbs up to 1 percent carbon from the carbon monoxide generated in the enveloping charcoal fire.

But iron did not offer up its virtues without a price, since all of the casting experience gathered from working with smelted copper and bronze was of no moment with a metal whose shape could only be changed by hammering. Moreover, the malleability of iron is less than that of copper for the same temperatures, which means that the smith has to work harder to promote a desired change in form. Stone hammers gave way to hafted bronze hammers, iron itself coming into use later. The first anvils—for copper and bronze—were convenient flat stones, to be

followed by increasingly larger cast-bronze models that in turn were superseded by rudimentary forms of the modern type, several pieces of iron being welded together. The earliest iron artifacts were of ruder appearance than the bronze articles that came before them.

A most valuable property of wrought iron is the ease with which two or more pieces may be united by hammering while the metal is at a high heat; iron has a high plasticity throughout a considerable temperature range below the welding, or white, heat. Even at the production stage, small pieces of spongy iron were united into larger blooms. Hammer-welding, however natural it was to ironworking, had been practiced before by goldsmiths and, in spite of the difficulties due to gassing, was even used for joining copper to make, for example, tape by welding together strips cut from plate. Welding became an essential production procedure. When iron tools had reached the end of a useful life, they could be used as bases for a new round of tools by welding the scrap into a blank and starting over, a process akin to the melting of copper and bronze scrap to cast new tools.

Iron led only slowly to generally superior tools and weapons, the gradual process being dependent upon the evolution of new techniques suited to fuller exploitation of its properties, the most singular being weldability. Like most new materials, iron had to overcome traditional usage and inertia and again, like most new materials, was not superior to the old in all points but only in some, involving a degree of compromise between its attractive and its less desirable qualities.

Iron compared to bronze

Iron ordinarily has twice the flexibility of bronze and is very much tougher, for a bar of iron can be bent back upon itself without fracturing, whereas a bronze bar (as a sword blade) breaks after only a light bend (bronze blades repaired by casting new metals into the factured sections are known). Bronze, in other words, is brittle when compared to iron, although copper is not. As the tin content of bronze rises, ductility is lost with the increasing hardness; with even 5 percent tin most of the malleability is missing from cold bronze, the ductility becoming practically nil at 20 percent tin content. The cutting edge of a hammered bronze tool is superior to that of its similarly treated iron counterpart, and is corrosion resistant.

In the Early Iron Age, when the metal was still in scarce supply, local armament makers were the chief consumers of the new metal. Agricultural tools, always subject to hard usage, entered the picture next, under the demands for adequate tools for forest clearance and subsequent cultivation. Axes, picks, and hoes were needed. Iron was smelted in the Near East before 2500 BC, but the Iron Age proper was 1,000 or more years in maturing. The long delay was due to several factors—tradition has already been mentioned, but there was the necessity for a strong air blast, and the time-consuming hammer forming as opposed to the easy process of casting. The fully developed Iron Age came with the discovery of hardening by carburization (addition of carbon) and heat-treating, which led to superior edge tools of great toughness.

II. Later development of hand tools

From the evolution of tools over more than 2,000,000 years with principal materials successively stone, bronze and iron, a number of particular tools emerged. Taken together, these specialized tools form an inverted pyramid resting upon the first general-purpose tool, the almost formless chopper. With the discovery of metals and the support of numerous inventions allowing their exploitation, first approximations to the modern forms of the basic tools of the craftsman established themselves, with the main thrust of further development directed at improving the cutting edges. Before speaking of specific tools, some generalities are necessary to provide an appreciation of the problems they were designed to solve.

A design is a concept that is only incompletely realized in physical form by reasons of necessary compromises imposed by the materials of the workpiece and tool, by the motion the workman can give the tool, and by the forces that can be applied. It is here that hafting, or the fitting of a handle to a cutting edge, provides the opportunity for the differentiation of tools and their wide variety of forms.

The need for specialized tools

The earliest tools were multipurpose in the same sense that an ordinary pocketknife is. Specialized tools were latecomers. A multipurpose tool, whereas able to do a number of things, does none of them as well as a tool designed or proportioned for one job and one material; compromise in function means compromise in product. The way in which a tool is hafted provides the primary distinction between knife, ax, saw, and plane. An application or craft is best served by a further specialization or form within a category: the knives of the butcher, woodcarver, and barber reflect their particular tasks. When confronted with the unusual, a skilled mechanic rises to the occasion with a special tool to cope with the situation. In the early 19th century, for example, a joiner had dozens of planes in his kit to deal with the many moldings, rabbets, and jointings he had to produce before the day of machine-made stock and mill-planed lumber.

PERCUSSIVE TOOLS

In looking at the tool spectrum, it will be seen that several involve a violent propulsion to deliver a telling blow. These have been named percussive tools; their principal representatives are the ax and hammer, one an edged tool and the other not. Under these two names are found an immense number of variations of which only a very few can be touched upon. The percussive group may also be called dynamic by reason of their swift motion and the large, short-term forces they develop. This means that mass and velocity and, hence, kinetic energy and momentum are factors related to the force generated or transmitted. Weight distribution between head and handle and mechanical properties of the head (suitability for cutting edge, lack of elasticity) must also be recognized in the design of a percussive tool. Obviously, these various influences were not formally considered during the ages-long trial-and-error evolution of a now successful tool, but their recognition now will aid in identifying the several evolutionary stages.

Percussive tools generally have handles that allow them to be swung; that is, they are endowed with kinetic energy by reason of their rapid motion. The attainable energy of a blow depends upon a number of factors, including the weight of the toolhead, the angle through which it is swung while gaining speed, the radius of the swing (handle length plus part or all of the arm length), and the muscle behind it all. There is a permissible energy level for a given task and tool, set by either the nature of the task or the material of the tool. Thus, a blacksmith flattening a one-inch iron bar wants a heavy, fairly long-handled hammer, whereas a light and short-handled hammer, used with wrist action, is appropriate to forging a small, soft-gold wire. A hafted flint ax is an effective tool, but may be destroyed if swung too hard or if twisted while in the cut. Bronze and steel axes can and do use longer handles than the stone ax and, being of tougher material, will not break under usage that would fracture a stone head.

The physics of percussive tools takes into consideration the centre of gravity and what is technically called the centre of percussion—*i.e.*, a unique point associated with a rotation, in this case the arc through which the tool is swung before delivering its blow and coming to rest. The tool's centre of gravity is readily found because it is the balance point, or location along the handle at which the tool can be picked up loosely and still remain in the horizontal position. The centre of percussion is the ideal point at which striking should occur on the toolhead to minimize the sting of the handle in the operator's hand as well as deal the maximum blow; this point will be outboard of the centre of gravity and should be as close to the centre of the head as possible. This last condition is best met with a light handle and heavy toolhead, placing the centre of gravity close to the head and the centre of percussion in an optimum location in the cutting edge. The centre of percussion is readily defined for a mechanical system in which there is a fixed pivot about which rotation

Centre of percussion

takes place; with the human body, where wrist, elbow, and shoulder joints can contribute rotations to the swing of the tool, the centre of percussion can be discussed only in principle—its location cannot be defined precisely.

It is apparent that the sheer weight of the head is of paramount importance in promoting the proper balance, or hang, to the tool. On this basis alone, the shift from stone axheads to metal was a step in the proper direction because metal heads of the same size as those of stone are about three times as heavy. With the heavier head, the centre of gravity of the hafted tool is further out, closer to the head, and the centre of percussion is more likely to be properly located in the head.

With the mallet and chisel, still other interrelations are involved. When working stone, a brittle material that responds to the sharp tool point by breaking out in small chips, the sculptor strikes many light blows to remove material. In consequence, mallets have short handles, and the amplitude of swing is small, allowing a succession of rapid blows without undue operator fatigue. To provide energy and momentum, the mallet head is heavy. Being of wood, it does not rebound in the manner of a metal head but stays on the chisel that transmits the blow to the cutting edge that focusses it into a small area of stone to be spalled off. The net effect of the proper combination of all elements, the properties of wood, chisel, and stone, the weight of the head (perhaps even heightened by reason of a lead-filled cavity), and short handle, is to waste the least energy. The wooden head is of course expendable, particularly if of one-piece clublike construction, for it becomes badly battered from contact with the metal chisel. A more refined mallet consists of separate head and handle, the head having a working face of end-grain wood.

Working metal with a chisel—the process is called chipping and was the only way of removing metal before the development of machine tools—requires that heavy blows be struck to enable the chisel to dig into the metal and lift out a chip. A steel hammer with a hardened face is used, and in this operation it is the soft end of the chisel that is battered and needs periodic dressing.

Hammers and hammer-like tools. "Hammer" is used here in a generic sense to cover the wide variety of striking tools distinguished by other names, as pounder, beetle, mallet, maul, pestle, sledge, and others, as well as the common hammer. The best known of the tools that go by the name hammer is the carpenter's claw type, but there are many others, such as rivetting, boilermaker's, bricklayer's, blacksmith's, machinist's ball- and cross-peen, stone or spalling, prospecting, and tack hammers. Each has a particular reason for its form. Such specialization was evident under the Romans, and a craftsman of the Middle Ages writes (AD 1100) of hammers having "large, medium and small" weight, with further variations of "long and slender" being coupled with a variety of faces.

Since a pounder or hammerstone was the first tool to be used, it may also have been the first to be fitted with a handle to increase the blow. Although some craftsmen of the soft metals still favoured the hand-held stone, presumably for its better "feel," hafting was an enormous technological advance. Yet it created a problem of major proportions that still persists: that the connection between handle and head must carry shock loads of high intensity, a situation even more complicated with the ax than the hammer because the ax may be subjected to twisting on becoming wedged in a cut. The most satisfactory solution for metal heads is the shaft hole in the toolhead; it is a poor solution for a stone tool because of the weakening effect, although it was tried, especially in stone imitations of bronze axheads.

The hafting problem in percussive tools

In hammer hafting, it is possible to distinguish between the long handles that allow tools to be swung to give them speed, and those simpler handles by which a tool such as a pavement tamper may simply be picked up so that it can be dropped. A long handle, even if not needed for dynamic effect (light blows), makes the tool easier to control and generally reduces operator fatigue.

The oldest form of hafted hammer, probably the miner's maul of Neolithic date, had a conical or ovoid stone head with a circumferential groove on a diameter at midheight; many such rilled stones have been found in flint, copper, and salt mines and elsewhere, though very few handles have survived. Such a stone could be bound to a short section of sapling with a branch (handle) coming off at an angle, twisted fibres or sinew serving as the ties (Figure 3A). With such a side-mounted head it is most likely that the handle's principal function was to lift the head that it might do its work by simply dropping, the binding-fastening being too weak to carry much extra shock produced by swinging the tool. Better shock resistance could come with bending a long flexible branch around the groove in the stone, secured with lashings (Figure 3B).

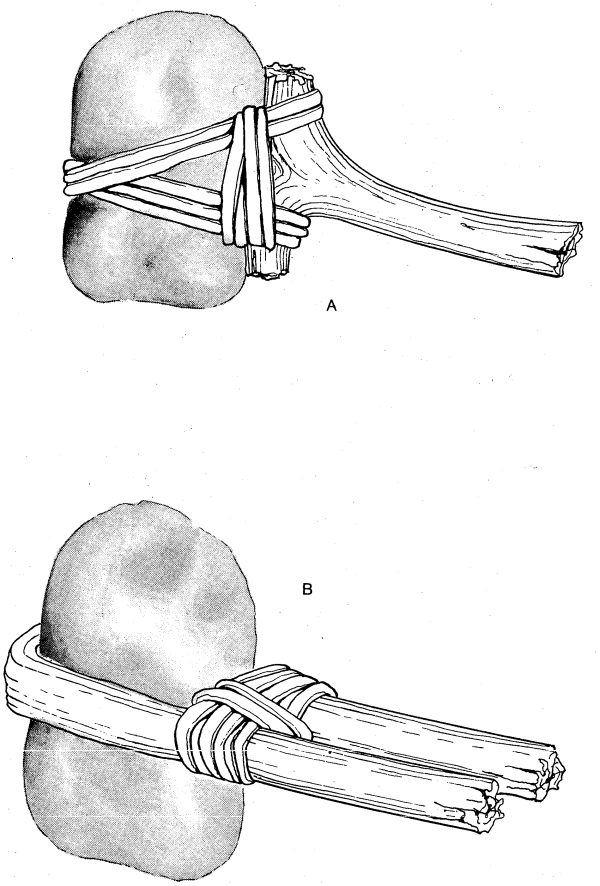


Figure 3: *Hafted hammers.*
(A) Hammer of rilled stone bound onto a knee-shaft handle, a device that was most satisfactory when merely lifted and dropped onto work. (B) Hammer bound into a bent branch; used when the work required a heavier swing.

Hammers and pounders of material other than stone were widely used; essentially clublike, they may be called self-handled. Clubs of hardwood might have one end thinned for grasping, or a mallet-like tool could be made from a short section of log with a projecting branch serving as handle. Similar mallets were made by piercing a short piece of wood and fitting a handle; this also gave an end-grain strike, more durable than a simple club. Antlers modified by trimming off tines are known from the Paleolithic Period. Such "soft" hammers were used for striking chisels of stone to prevent the destruction of the more valuable tool. Such tools, especially the wooden mallet, were used on metal chisels as well, particularly by stonecutters, because a very heavy blow on a light tool does not necessarily remove more stone than a moderate blow. There is a good deal of evidence that bone, antler, and flint wedges were used to split wood; here the use of a soft hammer would have been imperative.

The hammer as it is best known today—*i.e.*, as a tool for nailing, rivetting, and smithing—originated in the Metal Age with the inventions of nails, rivets, and jew-

elry. For beating lumps of metal into strips and sheet, heavy and compact hammers with flat faces were needed. These, in lighter form, were suited to rivetting and driving nails and wooden pegs.

In the beginning, hafting followed the stone-tool tradition. The first step away from lashing came with casting a socket opposite the head into which the short end of an L-shaped wooden handle was fitted and further supported by lashings. Such a tool was necessarily light. Ultimately the idea of piercing the head with a shaft hole (acceptable in metal where it had not been for stone) for a handle occurred to the Europeans in the Iron Age. This was several hundred years after it had become common practice among the bronzeworkers of the Near East. The shaft hole, whereas posing fastening problems that still exist, allowed heavy hammers—mauls and sledges—to be made for smithing iron.

Shaft holes for handles

The familiar claw hammer that can pull bent nails is known from Roman times in a rather well-proportioned form, for the expensive handmade nails of square or rectangular cross section did not drive easily and needed to be set into a hole prepared with an awl.

Aside from the claw hammer, other special forms of the peen—the end opposite the flat face—were developed. Hemispherical, round-edged, and wedgelike shapes helped the metalworker stretch and bend metal or the mason to chip or break stone or bricks. An especially important hammer was the filemaker's, equipped with two chisel-like heads used to score flat pieces of iron (file blanks) that were subsequently hardened by heating and quenching.

Ax and adz. The ax and adz are similar enough to be considered together. This is especially true of ancient tools that were small and ineffective because of either brittle stone or unsatisfactory hafting. The generic difference between the tools lies in the relation of the cutting edge to the handle. In the ax, the cutting edge and handle are parallel, whereas in the adz they stand at right angles. The ax and some adzes chop diagonally across the grain of the wood, but the developed adz with long handle cuts with the grain; the nature of the chips is quite different. The ax is for felling or cutting through, whereas the adz is for smoothing and levelling, although some forms were developed to scoop out gutters or dig out logs to make canoes. The great problem of both tools is satisfactory hafting; the shock impact between toolhead and handle threatens any type of connection, however ingenious.

The celt, a smooth chisel-shaped toolhead that formed either ax or adz, dates from the inventions of agriculture and animal domestication. By 6000 BC north European man was also tending toward an agricultural life; for him forest clearance was a necessity and possible only with the ground and polished ax that has been made the principal index of the Neolithic Period. The earliest true axheads, made of fine-grained rock with ground edges, are of Swedish provenance and date from about 6000 BC. Even earlier, self-handled axes, made of reindeer antler, were used. The brow tine, an antler branch running nearly at right angles to the main stem (beam), was sharpened, giving a small ax, with a haft of about eight inches (Figure 4A). By sharpening the tine the other way, a tiny adz was created. Some of these small bone implements have survived as the Lyngby tools, from a Danish site of perhaps 8000 BC.

A subsequent design socketed a stone blade in a short length of antler that was perforated for a handle (Figure 4B). This Maglemosian style, from a Danish site of about 6000 BC, was a popular model for several thousand years despite its narrow cutting edge and length of about 20 inches (50 centimetres). For comparison, it may be noted that a modern camping ax is about 12 inches (30 centimetres) long, that axes for ordinary home use are 28 inches (70 centimetres) overall, and that the professional's tool has a length of 36 inches.

The desire for a better feel or a longer cutting edge or perhaps the shortage of antlers led to a great variety of haftings. A common arrangement involved fastening heavy celts to knee-shaft handles, made from branched tree sections, by lashings. To permit the use of larger

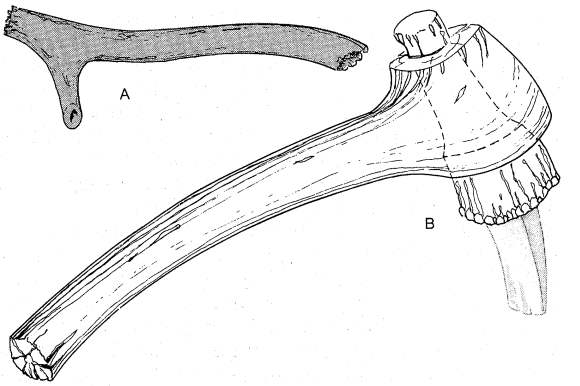


Figure 4: *Ax and adz.*
(A) Lyngby ax, a self-handled ax made from a reindeer antler.
(B) Stone adz head with a socket of antler inserted in a wooden haft.

celts, the stone was sometimes fitted into a wooden handle, but this at once created the danger of failure of the handle due to the weakening hole. Heavy, clublike handles with ample strength at the hole gave the tool an unfavourable balance.

Surviving examples of celts of soft stone are believed to have been restricted to nonwoodworking axes, such as for killing game or perhaps for certain ritual purposes. Hard-stone axes with shaft holes, often obvious imitations of bronze axes, are associated with the Bronze Age. They are among the supreme examples of stoneworking and are products of the pecking technique. From their very delicacy it may be inferred that these axes were not for the working of wood.

Parallel to the development of the ax ran that of the adz. It was often shorter handled than the ax and, because of this, was essentially a chipping tool rather than the shaving tool it became when the handle was lengthened.

The first adzes

An Egyptian relief of about 2500 BC, the pyramid-building era, shows a metal ax (copper or bronze) of curious shape, almost semicircular, lashed to a wooden handle along its diameter. The same picture shows a knee-shaft adz whose metal blade makes an angle of about 30° with the handle. If the number of pictures and artifacts of the adz is meaningful, the adz was more widely used than the ax. Generally speaking, the adz has a short handle, with angles of the order of 60° between blade and handle. Although the Egyptians became skilled metalworkers, this was not reflected in their tools, the designs of which hardly changed over 2,000 years.

On the other hand, bronze axes and adzes from Mesopotamia of even the period 2700 BC are shaft-hole types, the hole for the handle being formed in the mold. Aside from eliminating the nuisance of lashing the blades, these castings meant a heavier head than the thin-bladed Egyptian models, with better dynamic characteristics.

In the Crete of 2000 BC shaft-hole axes and adzes were also being cast, with a distinctly new tool being added to the kit. The double-bit (two-bladed), shaft-hole ax, classically associated with the Minoans, was first known about 2500 BC as a votive ax, a piece of tomb furniture made of riveted bronze plates.

The emblematic double-bit ax became a working tool on being cast in bronze with a shaft hole about 2000 BC. Double-bit adzes are also known from this time, as are ax–adz combinations. Succeeding civilizations, the Mycenean, Greek, and Roman, carried these designs forward. According to Homer, Ulysses used a double-bit ax, a type that disappeared with bronze. Illustrations or artifacts from the Middle Ages reveal only iron single-bit types, in a bewildering variety of profiles. By mid-19th century the double-bit was again in use, principally as a lumberman's ax in the United States, where it was called a back wounder by the awkward. Canada and Australia know this type, too, and it is still marketed.

The ax of Ulysses

In western Europe the advent of metal was about 500 years later than in the Middle East. In making the transi-

tion from stone to metal, the European continued the tradition of knee-shaft handle, but not in the Egyptian manner unless a flat blade (celt) had been cast. Another and more popular line of castings was given a wide slot by either forging or casting into which a cleft knee-shaft was fitted and lashed. This was the palstave. To minimize splitting of the shaft a stop was later cast in at the bottom of the slot. Subsequently, one or two eyes, or loops, were furnished in the casting to allow firmer lashing.

The socketed head, perhaps carried over from spearhead technology, was an improvement because the knee-shaft stub sat in a socket with greater security, although still requiring lashing. Like its predecessors, this tool was small, almost toylike; the cutting edges of the order of 1½ inches and short handles suggest a one-hand operation. Adzes were similarly proportioned, as were hammers.

The Bronze Age smiths of Europe were slow in inventing the shaft hole that those of the Middle East had developed in an earlier millennium. The knee-shaft tradition with socketed head entered even the Iron Age in which, at long last, the shaft-hole tools appear in Europe. To forge a socket is a difficult enough operation with even modern equipment. A shaft hole, however, is fairly simple to make by folding a rectangular plate around an iron bar, creating a roughly V-shaped forging that needs some further work to become a shaft-hole tool with cutting edge parallel to the eye, the weld being at the lower part of the eye. Such shaft-hole tools appeared in northern Europe well after the Iron Age was under way, perhaps after 500 BC. By this time, expensive bronze had been supplanted by plentiful iron for tool use.

The bronze tools had been relatively delicate in design; their iron successors soon gained size and developed in character and effectiveness to display specialized forms. Of these, two are especially important. First, there was the felling ax of the woodcutter, the blade bevelled on both sides for symmetry and often fitted with a flat end suited to driving splitting wedges. There were numerous variations within this form as it evolved toward the modern conformation, a tool distinguished by its fine balance.

The iron ax had little advantage over its bronze forerunners until smiths discovered carburization and could produce a temperable steel along the cutting edge. This must have occurred early, for repeated heating of the edge region in forging would draw at least small quantities of carbon from the charcoal of the fire. A number of Roman axes subjected to analysis have been found to contain steel.

Steeling the axhead

Steeling, or the welding of strips of steel to the iron head, was invented in the Middle Ages. The head had been rough forged by bending a properly shaped piece of flat iron stock around an iron handle pattern to form the eye. Steeling could take one of two forms. In the first, a strip of steel was inserted between the overlapping ends and the whole welded into a unit (inserted steeling). For the second, the overlapping ends were welded together and drawn to a V-shape over which a V-shaped piece of steel was then welded (overcoat, or overlaid, steeling). Inserted steeling was regarded as superior because it furnished about three times as much steel to resist loss of metal in repeated grinding and sharpening. The manufacture of steeled, or two-piece, axes ceased in the early 20th century, the heads then being made of a single piece of high-carbon steel whose properly tempered edge is backed by a tough body.

To convert felled timber to squared timber, special tools were required. As the log lay on the ground or on low blocking, vertical sides were developed by the broadax, or side ax. Somewhat shorter handled than the felling ax, it had a flat face, the single bevel being on the opposite or right side; it sliced diagonally downward as the carpenter moved backward along the log. The head is rather heavy, about twice that of a felling ax, for, whereas it is a two-handled tool, the broadax is never swung in the manner of a felling ax but raised to waist height and allowed to fall with minimum added pressure. The handle was bent, or offset to the right, to give finger clearance when hewing to the line on a debarked log. Deep vertical cuts had been made to this line with a felling ax—"scoring to the line"—after which the broadax split off the wood between score marks while hewing to the line. Hewn timber found in old buildings often carries the faint marks of the scoring.

If the timber was to be presented to view it was smoothed by the adz that removed the last of the score marks to leave a kind of ripple finish. For this purpose a long-handled adz was used, the radius of its gentle swing originating in the carpenter's shoulder. The blade was bevelled on the inside and removed material in the same sense as does a plane.

The adz was once an indispensable tool of general utility, for, in addition to surfacing, it was particularly useful for trueing and otherwise levelling framework such as posts, beams, and rafters, in setting up the frames of wooden ships and in dressing their planking. For special purposes the blade was round instead of flat, allowing the adz to cut hollows such as gutters. Dugout canoes, log coffins, and stock watering troughs, all cut from a single log, were products of the adz. Short-handled adzes, reminiscent of the ancient patterns, were used by coopers and makers of wooden bowls.

CUTTING, DRILLING, AND ABRADING TOOLS

Knife. The same jagged crest on the Paleolithic chopper that developed into the ax, in another direction developed into another broad tool category, the knife, a combination of a uniquely shaped sharp blade and a handle that optimizes the position of the cutting edge. The motion of a knife is in the direction of its edge for slicing action, in contrast to the blades of ax, adz, chisel, or plane.

The jagged edge of the chopper could slit, sever, and cut after a fashion; these operations became more effective as the bifacial-flaking technique developed to produce the Acheulean hand ax, the blade of which, whereas still short, had many small serrations instead of only a few because of the advance in flaking control. A continuous edge was achieved with the blade technique that separated slivers of stone carrying a long edge.

The knife handle

The first hafting may have taken the form of a protective pad of leaves or grass. Next, pieces of flint were set into grooves of wooden handles by cementing with resin or bitumen to leave sharp cutting edges exposed. The handle at the end of a blade had to wait until the Metals Age, which produced a longer and tougher blade that could be set into a handle, or riveted to a handgrip. Some knives, as surgical types and razors, were cast with a handle (self-handled). Copper, bronze, and iron blades were hammered to produce a locally hard edge.

Aside from the knife's utilitarian use in the field, kitchen, and workshop, variations giving it the status of a weapon appeared in the form of daggers and short and long swords. The stabbing dagger probably had its conceptual origin in the Neolithic, although an effectively thin and adequately strong blade had to wait for the Iron Age.

As various individual crafts emerged, an impressive number of convenient but single-purpose knives were fashioned to suit particular tasks, such as those of the goldbeater, the farrier, and the shoemaker, not to mention those adapted for agricultural purposes. The first illustration showing a knifesmith forging blades comes from the early 15th century.

Hunting knives, equally useful as fighting knives, developed an overall style, proportion, and balance that changed little over the centuries after the introduction of iron. The first folding knife is a Roman model of the 1st century AD. Beginning in the late Middle Ages many improvements in detail were introduced. These included fancy handles and springs and locks for the blade. Folding knives were made in many configurations from large to small, and some have come to carry scissors, bottle openers, files, and other accessories.

Drilling and boring tools. The terminology related to making holes with revolving tools is confused. A hole may be drilled or bored; there is a tool called a drill but none known as a borer; and awls, gimlets, and augers also produce holes. An awl is the simplest hole maker,

for, as a pointed instrument such as a needle, it simply pushes material to one side without removing it, whereas drills, gimlets, and augers have cutting edges that detach material to leave a hole. A drilled hole is ordinarily small and usually in metal; a bored hole large and in wood or, if in metal, is usually done by enlarging a small hole. Drilling usually requires high speed and low torque (turning force), little material being removed during each revolution of the tool. Low speed but high torque are characteristic of boring because the boring tool is of larger radius than a drill.

The Upper Paleolithic furnishes the first perforated objects of shell, ivory, antler, bone, and tooth, although softer, perishable materials, such as leather and wood, were undoubtedly provided with holes by the use of bone or antler splinters. How holes were made in harder materials is subject to speculation; it has been suggested that flint blades were trimmed to sharp points by bilateral flaking and that these points were turned by hand, a very slow process. Another scheme, evident in later time, may be ancient; it involves the use of an abrasive sand under the end of a stick that is twirled back and forth between the palms. At some unknown time, more efficient rotation was provided by wrapping a thong around the stick or shaft and pulling on the ends of the thong; this gave rotation first in one direction and then in another. Such an arrangement, the strap, or thong, drill, could be applied to drilling either with abrasive or with a tool point hafted to the end of the stick. The upper end of the shaft required a pad or socket (drill pad) in which it could rotate freely.

Invention of the bow

After the invention of the bow, sometime in the Upper Paleolithic, the ends of the thong were fastened to a bow or a slack bowstring was wrapped around the shaft to create the bow drill. This simple and effective drill was widely distributed. Because of its simplicity, it maintained itself in Europe in small shops down to the present century and is still in use in parts of the world.

Abrasive drilling in stone was well suited to the high-speed bow drill. For larger holes the amount of material that had to be reduced to powder led to the idea of using a tube, as of rolled copper strip, instead of a solid cylinder. This is the core drill, for, with abrasive trapped between rotating tube and stone, a ring is ground out, containing a core that can be removed.

A new and more complicated tool, the pump drill, was developed in Roman times. A crosspiece that could slide up and down the spindle was attached by cords that wound and unwound about it. Thus, a downward push on the crosspiece imparted a rotation to the spindle. A flywheel on the spindle kept the motion going, so that the cords rewound in reverse to raise the crosspiece as the drill slowed, and the next push down brought the spindle into rotation in the opposite direction. Whether such a drill cut while turning in either direction depended upon the cutting edge. If the edge had a double bevel, it cut in both directions.

The earliest (perhaps Bronze Age) drill points were rather casual, of more or less pointed form with sharp edges that ultimately developed into conscious arrow shapes with two distinct cutting edges. This shape was quite effective, especially when made of iron or steel, and remained popular until the end of the 19th century, when factory-made, spiral-fluted drills became available at reasonable cost to displace the blacksmith-made articles.

The basic auger originated in the Iron Age; a tool for enlarging existing holes, it had a crossbar so that it might be turned with two hands. It resembled a half pipe and was sharpened in several ways, as on the inside of the semicircular end, along the length, or both. The end might be forged into a spoon shape and the edges sharpened so that cutting might take place at the bottom of the hole in addition to the sides. To clear the hole of parings it was necessary to pull the auger and turn the workpiece over. Augers with spiral or helical stems that bring the shavings or chips to the surface are an invention of the Middle Ages, though one example is known from Roman Britain.

The so-called breast auger is a short version in which the operator could apply pressure by leaning on a drill pad or socket in which the upper end of the auger could swivel freely.

The brace and breast swivel

The familiar and common brace, a crank with a breast swivel at one end and a drill point at the other, is first seen in a painting of about 1425 that shows Joseph at his bench. This brace, as well as later but still early examples, is shown fitted with a bit of small diameter. It has been suggested that the function of the new tool was to make the small, or pilot, holes for the larger auger bit. This is a reasonable assumption, for the crank, fashioned from a wide board, had insufficient strength because of the cross grain to drive a large bit (see Figure 5A). This

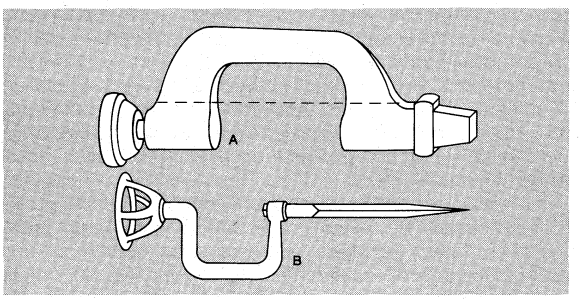


Figure 5: *Braces.*
(A) Early one-piece wooden brace showing weak sections (dotted lines). (B) All-metal blacksmith's "wimble" of late 17th century; the metal crank was not generally applied to the woodworker's brace until two centuries later.

weakness was later counteracted by reinforcing the two weak sections with metal plates, a practice continued until about 1900 despite the commercial introduction of iron sweeps (cranks) about 1860 (Figure 5B). With this, holes up to one inch in diameter can be bored with the one-handed operation; larger holes still require the two-handed augers. An iron sweep is noted in a German manuscript of 1505, and an English book of 1683 has a metal brace as part of a blacksmith's kit.

The early wooden braces were equipped with a large socket into which bits with appropriate shanks could be fitted to give interchangeability. When the sweep came to be made of iron, the bits were given square shanks that fitted into simple split chucks (holders) secured with a thumbscrew, and it was not long before the screwed shell chuck and ratchet appeared to set the standard for the modern tool. By 1900 the swivel was turning on ball bearings instead of a leather washer, and the metal parts were nickel-plated.

The bow and pump drills, suitable only to small work, required two hands, one to steady the tool, the other to operate it. One-hand drills began to appear about 1825. Their essential elements are a steeply pitched screw and a nut that mates with it; when the latter is pushed down, the screw and attached bit turn. Many variations of the principle were offered before the modern push drill assumed its present, convenient form. It is still suitable for only light work in wood.

Both the bow and pump drills remained the metalworker's prime tool for small holes until the beginning of the 19th century, when the first geared hand drill was invented (1805). Like every other tool, it underwent many improvements before acquiring its present rugged simplicity. Its great advantage lies in its unidirectional motion and the gearing that rotates the drill faster than the rate at which the crank is turned. The one-directional motion allowed better drills to be designed, and, with their greater efficiency in chip production, it was not long (1822) before drills with spiral flutes were proposed. Again there was a manufacturing problem—the flutes had to be hand filed—that was not solved until the 1860s and the invention of a milling machine for making the now universal twist drills.

Augers were used for boring both across the grain of wood and along the grain. The latter operation produced wooden pipes and pump casings or wheel hubs; special bits of many forms were designed for these purposes.

The modern auger

The common use of the auger or bit was in the cross-grain direction to make holes for wooden pins (treenails, or trunnels) or bolts for connections. The modern auger bit has a screw ahead of the cutting edges to pull the auger into the work. This screw provides an automatic feed and relieves the workman of the necessity of pushing the tool. The idea appeared in the mid-16th century but was long in being generally applied, for the screw had to be hand filed, a disadvantage not overcome until the advent of screw-making machinery in the mid-19th century.

Saw. The chipped flint knife with its irregular edge, as differentiated from a straight-edged blade tool, has been called a saw because of the serrations' superficial resemblance to sawteeth. A little reflection will show that the tool was not a saw in the proper sense at all, for though it could sever wood fibers and gash bone or horn, it could not possibly remove small pieces of material in the manner of a saw. Furthermore, the necessarily broad V-shaped profile of this flint "saw" severely limited its penetration into the workpiece; an encircling groove on a branch or a notch on something flat defined the nature of the cut.

The real saw, a blade with teeth, one of the first great innovations of the Metals Age, was a completely new tool, able to cut through wood instead of merely gashing the surface. It developed with smelted copper from which a blade could be cast. Many of the early copper saws have the general appearance of large meat-carving knives, even to bone or wooden handles riveted to a tang at one end. Egyptian illustrations from about 1500 BC onward show the saw being used to rip boards, the timber being lashed to a vertical post set into the ground. The teeth were raked to cut on the pull stroke, for the hooks of the teeth point toward the handle.

The use of relatively narrow, thin, and not quite flat blades, made of a metal having a tendency to buckle and with poorly shaped teeth with high friction, required that the cutting take place on the pull stroke. In this stroke the sawyer could exert the most force without peril of buckling the saw. Furthermore, a pull saw could be thinner than a push saw under the circumstances, making for less waste of material.

The familiar modern handsaw, with a thin but wide steel blade, cuts on the push stroke, which permits downhand sawing on wood laid across the knee or on a stool, the sawing pressure helping to hold the wood still. With this sawing on the push stroke, operator control is superior, and, because the line being sawn to is not obscured by the fuzz of undetached wood fibers or sawdust, greater accuracy in following the line is possible. Some tree-pruning saws have teeth raked to cut on the pull stroke to urge the branch toward the operator instead of away from him. Blades that are thin and narrow, as in the coping saw (fretsaw or scroll saw), are pulled through the work by the frame holding the blade. The thin hacksaw blade is pulled by the frame the operator pushes. Electric reciprocating and sabre saws, the narrow blades of which are supported at only one end, pull the blade when cutting to prevent buckling. The carpenter's pull saw for wood was for a long period forgotten by the Western world but kept alive in China and Japan, where some craftsmen still favour it. The craftsman sits on the floor and uses his feet to restrain the wood. The pull saw is also used in some parts of Albania and Greece.

Although there is no positive evidence of either the method or the saw, the Egyptians were able to saw hard stone with copper and bronze implements. The blade, probably toothless, rode on an abrasive material such as moistened quartz sand. The 7½-foot (two-metre) granite coffer still in the Great Pyramid carries saw marks.

During the Bronze Age, the use of saws for woodworking was greatly extended, and the modern form approached. There are some saws with narrow blades looking very much like hacksaw blades, even to the holes at either end. It is supposed that they might have been held in a frame or pinned into a springy bow of wood, either arrangement keeping the narrow blade taut and furnishing a handle as well.

Iron saws, looking much like those of copper or bronze, are known from the middle of the 7th century BC. A major contribution to saw design was noted in the 1st century AD by Pliny the Elder, whose works are one of the major, though often fanciful sources, on the technology of the ancients. Pliny observed that setting the teeth—that is, bending alternate teeth to one side or the other, creating a kerf, or saw slot wider than the thickness of the blade—helps discharge the sawdust. He seems to have missed the more practical point that the saw also runs with less friction in the now wider slot. The Romans, always ingenious mechanics, added numerous improvements to both simply handled saws and frame saws but did not make push saws despite the advantage of the kerf that made the saw easier to work and less liable to buckle. Roman saw sets and files have been found in substantial numbers. The small handsaws were sometimes backed with a stiffening rib, to prevent the buckling of thin blades; today's backsaw still carries the rib. Frame saws, in which a narrow blade is held in tension by a wooden frame (see Figure 6), were exploited in

Roman improvements in saws

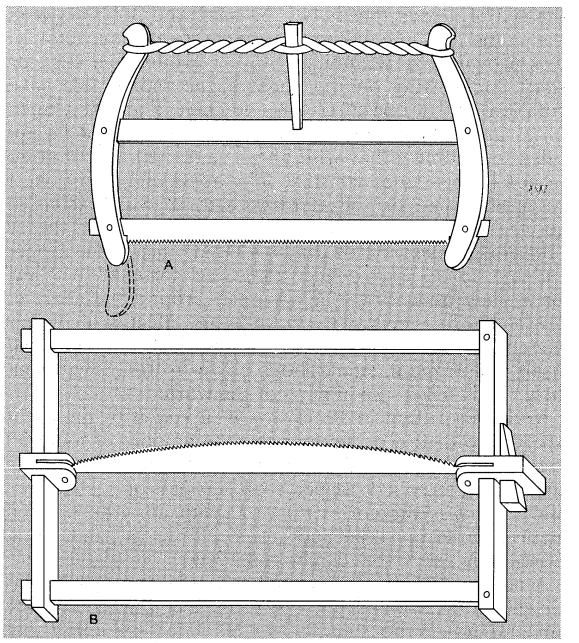


Figure 6: Framed saws in which a thin blade is put under tension (A) by a cord and twisting stick and (B) by a wedge.

many sizes, from the small carpenter's saw to two-man crosscut saw and ripsaws for making boards.

The time and provenance of the push saw are uncertain, although it appears that it may be dated from the end of Roman times, well before the Middle Ages. Nevertheless, after the decline of the Roman Empire in the West, the use of the saw seems to have declined as well. The ax again became the principal tool on the return to the more primitive state of technology. Saw artifacts are very few in number, and even the Bayeux tapestry of about 1100 shows no saw in the fairly detailed panels dealing with the construction of William the Conqueror's invasion fleet: only ax, adz, hammer, and breast auger are among the woodworking tools. It may be remarked that pioneers have always found the ax a universal tool, able to take care of the fundamental needs and, whereas wasteful of material, rugged and easily cared for.

With the Middle Ages came the search for a nonclogging tooth for use when crosscutting green and wet wood. The saws themselves were long, with handles at both ends, so that two men might each pull, alternate teeth being raked in opposite directions. To provide space for the cuttings, M-shaped teeth with gaps (gullets) between them were developed; this tooth conformation, first noted in the mid-15th century, is still modern practice in crosscut saws for coarse work and heavy timber.

Perhaps even more important than crosscutting was the need to rip a log lengthwise to produce boards. Saws for

this purpose were generally called pit saws, for they were operated in the vertical plane by two men, one of whom, the pitman, stood sometimes in a pit below the timber, or under a trestle supporting the timber being sawn. His mate stood on the timber above, pulling the saw up; the pitman and gravity did the work of cutting on the downstroke for which the teeth were raked. A pit saw might be just a long blade with two handles, giving the so-called whipsaw, but more often it was constructed as a frame saw, which used less steel and put the blade under tension.

A new invention, a result of progress in spring-driven clocks, came in the mid-16th century. It was the fretsaw, in which a U-shaped metal frame stretched a narrow blade made from a clock spring, the best and most uniform steel available, for it was not forged but rolled in small, hand-powered mills. These very thin (for the time) blades with their fine teeth were well suited to cutting veneer stock from decorative wood for furniture of all kinds.

Water-powered rolling mill products

By the middle of the 17th century, larger, waterpowered rolling mills in England and some parts of the Continent were able to furnish broad strips of steel from which wide saws could be fashioned in many varieties. In particular, the awkward framed pit saw was largely replaced by a long, two-handled blade of adequate stiffness. Smaller general-purpose saws, especially in England, developed from the rolling-mill stock into the broad-blade saws of today. The modern broad-blade handsaw is taper ground, that is, the blade is not of uniform thickness, but is several thousandths of an inch thinner at the back than at the toothed edge. This makes for no-bind cutting, and such saws require little set for fast and easy cutting. The Continental craftsman, however, is fond of the frame saw for even his benchwork; the only purchased part is the blade itself, the worker often making his own wooden frame, which is tightened by twisting a cord with a short stick.

File. The file's many tiny chisel-like teeth all point in the direction in which it must be pushed in order to be effective. Because little material is removed with each stroke, the tool is well suited to smoothing a rough workpiece or altering its shape in substantial detail. The file was unknown in early antiquity, during which smoothing was done with abrasive stone or powder or with sharkskin, the granular surface of which approximates sandpaper or abrasive cloth. In Greek, an older form of the word file is also the name of a fish having a strong skin that is used for the finishing of wood or marble.

Files of copper are unknown, but bronze, the inherently hard derivative of copper, was shaped into flat files in Egypt about 1500 BC. A combined round and flat file of bronze is known from Europe by 400 BC. The file became popular in the Iron Age, a number of specimens surviving from Roman times. The longest is flat, about 15 inches (38 centimetres) long including the handle, leaving about eight inches (20 centimetres) of working length one inch (2.5 centimetres) wide. A number of shorter files of about four-inch (ten-centimetre) working length are particularly interesting because of the notch they carry near the handle. This notch, together with their V-shaped cross section (called knife-shaped today) indicates that these files were intended for dressing sawteeth, the notch enabling the workman to set the teeth—*i.e.*, bend successive teeth to alternate sides to gain a free-running saw. These files had straight-across and coarse toothing, but the advantages of obliquely cut teeth and of the double-cut (intersecting) teeth were appreciated early.

A treatise of 1100 AD mentions files of square, round, triangular, and other shapes. At this time files were made of carburized steel that could be hardened after completion of the cutting, which was done with either a sharp, chisel-like hammer or chisel and hammer. An illustrated manuscript of 1405 shows a file of polygonal section that is copied by a succession of later authors; the screeching of the filing operation is commented upon, too, with the curious suggestion that files be made hollow and filled with lead to eliminate the noise. A writer on things mechanical in 1578 asserted that the only way in which threads could be cut in screws was with the file.

Even earlier Leonardo da Vinci had sketched a file-making machine to do a better job than the freehand technique could produce. The first file-making machine, however, has a date of 1750, and it was a century later before machine-cut files replaced those cut by hand. Power-driven, hand-cut rotary files are still used on dense metals because the hand-formed and discontinuous teeth dissipate the heat better.

The ordinary file, in terms of its material and cut, is for use on cast iron and soft steel. Other materials—various nonferrous alloys, stainless steels, plastics—are better accommodated with files of special composition and tooth formation (cut). A wide selection is manufactured.

Rasps, or, more correctly, rasp-cut files, have a series of individual teeth produced by a sharp, narrow, punchlike chisel. The very rough cut is suited to soft substances, such as wood, hoofs, leather, aluminum, and lead, for fast removal of material.

Chisel. It may be speculated that the remote origin of the chisel lay with the stone hand ax, the almond-shaped tool that was sharp at one end. Chisel-shaped flints, long and of rectangular cross section, appear about 8000 BC. The somewhat later Neolithic time brought a more workmanlike version that was finished by grinding. With care, flint and, especially, obsidian chisels can be used on soft stone, as intricate sculptures in pre-Columbian South and Central America testify. Gouges—*i.e.*, chisels with concave instead of flat sections, able to scoop hollows or form holes with curved instead of flat walls—are also known from this period. Chisels and gouges of very hard stone were used to rough out both the exteriors and interiors of bowls of softer stone such as alabaster, gypsum, soapstone, and volcanic rock, with final finish by abrasion and polishing.

The advent of smelted copper meant that a tough and readily sharpened material could replace the brittle stone. Cutting edges were hardened by hammering. The earliest copper chisels were long in the manner of their flint forebears. Such so-called solid chisels of copper (and later of bronze) were used for working not only wood but soft rock as well, as many magnificent Egyptian monuments of limestone and sandstone testify.

Copper chisels

With bronze, a better casting metal than copper, and with the experience of making molds, it was possible to economize on metal by hafting a short chisel to a wooden handle, which was also less injurious to the mallet. The round handle was either impaled on a tang with a cast-on stop (tanged) or set into a socket (socketed); both forms of hafting presage modern forms. The Egyptians used the chisel and clublike mallet with great skill and imagination to make joints in the construction of small drawers, panelled boxes, furniture, caskets, and chests.

The use of iron meant that tools had to be forged; no longer were the flowing lines and easily made cavities of casting available to the toolmaker. In consequence, early iron chisels are rude and solid. Tanged chisels were made with less trouble than socketed chisels, for which the socket had to be bent from a T-shaped forging. Hardenable steel edges were developed at first accidentally and then deliberately by giving the iron repeated contact with carbon from the charcoal of the forge fire.

Chisels and gouges came to be made in great variety in later centuries as generally increasing wealth brought a demand for more decoration and luxury in both churchly and secular fittings and furniture. The rough and heavy tools of the carpenter were refined to more delicate models suited to the woodcarvers, to the joiners who did wall panelling and made stairs, doors, and windows, and the cabinetmakers. In the 18th century a woodcarver's kit might contain upward of 70 chisels and gouges. Some special-purpose chisels of cabinetmaker, coachmaker and patternmaker had quite long handles and were never used with a mallet but worked from the shoulder.

Plane. The plane is a cleverly hafted cutting edge the function of which is to skin or shave the surface of the wood (see Figure 7). Its purpose is to finish and true

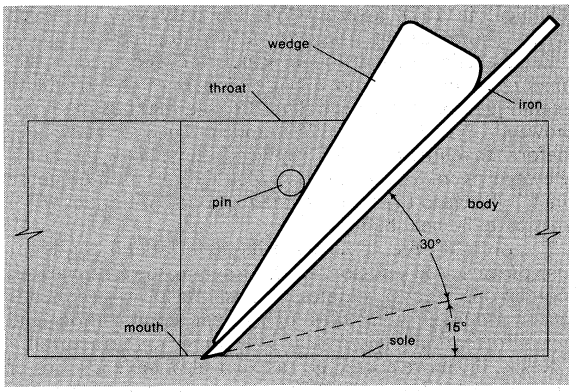


Figure 7: *Essential features of a plane.*
The body is pierced by a wide throat with a narrow slit (mouth) on the bottom, or sole. In this case, the iron (blade) is held in place by a wedge and pin, an arrangement common from Roman times to the 16th century. The cutting edge makes a 15° angle with the horizontal.

a surface by removing the marks of a previous tool (adz, ax, or saw), leaving the surface smooth, flat, and straight. Two tools, the plane and its kin the spokeshave, are unique, since both depend upon a constant depth of cut that is given by the slight projection of the blade beyond the sole, or base. The plane is a luxury item associated with the good life; while convenient to have, it is not an imperative tool such as the ax, adz, chisel, saw, and even drill.

The plane is an anomaly, for which no line of descent has yet been identified. Pliny the Elder ascribes the invention of the plane to Daedalus, the mythical Greek representative of all handiwork.

Origins of the plane

It has been suggested that the Paleolithic unifacial (flat) scraper is the remote ancestor of the plane. It is true that localized planing of a very poor sort, as in removing high spots, can be done with such a scraper, but the difference in design and action between the scraper and plane is too great to permit invoking the scraper as forerunner of the plane. The adz seems a more likely progenitor. Early adzes were bevelled (sloped) on the outside, although later on, with better hafting and longer handles, the bevel was moved to the inside. The blade and handle of an outside-bevelled adz could be used in a plane-like fashion to lift a shaving; the tool would chatter badly and cut discontinuously but might lead to the provision of a sole from which only a small portion of the blade projected and which would provide support. The control of the blade projection, or depth of cut (or thickness of shaving), is critical to the concept of the plane and is met with in only one other tool, the spokeshave.

The earliest illustrations of wood finishing, the surfacing of pieces of furniture, are Egyptian and show the surfaces being scrubbed with flat objects taken to be abrasive stone or blocks riding on abrasive sand.

Presumably the surfaces had been dressed by an adz, but now the marks of this tool needed to be erased. Stone scrapers are not in evidence, and, although the adz is shown, it is being used as an adz, not as an improvisation of a plane.

The Romans are the first definitely known users of the plane, the earliest examples coming from Pompeii. In a manner of speaking, these planes are full-blown, without a prehistory, without even vague antecedents. The modern plane differs in details, but not in principle nor in general appearance.

These Pompeii planes are of comfortable size, being about eight inches (20 centimetres) long and 2¼ inches (57 millimetres) wide. The blade was relatively narrow, about 1½ inches against the modern width of two inches (five centimetres). The sole was made of one-quarter-inch-thick iron bent to form a shallow box that was filled with a wooden core, cut away at the back to form a handgrip, while the mouth was cut in about one-third of the way from the front. The blade, or plane iron, was held in position by a wooden wedge tapped under an iron bar across the mouth. Frontier posts in Great Britain and Germany have yielded nearly a dozen Roman planes, ranging in length from 13 to 17 inches (33 to 43 centimetres). Three constructions are represented: iron sole plus wooden core, all wood, and wood reinforced with iron plates at the sides of the mouth.

Planes may be divided into two main categories, the first, the common bench plane, with straight-across iron and flat sole for working flat surfaces such as panelling, and the second, all others, in a wide variety defined by the profile of the iron and sole. If the iron has a concavity, a projection or molding will remain; if the iron has a projection, a groove will be dug. Generally speaking, planes with profiled irons and correspondingly fluted soles are molding planes, with each molding identified by its own name. Some of the Roman planes show irons for cutting rectangular grooves.

After the decline of the Roman Empire, the plane almost disappeared from view, although there is mention about AD 1100, of iron strips being planed to create a good, lighttight match when laid together. Practically no planes and only a very few other tools have survived from the period AD 800–1600. Secondary sources, such as illuminated manuscripts, legal documents, carvings, and stained-glass windows do provide some information, but details are obscure either because of the coarseness of the medium or the fact that the artist did not understand the fine points of the craft he was representing.

Categories of planes

By the late 17th century the plane was again firmly established, with the distinction between bench planes and the myriad molding planes being quite clear. Bench planes might be called the common planes; they were used for surfacing panels, or for creating straight edges on boards so that two or more might be made up into a wide panel. Boards were sawed or split (riven) from the log and were, in consequence, quite rough. The first planing operation was done with the roughing, or fore, plane of medium length, possibly 16–18 inches (41–46 centimetres). This fore plane had a slightly convex iron that removed the saw and adz marks but left hollows. The fore plane was followed by planes with straight irons that levelled the hollows. If the workpiece was long, a long-bodied, or trying, or jointing plane of, say, 30-inch (75-centimetre) length was necessary to remove large curves in the wood. Short planes—a common length was of the order of nine inches (23 centimetres)—were called smoothing planes for the final finish they gave.

Planes with straight irons and flat soles could be made by the craftsman himself without undue expenditure of time. There was, however, a change in taste and fashion in the 17th century, wood carving giving way to decorative features such as moldings and beadings; also the generally lighter furniture and a demand for wooden window sash for larger panes of glass called for molding and grooves. The proliferation of plane types established plane making as an industry.

The indispensable common plane with the straight iron was improved in a number of details to ease the working and increase the accuracy of the tool. When speaking of Roman planes, it was noted that the wedge holding the iron was jammed against a cross bar through the mouth of the plane. This feature, awkward because it impaired the free escape of the shaving, was eliminated in the 16th century by seating the wedge in tapered grooves let into the sides of the mouth. An early attempt to dispense with the wedge altogether by securing the iron with a thumbscrew remains an isolated case. In England and America wooden planes were gradually displaced by cast-iron bodies with wooden handles as advanced metallurgy and machine tools allowed good castings to be machined accurately on a mass-production basis. Metal predominated by the mid-19th century, but this was also the beginning of the decline of the molding planes, for machinery was now making millwork, tongue-and-groove flooring, door and window moldings, picture rails, baseboards, sash stuff, stair work, blinds, and much more.

The top iron, an English innovation

A capital improvement was the invention of the top iron, apparently an English innovation toward the end of the 18th century. This top iron, or chip breaker,

is like an inverted plane iron placed over the cutting iron; it limits the thickness of the shaving and helps it to curl out of the mouth. Now called the double iron, it is a feature of all but the smallest of modern planes.

The 19th century saw much effort, again in England and America, aimed at eliminating the wedge and the need for hammer adjustment of the iron. Various methods for easy and accurate setting of the iron, as well as its quick removal for sharpening, culminated in the screw and lever adjustments of the iron and the cam-actuated cap. This final evolution was completed about 1890, and changes since that time have been trivial. Despite their advantages, continental Europe has not been partial to iron-bodied planes with screw and lever adjustments, such tools commanding a much higher price than the still common wooden plane with wedge and hammer adjustment.

The spokeshave, which may be likened to a short-bodied plane with a handle on either side allowing the tool to be pulled toward the operator, has left little in the way of a record. The term has been known some 400 years, but the earliest example seems to be only half as old. Both the English word and the German *Speichenhobel* suggest that the tool was originally a wheelwright's specialization that became generalized for use on convex surfaces. As with the plane, the cutting blade (iron) projects only slightly from the short sole to regulate the depth of cut.

The drawknife is a handled blade that is pulled toward the operator. It may be considered as a rather questionable relative of the plane, for it lifts shavings in a similar manner, without, however, having the positive thickness control of the plane. In the modern form, the tangs at the ends of the knife are bent at right angles in the plane of the blade, and, while operating in much the manner of a spokeshave, it is a roughing tool for quick removal of stock. Skill is required in its use, for the blade tends either to follow the grain or to skip over it, the depth of cut being regulated by the tilt of the blade, until the wood grain asserts itself. The drawknife appears to be an older tool than the spokeshave and has undergone a change since the Viking time from which it is first known. Under the Viking craftsmen the handles were bent at right angles to the plane of the blade, and the tool seems to have been used for smoothing axed or adzed timber, in medieval Scandinavia, Russia, and elsewhere.

TOOL AUXILIARIES

Workbench and vise. Workbench and vise form an organic unit, for the vise, a fixture, is either part of the carpenter's bench or is attached to the machinist's bench; in consequence, the stories of bench and vise merge, as do those of the benchwork of the woodworkers and metalworkers.

Neither bench nor mechanical fixture provides an advantage when chipping or flaking stone. On the contrary, complete freedom in the positioning of workpiece and hammer is essential in such stoneworking to give the most advantageous apposition of stone and hammer, for many relatively small yet discretely placed and directed blows are the crux of fashioning stone tools. When large and unidirectional forces need to be applied, as in woodworking and many phases of metalworking or even in the manipulation of bone and horn, the advantage of a bench or a fixed rest of some sort becomes apparent.

Wood assumed its important role in structures and furniture and fittings with the polished stone tools (ax and chisel) of the Neolithic Period and was skillfully exploited for finer work under copper and bronze tools. Most of the furniture of ancient time no longer exists because of fire or rot. There is, however, a good deal of visual evidence, provided largely by sculptures, representations on vases, mosaics, and wall frescoes, that shows all manner of furniture: thrones, stools, benches, footstools, couches, cupboards, tables, chests, beds. Oddly enough, a stout table or workbench is missing from the renderings of busy Egyptian shops. The workpieces are on the floor with the craftsman kneeling or bending over his work or sitting on a low stool, even in those scenes in which tables are being finished. It is not known how the joinery, including the mortise-and-tenon joints and other niceties, was carried out; perhaps the craftsman used his feet to position the work on the floor while using chisel and mallet, a practice still known in some areas. Notably, a sloping board supported by a prop, probably a tanner's beam, is depicted in a leatherworker's shop. Timber being ripped is shown lashed to a vertical post; perhaps this may be viewed as a primitive vise. Little is known of Egyptian use of the copper pull saw for crosscutting in the absence of pictures, but, if the tool was used, the workman probably sat on the floor, bracing his feet against the timber that the saw was urging toward him.

Evidences of ancient furniture

Evidence in Europe for even Neolithic time suggests that the woodworker made use of the table, the distinction between a table and a workbench perhaps depending upon the kind of litter on the top or the decorative quality of the feet. Such a table bench was of the simplest form, a short length of heavy board split from a trunk and supported on four legs made of saplings set into bored holes.

This style of bench becomes quite evident in Roman times, often being only a length of half log with four legs somewhat splayed for greater stability. It is with the Romans that a stout bench became a necessity, since they were the first users of the plane, and truing a surface without a bench on which to lay and secure the wood is nearly impossible.

Two methods, still in use, were devised for holding the workpiece. The simplest procedure was to use wooden pegs set into holes in the bench top; the other was to use what are variously known as bench stops, holdfasts, or dogs, the stem of such more or less T-shaped iron fittings also being set into holes, one end of the horizontal part of the T being sharp to engage the wood.

Other arrangements came into use, as trestles for supporting wood to be sawed with the improved saws or specialized benches—horses—on which the leatherworker or coppersmith sat while facing a raised section arranged for his particular job. The small workpiece was often held by means of a strap that was made taut by a foot in the loop that formed the free end. Such horses proliferated from medieval times onward as new specialties developed; modified versions often serve today as cocktail tables.

The anvil and horn

A frequent accessory of the metalworker's bench was the anvil, still informally present on many machinist's vises in the form of a rudimentary anvil and horn suited to light work. Aside from making castings, metalworking was largely concerned with forging. The earliest anvils were convenient flat stones, usable for only the simplest kind of flat work. With increasing skill, anvils with the characteristic overhang, or horn, were first cast in bronze and, later, welded from short lengths of iron. Bench anvils were necessarily small, since large free-standing specimens of the smith had to await the cast iron that made larger masses of metal conveniently available.

Pictorial records are scant until the 15th century, when it becomes evident that the medieval carpenter's bench was still very much like the Roman, with pegs serving as end fixtures. The metalworker, especially the man using the by now indispensable file to shape and clean small forgings and castings (harness gear, buckles, and so on), used a simple rest, essentially a notch cut into the top of a post driven into the ground in front of his bench, to support his workpiece.

Within a century, according to the pictorial record, the metalworkers' rest disappears, being replaced by a screw vise, at first quite small. This vise was like a hinge, one leaf or jaw being fastened to the bench, with the other pulled up to clamp the workpiece by tightening the nut on a bolt passing through the "hinge" at midheight. The square nut was on the far side of this vise and tightened by means of a box wrench. Even clamp-on vises—the portable kind that can be applied to a handy plank or tabletop or bench top—are known from 1570.

It was, of course, slow and awkward to close the vise by running up the tightening nut with a wrench, and,

by the end of the 16th century, the inversion had taken place—the screw was now turned, from the front, by means of the T-handle that is part of every modern vise. The vise now had a form that would remain an integral element of every smithy as long as the trade lasted (see Figure 8).

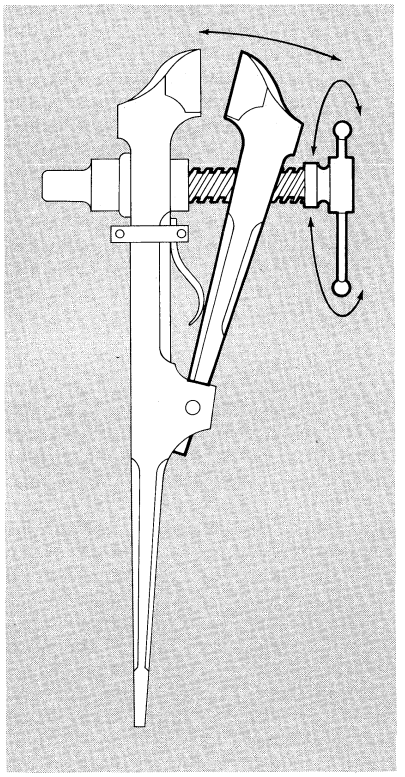


Figure 8: Blacksmith's vise, with long spike-like foot for extra support.

The modern machinist's vise has jaws that run parallel, and the vise as a whole may be turned about a vertical axis (swivel-base vise). Both of these features came before the end of the 18th century.

To return to the carpenter's bench, it remained in the 15th century equipped only with pegs to restrain the wood being planed, the older (by well over 1,000 years) Roman iron dogs, or holdfasts, not having been re-invented. Real fixturing of the workpieces is possible only with a screw arrangement of some sort, and, while all the necessary elements were described as early as 1505, nothing came of this, the approach to bench vises and indeed the use of the screw, well-known to the woodworker, taking long to mature.

Two kinds of vises are necessary to the woodworker. One is to hold (clamp) the board so that its long edges may be trued and planed; custom places this vise at the left front of the bench, a location convenient for the right-handed workman. The second vise is at the right end on the side; its moving jaw has a bench stop, and long pieces of wood to be surfaced are caught between this adjustable stop and a fixed stop in the bench top at an appropriate place. Both types of vise had been developed and even made part of the same bench by the early 19th century.

Tongs, pincers, and pliers. Tongs, pincers, tweezers, and pliers have in common the task of holding or gripping something for ease in handling. Usually, but not always, the arrangement is such that the gripping force is larger than the operator-applied force.

Handling hot coals

After man came to use fire he acquired a new problem, that of handling hot coals. Two sticks probably first served as uncertain holders. An Egyptian wall painting of about 1450 BC shows a crucible that is supported between two bow-shaped metal bars, bronze having replaced wooden sticks for this purpose about 3000 BC. The same painting shows a craftsman with a blowpipe in his mouth holding a small object over a fire with a tweezer-like instrument about eight to ten inches (20 to 25 centimetres) long. It was at about this time that bronze loops that were able to cope with large and heavy crucibles made their appearance.

Spring-back, or tweezer-like, tongs were the model used by the early ironsmith. The change to the mechanically more effective hinged tongs was slow, and it was not until 500 BC that they became common enough to give the blacksmith's kit, as shown on the vases from the classical age of Greece, a quite modern look. Pivoted tongs, with short jaws but long handle, have quite a mechanical advantage, a pair of 20-inch (50-centimetre) tongs being able to exert a gripping force of say 300 pounds (135 kilograms) for a 40-pound (18-kilogram) squeeze from the smith's hand. Such tongs had one handle slightly shorter than the other so that an oval ring could be slipped over the two to help secure the grip. This old feature of locking the tongs onto the workpiece is in common use today.

Tongs both large and small, the latter often called pliers or forceps, proliferated into many uses. A significant variation in which the jaws were sharpened were the cutters known from Roman remains, as well as the pincers that were useful for pulling bent nails because of the leverage they were capable of exerting. Although they were originally a carpenter's tool, pincers were a principal tool of the farrier, since the old nails had to be pulled from horses' hoofs before new shoes could be fitted and nailed on.

SCREW-BASED TOOLS

Invention of the screw. Before speaking of either wrench or screwdriver one must glance at the screw, the reason for both tools. Archimedes in the 3rd century BC is credited with the invention of the screw. But Archimedes' screw was not the ordinary fastener now known so well in its many varieties but had two other forms. One was a kind of water pump, still used today for large-volume, low-lift applications, seen in its essentials in the inclined screw conveyor of industry. The second was the "endless screw," actually the worm of a worm and gear set, one of the ancients' five devices for the raising of heavy weights. With the state of the mechanic arts as it was then, this concept of the screw, really a motion-transforming device, was more hypothetical than practical.

By the 1st century BC, heavy wooden screws were elements of presses for making wine and olive oil and pressing cloth. The character of the screw had been given a new dimension, for these screws were used to exert pressure; their modern counterparts are called power screws. Such press screws were turned by means of hand-spikes thrust into radial holes in the cylindrical end. The principal problem, making the nut with its internal thread, would defer the use of small threaded fasteners in metal construction, where either nuts or threaded holes are needed. The external thread was readily if tediously made by filing.

Problem of the nut

Metal screws and nuts appear in the 15th century, the square or hexagonal head or nut being turned with an appropriate box wrench; a T-handled socket wrench is also known from the 16th century. Some screws to be seen in 16th-century armour have slots (nicks) in which a screwdriver may have been used, although this tool is not shown. Deep notches on the circumferences of the heads of other armour screws suggest that some kind of pronged device was used to turn them. A drawing of a machine—again, 16th century—shows slotted, round-headed screws but, of course, no tool to turn them is depicted. These metal screws would be called machine, or machinery, screws since they are of metal and mate with threaded holes. It is worthy of mention that few screw-and-nut-fastened clocks are in evidence earlier than the 17th century, parts until that time being secured by little wedges driven into holes.

The wood screw differs from the machine screw in that the wood into which it is turned is deformed into a

nut. It must, however, be started in a hole made by awl or drill. Aside from a few and sometimes doubtful artifacts from Roman times, the wood screw does not appear until the mid-16th century, and then it appears in a mining treatise. Here a screw tapered to a point and, carrying a slotted head, looking very familiar except for its left-handed thread, is described so casually as to suggest that it was a common article. It is remarked that the screw is superior to the nail, which is also shown being driven by a claw hammer. There is no mention of a screwdriver.

Screwdrivers and wrenches. The simple screwdriver is preceded, at least in illustration, by a flat-bladed bit for the carpenter's brace (1744). The handled screwdriver begins to be shown on the woodworker's bench after 1800 and appears in inventories of tool kits from that date. Screwdrivers did not become common tools until automatic screw machines, about 1850, began the mass production of tapered, gimlet-pointed wood screws. In its early form, the screwdriver was made from flat stock, with sometimes scalloped edges that contributed nothing to function. Being flat, the blade was easy to haft but weak when improperly used for prying. This and the easy availability of round-wire stock led to the present form, round and flattened only at the end.

Origin of the screwdriver

Early box and socket wrenches (see Figure 9) fit only

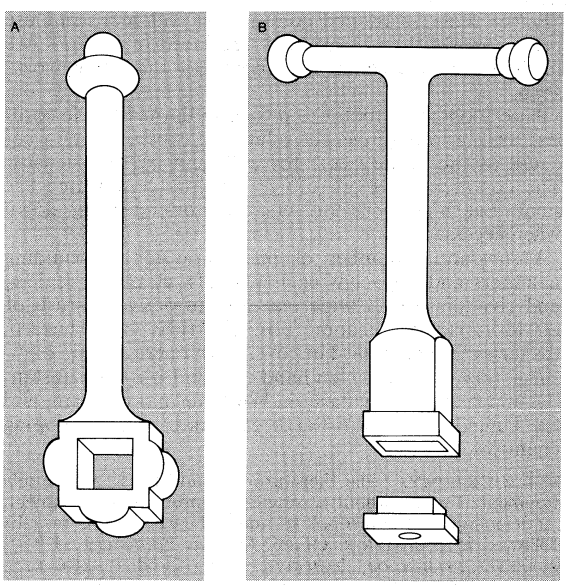


Figure 9: (A) Box and (B) socket wrenches; 16th century.

a particular nut or screw with flats on the head. The open-end wrench may have rectangular slots on one or both ends. In their earliest forms such wrenches, with either straight, angled, or S-shaped handles, were of wrought iron, cast iron entering the picture around 1800. Modern wrenches are drop forgings and come in many formats.

The very real limitations of fixed-opening wrenches were challenged in the early 19th century by sliding-jaw types able to accommodate to a range of flats. In these, the end of an L-shaped handle provided the fixed jaw, a parallel jaw being arranged to slide along the handle to engage the flats. With the first models of the 1830s, the sliding jaw was fixed in position by a wedge that was hammer tapped into place. By 1835 patents for screw wrenches began to proliferate: the sliding jaw was positioned and held by means of a screw whose axis was parallel to the handle. The most familiar example is the monkey wrench, whose name first appeared about 1858. A convenient variation is found in the thin and angled crescent wrench, a fairly recent innovation.

Although not a carpenter's tool, a serrated-jaw variation of the monkey wrench, with the additional feature of a pivotable movable jaw and called a pipe wrench by the plumber, is of interest since it can engage rounds, such as rod and pipe.

MEASURING AND DEFINING TOOLS

Plumb line, level, and square. A plumb line is a light line with a weight (plumb bob) at one end which, when suspended from the other, defines a vertical. "Plumb" comes from the Latin *plumbum*, or "lead," a material that replaced stone as the weight for the bob or plummet. Modern plumb bobs are often nickel-plated cones whose hollow interiors are filled with mercury, the greater weight allowing such bobs to come to rest more quickly in a breeze or draft.

While an end-weighted string defines a vertical, its direct use for plumbing walls (making them vertical) is awkward. The Egyptians devised a tool like a letter E, a plumb line being suspended from the upper outboard part of the E. With the E placed against a wall, the wall was vertical when the string just touched the lower outboard part of the E. The middle bar of the E can be dispensed with. Oddly, this useful tool was apparently forgotten for many centuries, reappearing only in modern times.

The tool for checking the horizontal direction is a level. The Egyptians used an A-frame, suspending a plumb line from the vertex of the A. The feet of the A were set on the surface to be checked, and, if the plumb line bisected the crossbar of the A, the surface was horizontal.

The A-frame level came to Europe and survived until the middle of the 19th century. Sometimes a variation is shown in which the frame is an inverted T with a plumb line suspended from the top of the vertical stem; the horizontal part was laid against the surface to be checked, and, if the line bisected the part, the latter was horizontal.

The hose level

Because the surface of a body of water is horizontal, a trough or channel filled with water can serve as reference for a limited number of situations. The hose level is described in 1629. In this, a length of hose is fitted with a glass tube at each end and water added until it rises in both the vertically held tubes, the free surfaces of the water in each tube coming to the same level. This excellent idea was impractical as long as only leather hose was available. Vulcanized rubber hose (after 1831) allowed the idea to be again put forward (1849). With this instrument, levels can be established under awkward circumstances, since the hose can be carried through holes in the wall, around partitions, and so on.

The spirit, or bubble, level, a sealed glass tube containing alcohol and an air bubble stems from 1661. It was first used on telescopes and later on surveying instruments and did not become a carpenter's tool until the factory-made article of the mid-19th century. Tubes with carefully made interior curvature date from 1812. The circular level, in which a bubble floats under a circular glass to indicate level in all directions, was invented in 1777. It lacks the sensitivity of the conventional level.

The square appears in the ancient Egyptian world as two perpendicular legs of wood braced with a diagonal member. In the following centuries many variations were designed for specific purposes, as, for example, a square with shoulders that allow it to also cast a mitre (45 degrees). Iron was rarely used before 1800. Factory-made metal squares appear around 1835. The adjustable, or bevel, square for angles other than 90 degrees may be noted for the 17th century. In the earliest, the thin blade moved stiffly, being rivetted into a slot in the thick blade. Later models, as from the 19th century, are equipped with a thumbscrew to fix the thin blade with respect to the thicker.

Compass, divider, and caliper. Compass, divider, and caliper are often given plural forms, frequently as "a pair of." Basically, all three instruments have two legs pivoted to each other at the top and are concerned with small-distance measurement or transfer. The compass and divider have straight legs, the caliper curved.

The terms compass and divider are often interchanged, for each instrument can be used to draw circles or mark out divisions (divide a given distance) or simply mark out a distance. Technically, a compass as drafting instru-

ment has one pen or pencil point, the other being sharp and set into the centre of the circle to be described; a divider has two sharp points, one for the centre, the other for scribing or marking. Caliper is a corruption of "calibre," the diameter of a hole (as in a firearm) or of a cylindrical or spherical body. Inward-curved legs mark the outside caliper for measuring the diameters of solids of revolution, as lathe-turned objects; legs with their tips bent outwards are inside calipers for dealing with bores.

Dividers and calipers were known to both the Greeks and Romans, though the caliper was uncommon. A divider with circular sector (or wing) connecting the two legs was sketched in 1245; its modern counterpart is the wing divider with thumbscrew clamp and screw for fine adjustment. The caliper is mentioned in the Middle Ages, a time when the divider was the principal tool of the architect for the full-scale layouts of stonework, as in the construction of a cathedral. Such dividers were large, often half as tall as a man. The divider underwent refinement to become a drafting instrument for Albrecht Dürer and Leonardo da Vinci; da Vinci suggested improvements such as the knuckle-joint hinge for increased rigidity and proportional dividers that were adjustable (Roman proportional dividers with fixed pivot gave only one ratio). Da Vinci's notes also show the beam compass with screw adjustment for large radii, as well as a compass with interchangeable points, one leg having a clamp for different drawing media, as graphite or chalk.

Proportional dividers

Chalk line. To "snap a line" is a technique familiar to ancient Egypt and to modern building construction. The principle is that a taut, chalk-covered line or cord stretched between two points on a surface will deposit a trace of chalk when the line is plucked and released (twanged) to snap onto the surface. About the only change in 5,000 years is that the Egyptians used wet red or yellow ochre, while the modern craftsman follows the usage of Greek masons who employed white and red chalks in addition to wet ochre.

Rules. The unit of linear measure in the ancient world, the cubit, derived its name from association with elbow and forearm as the length from elbow to the extremity of the middle finger. This gave an order of magnitude but hardly a standard, and, in fact, the length of the cubit varied widely in different times and places.

To speak of but one of many, the royal Egyptian cubit had a length of 20.64 inches (52.4 centimetres). It was divided into seven palms (measured across the fingers, not knuckles) making a palm very nearly three inches. Each palm corresponded to four digits of about three-quarters of an inch. Thus, 1 cubit = 7 palms = 28 digits. On occasion, digits were subdivided into tenths, fourteenths or sixteenths.

The common rule of the Egyptian masons and carpenters was of wood, with narrow cross section and one bevelled edge, the two left-hand palms carrying the smaller divisions of digits. Stone rods of late Egyptian date are known, with digits divided into 16ths. It is supposed that these were ceremonial rods or perhaps master gauges for calibration and comparison; their brittleness would make them unsuitable for the rough handling of mason's tools.

The Romans introduced folding rules of bronze in 12- and six-inch (30- and 15-centimetre) sizes. It is suspected that these were probably "pocket" instruments for officials, since their high cost would argue against their use by ordinary craftsmen, who probably used plain strip rules.

There is only the scantiest evidence for the Middle Ages and Renaissance in the way of graduated rules, plain straightedges being prominent. Paintings and sculptures show rather odd and uncertain markings. By 1683 an English writer speaks of foot rules having ⅛-inch subdivisions. The folding rule, but this time of wood, reappears at the end of the 17th century.

Because measurement was long characterized by great national and even regional differences, with every large city in Europe and most towns having a different but locally standard "foot," rules with four different graduations, one on each face, were made. When the metric system came in at the end of the 18th century, special rules were created to bridge the transition.

Craftsmen made rules to suit their own convenience, choosing a hard and well-seasoned wood for the purpose. The folding foot rule with brass tips and hinges may be dated about 1813. Mass production of two-foot rules with three joints and brass tips seem to have started around 1840.

Power tools. Power tool is the popular name for what is technically a power-driven hand tool or, sometimes, portable power tool, to distinguish it from the stationary power tool such as the drill press.

While power tools are generally driven by electricity, the category also includes small pneumatic tools driven by compressed air, such as air impact wrenches and hammers. Gasoline-engine-driven tools (chain saws, gas-powered drills) are not included.

The two most popular power tools

The most popular power tools are the electric drill and the electric circular saw. The drill rotates a tool bit, but the circular saw has no manual prototype. Jigsaws, sabre, and reciprocating saws have familiar blades and so do electric screwdrivers, but many of the power tools are recent creations built around the ubiquitous electric motor. Among the new tools are polishers, several kinds of sanders (circular, belt, oscillating, and reciprocating), shears, and nibblers.

Power tools, in limited commercial and industrial use before World War II, are now produced in the millions in the developed countries, largely for the home workshop.

Battery-operated tools are presently limited to the cordless drill, an attractive but relatively expensive item because of the battery cost for sufficient current density. The common tools operating from the power line use small but high-speed universal motors, gearing down when necessary.

A substantial number of pneumatic tools, including hammers and impact wrenches, drills, chippers, scalers, and rivetters are in wide use. An attractive feature of air tools is the freedom from electrical-shock hazard, the largest single hazard in power-tool operation.

The total power-driven-hand-tool industry is substantial and rising, particularly in West Germany, Japan, and the United States, with large exports to the developing countries.

BIBLIOGRAPHY. The literature on hand tools is generally fragmented and without a single comprehensive treatment. Archaeology and anthropology and the earliest tools may be followed in: CHESTER S. CHARD, *Man in Prehistory* (1969); ROBERT J. BRAIDWOOD, *Prehistoric Men*, 7th ed. (1968); KENNETH P. OAKLEY, *Man the Tool-Maker*, 3rd ed. rev. (1966); and F. CLARK HOWELL, *Early Man* (1965). Specific treatments are given in FRANCOIS BORDES, *Typologie du Paléolithique ancien et moyen* (1961; Eng. trans., *The Old Stone Age*, 1968); and JACQUES BORDAZ, *Tools of the Old and New Stone Age* (1970). The rise of metal tools is found in LESLIE AITCHISON, *A History of Metals*, 2 vol. (1960); and THOMAS A. RICKARD, *Man and Metals*, 2 vol. (1932). From Roman times onward, WILLIAM L. GOODMAN, *The History of Woodworking Tools* (1964), is definitive and comes to the present; but HENRY C. MERCER, *Ancient Carpenters' Tools*, 3rd ed.(1960), should not be missed; and PETER C. WELSH, *Woodworking Tools, 1600–1900* (1966), and FRANK H. WILDUNG, *Woodworking Tools at Shelburne Museum* (1957), include a wealth of illustrations. CHARLES SINGER *et al.* (eds.), *A History of Technology*, 5 vol. (1954–58); and MAURICE DAUMAS (ed.), *Histoire générale des techniques*, 3 vol. (1962–68; Eng. trans., *A History of Technology and Invention*, 2 vol., 1969–70), give wide but unconnected treatments. Illustrated by original drawings, ERIC SLOANE, *A Museum of Early American Tools* (1964); and EDWIN TUNIS, *Colonial Craftsmen and the Beginnings of American Industry* (1965), are highly informative of their period since techniques and products are shown in addition to tools. Articles relevant to this topic may also be found in serial publications such as *Beiträge zur Geschichte der Technik und Industrie*, now simply *Technikgeschichte* (quarterly); the *Transactions of the Newcomen Society* (annual); and *Technology and Culture* (quarterly). The *EAIA* (Early American Industries Association) *Chronicle* frequently contains articles on the history of hand tools.

(R.S.H.)

Han-fei-tzu

Han-fei-tzu (in Pin-yin romanization Han-fei-zu), ancient Chinese philosopher, is considered the greatest of the Legalist thinkers. He brought this school of thought, which contributed to the emergence of China's autocratic government, to the height of its development. The book that goes by his name comprises a synthesis of legal theories up to his time.

Life

Little is known of Han-fei's life. Member of the ruling family of Han, one of the weaker Warring States that were in conflict during the 5th–3rd centuries BC, he studied under the Confucian philosopher Hsün-tzu but deserted him to follow another school of thought more germane to the conditions accompanying the collapse of the feudal system in his time. Finding that his advice to the ruler of his native state went unheeded, he put his ideas into writing. A speech defect is also reputed to have induced his recourse to writing. King Cheng of Ch'in, a western state (who became the First Emperor of the Ch'in dynasty in 221 BC), read and admired some of his essays. When in 234 BC Cheng launched an attack on Han, the ruler of Han dispatched Han-fei to negotiate with Ch'in. Cheng was delighted to receive him and probably planned to offer him a high government post. Li Ssu, the chief minister and a former schoolmate of Han-fei's, presumably afraid that he might gain the King's favour by virtue of superior erudition, had him imprisoned on a charge of duplicity. Complying with Li Ssu's order to commit suicide, he drank the poison Li Ssu sent him and thus ended his life (233 BC).

To Han-fei it was axiomatic that political institutions must change with changing historical circumstances. It is folly, he said, to cling to outmoded ways of the past, as the Confucians did. It was also axiomatic that political institutions must be adapted to the prevailing pattern of human behaviour, which is determined not by moral sentiments but by economic conditions. In a year of famine people do not feed their own kin, while in a year of plenty they feast casual visitors—not because they are alternately heartless and generous but "because of the difference in the amount of food to be had." In ancient times, when goods were abundant, men made light of them, but increased population pressure on resources brought economic scarcity; consequently, "men of today quarrel and snatch." The ruler, therefore, should not try to make men good but only to restrain them from doing evil. Nor should he try "to win the hearts of the people" because, selfish as men are, they do not know their own true interests. The people's mind is as undependable as an infant's.

Royal authority: the notions of *shih* and *shu*

According to the Confucians, as virtue confers on a king the right to rule, misrule voids that right. Han-fei thought differently. Whatever the ruler's moral qualities and however he rules, possession of authority (*shih*) carries the inalienable right to exact obedience. "Subject serving ruler, son serving father, and wife serving husband" together constitute "an immutable principle of the world." Even if a lord of men is unworthy, no subject would dare to infringe his prerogative. Moreover, political duty takes precedence over other duties. A soldier, it was said, ran from battle because he thought that, if he was killed, he could no longer serve his father. Han-fei commented: "A filial son to his father can be a traitorous subject to his ruler."

Authority should be wielded not whimsically but through laws (*fa*) that the ruler promulgates and that all must obey. "The intelligent ruler makes the law select men and makes no arbitrary appointment himself; he makes the law measure merits and makes no arbitrary judgment himself." He can abrogate a law, but, so long as he allows it to stand, he should observe it.

To insure an effective bureaucracy and to protect his authority from encroachment or usurpation, the ruler must make use of *shu* ("administrative techniques" or "statecraft"). Rulers of the Warring States found it advantageous to employ men skilled in government, diplomacy, and war. But how to separate solid talent from idle chatter became a serious problem. *Shu* was Han-fei's answer to the problem. After assigning posts according to individual capacities, the ruler should demand satisfactory performance of the responsibilities devolving on their posts and punish anyone who is derelict of duty or oversteps his power. The ruler may authorize an official to carry out a proposal he has submitted. He should punish him not only when the results fall short of the stated goal but also when they exceed it.

Shu is also Han-fei's answer to the problem of usurpation, through which more than one ruler had lost his throne. The interest of the ruler and ruled are incompatible: "Superior and inferior wage one hundred battles a day." Therefore, it behooves the ruler to trust no one; to be suspicious of sycophants; to permit no one to gain undue power or influence; and, above all, to use wile to unearth plots against the throne.

Theories of military power and economic strength

With supreme authority secure and good order prevailing, the ruler proceeds to aggrandize his realm by means of military power. Might is the decisive factor in interstate relations. Military power is inseparable from economic strength. Farming being the only productive occupation, all other callings, especially that of the scholar, should be discouraged. Giving relief to the destitute is both unwise and unfair. To collect taxes from the rich in order to help the poor "is robbing the diligent and frugal and indulging the extravagant and lazy."

BIBLIOGRAPHY. Han-fei-tzu's writings, presumably compiled after his death, are entitled the *Han Fei Tzu*, comprising 55 sections of varying lengths. W.K. LIAO (trans.), *The Complete Works of Han Fei Tzu*, 2 vol. (1939–59), is the only complete English translation; BURTON WATSON (trans.), *Han Fei Tzu: Basic Writings* (1964), translates 12 of the more important sections.

Studies of Han-fei-tzu and his school of thought include: H.G. CREEL, "The Totalitarianism of the Legalists," in *Chinese Thought from Confucius to Mao Tsê-tung* (1953), a perceptive account of legalism; FUNG YU-LAN, "Han Fei Tzŭ and the Other Legalists," in *A History of Chinese Philosophy*, 2nd ed., vol. 1 (Eng. trans. 1952); "Han Fei Tzu and the Legalist School," in *A Short History of Chinese Philosophy* (1948); LIANG CH'I CHAO, "The Legalist School," in *History of Chinese Political Thought During the Early Tsin Period* (1930, reprinted 1968); LIN MOU-SHENG, "The Realistic School: Han-tzu," in *Men and Ideas: An Informal History of Chinese Political Thought* (1942), a convenient summary of leading ideas; ARTHUR WALEY, "The Realists," in *Three Ways of Thought in Ancient China* (1939); and JOHN C.H. WU, "Chinese Legal Philosophy: A Brief Historical Survey," *Chinese Culture*, 1:7–48 (1958).

(K.-c.H.)

Hannibal

One of the greatest military leaders of antiquity, Hannibal commanded, almost single-handedly, the Carthaginian forces against Rome in the Second Punic War. Hannibal was born in late 247 BC, the son of the great Carthaginian general Hamilcar Barca. According to Polybius and Livy, the main Latin sources for his life, he was taken to Spain by his father and at an early age was made to swear eternal hostility to Rome. From the death of his father in 229/228 until his own death in 183/182, Hannibal's life was one of constant struggle against the Roman republic.

Early career

His earliest commands were given to him in the Carthaginian province of Spain by Hasdrubal, son-in-law and successor of Hamilcar; and, although nothing is known of the details of his earliest campaigns, it is clear that he emerged as a successful officer, for, on the assassination of Hasdrubal in 221 BC, the army proclaimed him, at the age of 26, its commander in chief, and the Carthaginian government quickly ratified his field appointment.

Hannibal immediately turned himself to the consolidation of the Punic hold on Spain. He married a Spanish princess, Imilce, then began to conquer various Spanish tribes. He fought against the Olcades and captured their capital, Althaea; quelled the Vaccaei in the northwest; and in 221, making the seaport Cartagena his base, won a resounding victory over the Carpetani in the region of the Tagus River.

In the spring of 219 BC Hannibal made an attack on Saguntum, an independent Iberian city south of the Ebro

River. In the treaty between Rome and Carthage subsequent to the First Punic War (264–241), the Ebro had been set as the northern limit of Carthaginian influence in the Iberian Peninsula. Saguntum was indeed south of the Ebro, but the Romans had "friendship" (though perhaps not an actual treaty) with the city and regarded the Carthaginian attack on it as an act of war. The siege of Saguntum was a protracted one (eight months), and in it Hannibal was severely wounded. The Romans, who had sent envoys to Carthage in protest (though they did not send an army to help Saguntum), after its fall demanded the surrender of Hannibal, though under what terms it is difficult to see since Carthage had not violated the letter of the treaty. Thus began the Second Punic War, declared by Rome and conducted, on the Carthaginian side, almost entirely by Hannibal.

The march into Gaul. Hannibal spent the winter of 219–218 BC at Cartagena (Carthago Nova, the capital of Carthaginian Spain) in active preparations for carrying the war into Italy. Leaving his brother Hasdrubal in command of a considerable army for the defense of Spain and North Africa, he crossed the Ebro in April or May of 218 and marched into the Pyrenees (the Romans, shortly before they heard of this, decided on war). There his army—which consisted, according to Polybius, of 90,000 infantry, 12,000 cavalry (Polybius' figures are probably exaggerated; a total force of about 40,000 is more likely) and a number of elephants—met with stiff resistance from the Pyrenean tribes. This and the desertion of some of his Spanish troops greatly diminished his numbers, but he was able to reach the Rhône River with but little resistance from the tribes of southern Gaul.

Elusion of Scipio

Meanwhile, the Roman general Publius Cornelius Scipio transported his army, which had been detained in north Italy by a rebellion, by sea to Massilia (Marseille). As Scipio moved northward along the right bank of the Rhône, he learned that Hannibal had already crossed the river and was marching northward on the left bank. Now realizing that Hannibal probably planned to cross the Alps, Scipio returned to northern Italy to await his descent.

A great deal of controversy has surrounded the details of Hannibal's movements after the crossing of the Rhône. Polybius states that it was crossed while the river was still in one stream at a distance of four days' march from the sea. Fourques, opposite Arles, is thought by some calculations to be a likely place, but others indicate a crossing north of the confluence of the Isère and the Rhône. Hannibal used coracles and boats locally commandeered; for the elephants he made jetties out into the river and floated the elephants from these on earth-covered rafts. Horses were embarked on large boats or made to swim. During this operation hostile Gauls appeared on the opposite bank, and Hannibal dispatched a force under Hanno to cross farther upstream and attack the Gauls in the rear.

After this successful operation and after receiving a deputation of friendly Gallic leaders headed by those of the north Italian Boii, whose superior knowledge of the Alpine passes must have been of the greatest value to Hannibal's plans, the Carthaginians crossed the Durance River (or more probably an ancient branch of it that flowed into the Rhône near Avignon) and passed into an area that both Polybius and Livy call "the island." This area, the identification of which is the key to Hannibal's subsequent movements on land, was, according to Polybius, a fertile, densely populated triangle bounded on one side by hills, on the second by the Rhône, and on the third by a river that is probably either the Aygues or the Isère. Here on the "island" a civil war was being fought between two brothers (of what tribe it is not clear). Brancus, the elder of these, in return for Hannibal's help, provided new equipment for the Carthaginian army, which, after marching about 750 miles in four months from Cartagena, was in sore need of new supplies.

The Alpine crossing. Hannibal's army approached the Alps either by the Col de Grimone or the Col de Cabre, then through the basin of the Durance, or else by the Genèvre or Mont Cenis passes into the upper Po Valley, descending into the territory of the hostile Taurini. Hannibal's first action, therefore, was to storm their chief town, the modern Turin.

A number of details have been preserved about this great feat of military enterprise. In the initial days of the Alpine crossing, the danger came from the Allobroges, who attacked the rear of Hannibal's column. (Along the middle stages of the route, other Celtic groups attacked the baggage animals and rolled heavy stones down from the heights on the enfilade below, thus causing both men and animals to panic and lose their footings on the precipitous paths. Hannibal took countermeasures, but these involved him in heavy losses in men.) He was able on the third day to capture a Gallic town and to provide the army from its stores with rations for two or three days. Harassed by the daytime attentions of the Gauls from the heights and mistrusting the loyalty of his Gallic guides, Hannibal on the seventh day bivouacked on a large bare rock to cover the passage by night of his horses and pack animals in the gorge below. It was already October, and snow was falling on the summit of the pass, making the descent even more treacherous. Upon the hardened ice of the previous year's fall, the soldiers and animals alike slid and foundered in the fresh snow. On the 12th day a landslide blocked the narrow track, and the army was held up for one day while it was cleared. Finally on the 15th day, after a journey of five months from Cartagena, with 20,000 infantry, 6,000 cavalry, and only a few of the original 38 elephants intact, Hannibal descended into Italy, having surmounted the difficulties of climate and terrain, the guerrilla tactics of inaccessible tribes, and the major difficulty of commanding a body of men diverse in race and language under conditions to which they were ill fitted.

Hazards of the crossing

The war in Italy. Hannibal's forces were now totally inadequate to match against the proconsular army of Scipio, who had rushed to the Po River to protect the recently founded Roman colonies of Placentia (mod. Piacenzia) and Cremona. The first action between the two armies took place on the plains west of the Ticino River under conditions that turned the battle favourably for Hannibal's Numidian cavalry. Scipio was severely wounded, and the Romans withdrew to Placentia. After a number of manoeuvres that failed to lead to a second engagement, the combined armies of Sempronius Longus and Scipio met Hannibal on the left bank of the Trebia River south of Placentia and were soundly defeated (December 218 BC). This victory brought both Gauls and Ligurians to Hannibal's side, and his army was considerably augmented by Celtic recruits. After a severe winter (in which he contracted an eye infection), he was able to advance in the spring of 217 BC as far as the Arno River. Although two Roman armies were now in the field against him, he was able to outmanoeuvre that of Gaius Flaminius at Arretium and reached Faesulae (modern Fiesole) and Perugia. By design, this move forced Flaminius' army into open combat, and, as it passed between the northern shore of Lake Trasimene and the opposite hills, Hannibal's troops from their prepared positions all but annihilated it, killing thousands and driving others to drown in the lake. Reinforcements of about 4,000 cavalry under Gaius Centenius were intercepted before they arrived on the battlefield and were also destroyed.

Battle at Lake Trasimene

This was one of the greatest defeats suffered by Rome, but the Carthaginian troops were too worn to clinch their victories and march on Rome. Hannibal, furthermore, nurtured the vain hope that the Italian allies of Rome, whose attachments he always underestimated, would defect and cause civil war.

The summer of 217 was spent resting at Picenum, but toward the autumn and during the winter he ravaged Apulia and Campania; during this time the delaying tactics of the army under Quintus Fabius Maximus Cunctator prevented anything but skirmishes between the two armies. Suddenly in early summer of 216 Hannibal moved southward and seized the large army supply depot at Cannae on the Aufidus River. There early in August

Battle of Cannae

the Battle of Cannae (mod. Monte di Canne) was fought. While the Gauls and Iberian infantry of Hannibal's centre line yielded (without breaking) before the drive of the numerically superior Roman infantry, the Libyan infantry and cavalry of Hannibal's flanks stood fast, overlapped the Roman line, and in a rear encircling movement turned to pursue the victorious legionaries in the rear.

This third of Hannibal's great land victories in Italy brought the desired effect: many regions began to defect from the Italic confederacy. But, even with this victory behind him, Hannibal did not march on Rome but spent the winter of 216–215 in Capua. Gradually the Carthaginian fighting strength weakened. The strategy suggested by Fabius was put into operation: to defend the cities loyal to Rome; to try to recover, where opportunity offered, those cities that had fallen to Hannibal; never to enter battle when the enemy offered it but rather to keep the Carthaginians alert in every theatre of war. Thus Hannibal, unable because of inferior numbers to spread his forces to match the Romans and unable to employ this concentrated strength in a decisive battle, passed from the offensive to a cautious and not always successful defensive in Italy, inadequately supported by the home government at Carthage and, because of the Roman command of the sea, forced to obtain local provisions for protracted and ineffectual operations.

Between 215 and 213 Hannibal, except for the capture of Tarentum (modern Taranto), gained only minor victories; neither Philip V of Macedonia nor the Syracusans in Sicily (who had alike risen against Roman domination on the inspiration of the Carthaginian victories) were able to render direct aid. Reinforcements from Carthage were few. In 213 Casilinum and Arpi (captured by Hannibal in winter 216–215) were recovered by the Romans, and in 211 Hannibal was obliged to march to relieve the Roman siege of Capua. Despite Hannibal's countermeasure of a quick march to within three miles of the strongly fortified walls of Rome, Capua fell. In the same year (211), in Sicily, Syracuse fell, and by 209 Tarentum, in south Italy, had also been recaptured by the Romans.

The wars in Spain and Africa. Meanwhile, after years of indecisive warfare in Spain, the Roman successes there dealt severe blows to Carthaginian power in the peninsula. In 208 Hasdrubal, detaching a force from the main Carthaginian army, crossed the Alps (probably by his brother's route) to go to Hannibal's aid. Hasdrubal's army was defeated, however, at Metaurus in north Italy (207) before the Carthaginian armies could effect a junction. His last hope of making a recovery in central Italy thus dashed, Hannibal concentrated his forces in Bruttium, where with the help of his remaining allies he was able to resist Roman pressure for four more years.

Abandonment of Italy

Scipio, however, fresh from his victories in Spain, with the reluctant permission of the Senate, struck at North Africa, breaking Carthage's principal ally, the Massaesylian Numidians, and endangering Carthage. In order to go to the help of his country, Hannibal abandoned Italy in 203. Although a preliminary armistice had already been declared and the Carthaginian armies had accepted Scipio's severe terms (winter 204–203), Hannibal concentrated the remnants of the Carthaginian forces at Hadrumetum (modern Sousse, Tunisia). Almost at the very moment when the ambassadors were returning from Rome with the preliminary peace proposals, the Carthaginians violated the armistice.

The details of the campaigns that followed differed greatly. Both Hannibal and Scipio, in order to link up with their respective Numidian allies, moved up the Bagradas River to the region of Zama Regia. The actual site of the battle varies among sources. Hannibal was now deficient in cavalry; the mercenary troops of his front line and the African infantry of his second line together were routed, and Scipio, seeing that Hannibal's third line, the veteran soldiers, was still intact, reformed his front and brought up the Numidian cavalry of Masinissa, his Numidian ally, in the Carthaginian rear. Hannibal lost 20,000 men in defeat, but he himself escaped Masinissa's pursuit.

Exile and death. The treaty between Rome and Carthage that was concluded a year after the Battle of Zama frustrated the entire object of Hannibal's life, but his hopes of taking arms once more against Rome lived on. Although accused of having misconducted the war, he was made a suffete (a civil magistrate) in addition to retaining his military command, and in this office he was able to overthrow the power of the oligarchic governing faction at Carthage and bring about certain administrative and constitutional changes. He thus became unpopular with a certain faction of the Carthaginian nobility, and according to Livy he was denounced to the Romans as inciting Antiochus III of Syria to take up arms against the Romans. Before the arrival of the Roman envoy at Carthage, Hannibal fled for refuge to the court of Antiochus at Ephesus (195), where he was welcome at first, since Antiochus was preparing war with Rome. Soon, however, the presence of Hannibal and the sound advice he gave concerning the conduct of the war became a source of embarrassment, and he was sent to raise and command a fleet for Antiochus in the Phoenician cities. Inexperienced as he was in naval matters, he was defeated by the Roman fleet off Side, in Pamphylia. Antiochus was defeated on land at Magnesia in 190, and one of the terms demanded of him by the Romans was that Hannibal should be surrendered. Again accounts of Hannibal's subsequent actions vary; either he fled via Crete to the court of King Prusias of Bithynia, or he joined the rebel forces in Armenia. Eventually he took refuge with Prusias, who at this time was engaged in warfare with Rome's ally, King Eumenes II of Pergamum. He served Prusias in this war, and, in one of the victories he gained over Eumenes at sea, it is said that he threw cauldrons of snakes into the enemy vessels.

Flight from Carthage

Finally the Romans, by what means it is unknown, put themselves in a position to demand the surrender of Hannibal. Unable this time to escape arrest, Hannibal poisoned himself in the Bithynian village of Libyssa. The year is uncertain but was probably 183 BC, when Hannibal was 64 years old.

Death by poison

Hannibal as a general. In the conduct of the war with Rome, the project of the Alpine crossing probably originated with Hamilcar Barca. Strategically it was entirely successful and came to the Romans "as a thunderbolt," in the words of Florus. In men, however, it cost Hannibal dearly, between 5,000 and 10,000. Although it is said that he later regretted it, his action in not marching on Rome after Cannae is understandable. The city was strongly walled and supplied by sea and river; Hannibal had no siege craft. Hannibal hoped by a series of lightning victories to detach Rome's allies from the Italic confederacy. This was his greatest mistake: he had little realization of the strength of the ties of the confederates to Rome. Perhaps his greatest weakness, however, was at sea. He had no supporting navy and appeared indifferent to that Roman naval supremacy that was able in the first place to cut off reinforcements and in the second to bring about unimpeded the invasion of Africa. Although his tactics in the field, as attested even by Scipio, were brilliant, and he himself by his personal appearances and quick marches up and down Italy dazzled the Romans and complicated their strategy, he was at a decided disadvantage as regards reinforcements and provisions. Neither Philip V of Macedonia, with whom he made a treaty, nor the Syracusans were of material help. The winter of 216–215 spent idly at Capua and the refusal of the Romans thereafter to allow Hannibal to play the war as an army game provided the turning points in his success. Much that weighed against him was circumstantial and beyond his control; for all that went amiss in the conduct and policy of the war he cannot be entirely blamed; in tactics in the field he retained the mastery.

Personality. It is not to be expected that his Roman biographers would treat him impartially, but Polybius and Dio Cassius give the least biassed accounts. In spite of the charges of Hannibal's cruelty put forth by the Roman authors, he did enter into agreement with Fabius for the return of prisoners and treated with respect the

bodies of Tiberius Sempronius Gracchus (consul 215) and Lucius Aemilius Paulus (216), the fallen enemy generals. Of avarice, the other charge commonly laid against him, no direct evidence is found other than the practices necessary for a general to finance a war: indeed, he spared Fabius' farm. Much that was said against him (*e.g.*, cannibalism by Polybius) might be ascribed to individual activities of his generals, but even this is uncertain. His physical bravery is well attested, and his temperance and continence were praised. His power of leadership is implied in the lack of rioting and disharmony in that mixed body of men he commanded for so long, while the care he took for his elephants and horses as well as his men gives proof of a humane disposition. His treachery, that *punica fides* that the Romans detested, could from another point of view pass for resourcefulness in war and boldness in strategem, and, in assessing Roman judgments of him, the wide gulf between Roman and Carthaginian moral standards must be borne in mind. Of his wit and subtlety of speech many anecdotes remain. He spoke Greek and Latin fluently, but more personal information is absent from his biographies. He is shown in the only surviving portraits, the silver coins of Cartagena struck in 221, the year of his election as general, with a youthful, beardless, and pleasant face.

BIBLIOGRAPHY. Pronounced attitudes toward Hannibal have been taken largely by the German historians, varying from THEODOR MOMMSEN's lack of enthusiasm in *Römische Geschichte*, vol. 1 (1854; Eng. trans., *History of Rome*, vol. 1, book 3, 1911), to the warmth of J. KROMAYER in *Roms Kampf um die Weltherrschaft* (1912); and G. EGELHAAFS, *Hannibal, ein Charakterbild* (1922). EDUARD MEYER, *Untersuchungen zur Geschichte des zweiten punischen Kriegs* (1924), stressed his charismatic leadership, whereas E. PAIS, *Storia di Roma durante le guerre puniche* (1927), argued his mercenary motives on behalf of Carthaginian trade. A convenient general account of his tactics in Spain and Italy and the opposition by Fabius Cunctator and Scipio Africanus is in H.H. SCULLARD, *A History of the Roman World from 753 to 146 B.C.* (1935; paperback, 1969); and in B.H. WARMINGTON, *Carthage*, rev. ed., ch. 8–9 (1969), where there is also valuable discussion of Hannibal's relations with the government of Carthage. A penetrating study of Hannibal's personal history together with a treatment of his political aims in Italy (largely on the basis of the terms of his treaty with Philip V) and his relations with the democratic element at Carthage has been made by E. GROAG, *Hannibal als Politiker* (1929). Classical sources were unsure of Hannibal's route to Italy. F. WALBANK, *A Historical Commentary on Polybius*, vol. 1 (1957), summarizes both the textual and topographical criticism. The two fullest English reconstructions of the Alpine crossing are C. TORR, *Hannibal Crosses the Alps* (1924), good on literary sources and a landmark in the debate; and G. DE BEER, *Alps and Elephants* (1955), a lively and practical approach not only to topography but also to the problems of elephant transport; but it is not entirely in agreement with sources on crossing the Rhône near Arles. The author tabulates all opinions on the Alpine route. G. DE BEER, *Hannibal* (1969), collects for the general reader photographs of topography together with cultural background material on Rome and Carthage in Hannibal's time. For the Barcid coinage of Cartagena, see H.H. SCULLARD, "Hannibal's Elephants," *Numismatic Chronicle*, 8:158–168 (1948); also W. GOWERS and H.H. SCULLARD, "Hannibal's Elephants Again," *ibid.*, vol. 10 (1950); and E.S.G. ROBINSON, "Punic Coins of Spain and Their Bearing on the Roman Republican Series," in R.A.G. CARSON and C.H.V. SUTHERLAND (eds.). *Essays in Roman Coinage Presented to Harold Mattingly* (1956).

(W.Cu.)

Hanoi

The capital of the Democratic Republic of Vietnam (see VIETNAM, NORTH), Hanoi is a deceptive city. Superficially, it retains even now the atmosphere and appearance of a sleepy, provincial French colonial capital, with wide tree-lined boulevards, the villas of former colonial administrators, and bicycles that move slowly and silently along the streets. But its quietude belies the dynamism that had characterized the transformation of the city's colonial economy to that of a socialist economy after the establishment of the country's independence in 1954. The name Hanoi, of comparatively recent origin, was given to the original city in 1831 by the French, who named it after Hanoi Province which surrounded the city. "Hanoi" is derived from the Chinese for "inside the river"; *i.e.*, the land lying within a loop of a river. The city stands on the right bank of the Red River; the city proper has a population of roughly 700,000 people, while greater Hanoi has an estimated population of about 1,000,000.

History. Hanoi has been an important location throughout much of Vietnamese history. During the second period of Chinese domination (AD 43–540), the Chinese established their capital near the site of present-day Hanoi. In 541 Li Bi, a local aristocrat, successfully defeated the Chinese and established the short-lived kingdom of Van Xuan with its capital at Long Bien (Hanoi). The Chinese regained authority in 547 and the site of the capital was shifted several times. It was not until 1010 that a Vietnamese emperor, Ly Thai To, moved the capital back to Thang Long (Hanoi city). Despite attacks by the Mongols in the 13th century and Chinese incursions in the 15th, Hanoi remained the main capital of the various Vietnamese governments until the 17th century, when it was replaced by Hue as the principal capital. European travellers to the city in the late 17th and early 18th centuries recorded that it was an important city with a walled citadel and with streets of artisans and merchants outside the citadel walls.

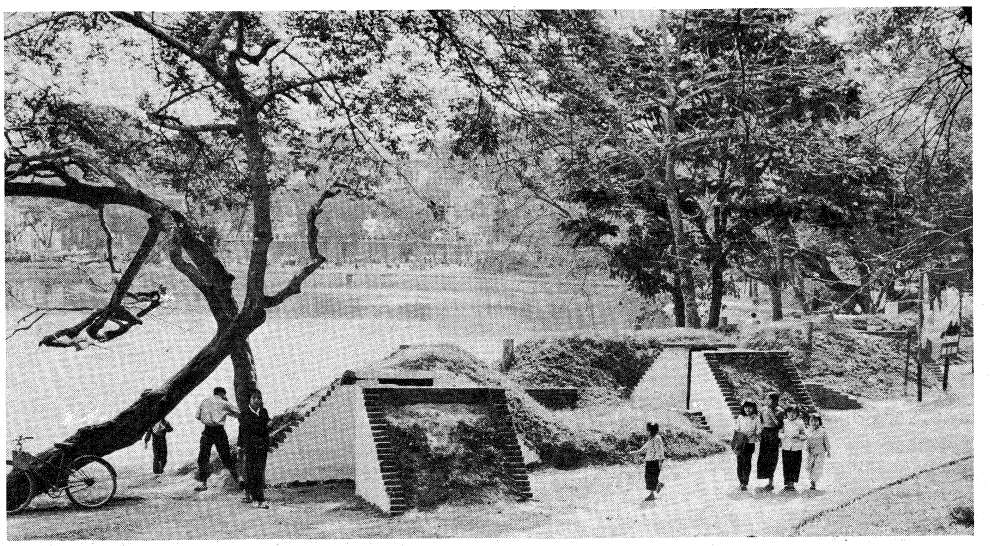
The French first occupied Hanoi in 1873; after a series of treaties with the Indochinese kingdom of Annam and with China, official French control began in 1883. The central government of French Indochina was finally organized in Hanoi in 1897. By 1902 Hanoi was considered to be the capital of French Indochina. The reason for Hanoi's emergence as the capital may be attributed to the greater importance which the French attached to Tonkin Province than to other provinces in Indochina; this was due to Tonkin's mineral resources and to French attempts to increase their influence in South China. In 1905 the population of Hanoi was recorded as 110,000—a figure which included about 103,000 Annamites, 2,000 Chinese, and 2,500 French. By 1936 the population had increased to 149,000, and the city had increased considerably in areal extent. To the city's administrative and commercial roles there was added an increasing transport function with the completion of the Hanoi–Saigon railway in 1936.

The capital of French Indochina

Hanoi remained the capital of the Indochinese territories during the Japanese occupation and the period of French–Japanese collaboration, which lasted from 1940 to 1945. In August 1945 the Indochinese Communist Party (the present Vietnam Worker's Party) seized power from the Japanese and launched a general insurrection. On September 2, 1945, the Democratic Republic of Vietnam was born when its president, Ho Chi Minh, read a declaration of independence in Hanoi's Ba Dinh Square. The French reasserted their control, however, and in 1949 Hue became capital of the short-lived State of Vietnam under a French-supported emperor, Bao Dai. Finally, on May 7, 1954, the French were defeated at Dien Bien Phu, and shortly thereafter Hanoi once again became the capital of the Democratic Republic of Vietnam.

The bombing by the U.S., which took place between 1965 and 1968 and again in 1972, caused some physical damage to Hanoi. The massive air raids of December 1972 were particularly destructive. During 1966 and 1967 one-man concrete bomb shelters were built at six-foot intervals along almost every street in the more populated areas.

The contemporary city. Hanoi is situated on the deltaic arm of the Red River, 75 miles (121 kilometres) upstream from the river's mouth. Much of the city is situated in the floodplain of the river, at an average elevation of 52 feet (16 metres) above sea level. Embankments, which are often as high as 45 to 50 feet (14 to 15 metres), protect the city from the floods that occur each summer. The city is bounded in the north by the Ho Tay (western) Lake; by a branch of the Red River, Song Hong, in the east; and by the Nhue River in the west; to the south the city merges into areas of agricultural land. A 1958 map of Hanoi, published by the North Vietnamese government, indicates that the city boundaries at that



Lake of the Returned Sword and park area in Hanoi.
Sovfoto

time encompassed most of the built-up areas, but by the early 1960s the boundaries of the city had expanded to include peripheral areas. Hanoi is linked with the left-bank suburb of Gia Lam, where railway-repair factories and the airport are located.

Climate. The city experiences a tropical monsoon climate (*i.e.*, a climate loosely characterized by seasonal rain-bearing winds). The coolest month is January, when the mean temperature is 63° F (17° C); the warmest is June, with a mean monthly temperature of 74° F (24° C). Hanoi experiences an average rainfall of some 70 inches (178 centimetres) a year; the rainy season falls between the months of May and September, and the driest months are from November to March. This dry period is often associated with a high degree of cloudiness, particularly in the months of February and March.

The city plan. The French left a mark on the physical appearance of Hanoi that 16 years of independence and war have not radically changed. With its tree-lined streets and many squares, parts of Hanoi resemble many French cities. The French were responsible for dividing the city into three main parts, or *quartiers*. The administrative quarter, located in the northern part of the city, stands on the site of the former Vietnamese citadel in the French period and contains most of the administrative buildings as well as the military barracks. Today it is still the main area for government buildings, focussing on historic Ba Dinh Square and the Ba Dinh Conference Hall; it is here that the main political functions are held. Extending southeastward along the arteries of broad streets are many of the major institutions, such as the university, the museum, the law courts, and the theatres. While some of the functions of these buildings have changed, the major educational and social institutions of the city are still located here. This was also the major residential zone of the French colonialists, who lived in stucco masonry villas. Since 1954 these villas have been taken over by the Vietnamese; they are often occupied by more than one family due to a shortage of housing.

To the north of this area lies the second quarter—a densely populated tenement zone in which the Vietnamese and Chinese were formerly the principal residents. A district of narrow, winding streets, it reflects the character of the city in its pre-industrial days.

The third quarter of the city grew up after 1928, south of the zone of French residential settlement. Vietnamese occupied this area almost exclusively, often living in housing resembling rural dwellings. Population densities in this quarter were lower than in the densely populated tenement zone to the north.

Since independence, housing for workers consisting of two- and three-story apartment houses have been built in some of the suburbs in areas such as Gia Lam; there has, however, been little transformation of the physical appearance of the city since the French colonial era, apart from the renaming of streets.

Transportation. Hanoi is an important communications centre. It has a railroad link with Haiphong, the major port (badly damaged by air raids), some 55 miles away. Railway lines also run northwest from Hanoi to Yünnanfu (K'unming, China), northeast via Lang Son to Nan-ning in the Chinese province of Kwangsi, and south to the demarcation line established in 1954 between North and South Vietnam. A road network radiates from Hanoi throughout the republic. Hanoi is served by an airport located at Gia Lam, which is linked to Hanoi by the Doumer Road and the one-mile-long rail bridge, which was completed in 1902. Internally, Hanoi has a system of streetcars, although most transportation is by foot and bicycle. The amount of car and truck traffic was reduced substantially between 1965 and 1972 during the bombing periods but is now increasing.

Road and rail networks

Demography. Reliable estimates of the population of Hanoi are difficult to obtain because it is often unclear whether these population statistics refer to the built-up area of Hanoi, the municipal area only, or to adjacent fringe areas as well. Most population figures suggest that the population of Hanoi city grew rapidly after independence, reaching 415,000 in 1960. This increased to approximately 600,000 by 1965, making Hanoi a crowded city with a density of 15,000 people per square kilometre (39,000 per square mile) in at least one area. The period of bombing, however, led to a considerable decrease in the city's population. From 1965 onward the compulsory evacuation of children and older people that took place may well have reduced the population by as much as one-third. Estimates of the city's population made in 1970, however, suggest that many people have now returned to the city and that the population, as mentioned, is now approximately 700,000.

Virtually the entire population is of Vietnamese origin, although there is a small Chinese community of largely Cantonese and Fukien origin.

Economic life. Since 1954 Hanoi has been transformed from a primarily commercial city into an important industrial centre. In 1954, when the French left, there were 12 large-scale manufacturing factories; by 1964, however, Hanoi had 88 industrial enterprises, about 750 cooperatives of artisans, and more than 1,000 handicraft cooperative cells employing over 88,000 workers. The larger factories were manufacturing such products as machine tools, electric generators and motors, plywood, textiles, chemicals, and matches. Despite relocation of industry, which took place as a result of the bombing, Hanoi is still estimated to contain one-third of North Vietnam's industry.

Industries

Commerce and retailing were considerably affected by U.S. bombing. Many shops were boarded up, and the state-run department stores were open for only two hours a day. The central market was vacated, and much of the selling had to be carried out from street markets. After the peace of 1973 the commerce of the city was reported to have returned to normal; many theatres reopened and outdoor recreation resumed.

Political and governmental institutions. Administratively, Hanoi is classified as a special municipality under direct central government control. As the capital of the Democratic Republic, it is the site of the central administrative offices of the republic.

Public utilities. The water supply of Hanoi is derived from deep wells. The system includes five treatment plants and a number of elevated storage tanks. Piped water is available in the built-up sections of the city and in the major city institutions and hospitals. Some areas of the city have a piped sewerage system, but there is also a collection system in operation. Electric power is provided by plants located in Hanoi.

Education. Hanoi contains several of the major educational institutions of the Democratic Republic. The University of Hanoi with 11 faculties was reported to have 20,000 students in 1963. The Hanoi Polytechnic College, financed in part by the Soviet Union, opened in 1956. There is a well-established system of primary and secondary schools.

Health. There are several large hospitals as well as the Medical and Pharmacological School of the University of Hanoi and an Institute of Traditional Medicine. The standard of medical treatment is generally high.

Cultural and recreational life. Hanoi is liberally provided with cultural and recreational amenities. There are over 34 libraries in the city; the National Library with over 800,000 books is the largest. There are three major museums: the Revolutionary Museum, the Historical Museum, and the Army Museum. In addition, there are many historical monuments, including the One Pillar Pagoda and the Tranvo Pagoda, both of which date from the 11th century. Popular leisure activities include visiting the numerous parks and gardens, the cinemas, and the theatres.

BIBLIOGRAPHY. There are no easily accessible book-length studies of Hanoi. The most comprehensive treatment of Hanoi in the period before Independence may be found in GEORGES AZAMBRE, "Hanoi: notes de géographie urbaine," *Bulletin de la Société des Études Indochinoises*, 30:355–365 (1955), which includes two maps of Hanoi in 1873 and 1954. Brief descriptions may also be found in PIERRE GOUROU, *L'Asie* (1953); CHARLES ROBEQUAIN, *L'Évolution économique de l'Indochine francaise* (1939; Eng. trans., *The Economic Development of French Indo-China*, 1944); and J. SION, "Asie des Moussons," *Géographie Universalle*, vol. 9 (1929). Information for the period since Independence is much more difficult to obtain, and usually takes the form of scattered references in more general texts. BERNARD B. FALL, *The Two Viet-Nams*, 2nd rev. ed. (1967); and P.J. HONEY (ed.), *North Vietnam Today* (1962), have some information on general features of Hanoi. LE-CHAU, *Le Viet Nam Socialiste: une économie de transition* (1966); and VO NHAN TRI, *Croissance économique de la République démocratique du Viet Nam (1945–1965)* (1967), have considerable information on economic developments in the city. HARVEY H. SMITH *et al.*, *Area Handbook for North Viet Nam* (1967), has useful information on public amenities and changes in the population of the city. More recent descriptions of the city may be found in journalists' reports occurring in periodicals such as *Le Monde diplomatique* (monthly); and *Far Eastern Economic Review* (weekly); and in information from official North Vietnamese sources.

(T.G.McG.)

Han Wu Ti

During his long reign, the self-willed and energetic Han emperor Wu Ti (in Pin-yin romanization, Han Wu Di) used and abused Imperial power in an unprecedented manner, raising to new heights the authority of the throne and extending Chinese influence abroad.

Han Wu Ti, named at birth Liu Ch'e, was born in 156 BC, probably the 11th son of Emperor Ching Ti, the fifth ruler of the Han dynasty. Not being the oldest son, he would normally not have ascended the throne, but relatives of the Emperor secured his designation as heir apparent at age seven. From his relatives and his teachers, the future Emperor absorbed influences from two basically antagonistic schools: the Taoists, inclined to the Legalist philosophy favouring an autocratic ruler guided by the rules of expediency, and the Confucianists, who sought through rituals and other means to check the growing power of the Han monarchs.

Emperor Wu Ti began his reign in 140 BC. During its early part he was under the moderating influence of relatives and court officials; however, by the late 130s he had decided that the essentially defensive foreign policy of his predecessors was not going to solve his foreign problems. In 133 he launched attacks on the nomadic Hsiung-nu, who constituted China's principal threat on the northern frontier, and thereafter he committed his realm to the expansion of the empire. By 101 his troops, spurred by an Emperor heedless of their hardships and intolerant of defeat, had extended Chinese control in all directions. South China and northern and central Vietnam were incorporated into the empire. North and central Korea, which had slipped from Chinese control in 128 BC, were reconquered and again administered by Imperial governors. Imperial troops were also sent across the Gobi desert in unsuccessful attempts to eliminate the Hsiung-nu threat. Han armies were farthest from home when they marched west to Fergana. The first expedition, in 104 BC, was a failure, but the Emperor refused to accept defeat. His intransigence stemmed from pride and his desire for horses. The horses Wu Ti wanted from Fergana were not principally intended for his war machine (although the Han armies suffered a chronic shortage of horses); rather, they were "blood sweating" horses (infected by a parasite causing skin hemorrhages), which for the Emperor had a mystical significance in that possession of them was considered a mark of Heaven's grace. The second expedition returned in 101 BC with some of the famous horses and the head of the ruler of Fergana; furthermore, the small states between China and Fergana had been humbled. Wu Ti had brought to submission all but the most distant parts of the world known to the Chinese.

Military campaigns

His wars and other undertakings exhausted the state's reserves and forced him to look for other sources of income. New taxes were decreed, and state monopolies on salt, iron, and wine were instituted. Yet by the latter part of his reign, his regime was in financial difficulties and confronted by popular unrest. The Emperor's economic controls were paralleled by his rigid control of the state apparatus. He created institutions for close supervision of the bureaucracy and drew into his personal service men who were outside the normal bureaucratic ranks and who made the bureaucracy more responsive to his will. He usually selected men whose behaviour was much like his own: harsh, demanding, and merciless.

Domestic policies

In spite of his aggressive policies, Emperor Wu Ti is also known for making Confucianism the state orthodoxy. Although he was unimpressed with the image of the ideal Confucian ruler as a benevolent father figure, he nevertheless appreciated the literary grace of the Confucianists and particularly the Confucian emphasis on ritual, which complemented his religious interests.

Most of the rituals performed by Emperor Wu Ti had a dual function; although of dynastic political and religious significance, they frequently manifested his ceaseless search for immortality. He rewarded richly men whom he believed could introduce him to immortals who would reveal their secrets to him. He sent men in search of the islands of the immortals and constructed elaborate palaces and towers designed to attract the spirits to him. At great expense he had conquered much of the world, and he invested heavily in the ardent hope that he would not have to leave it.

The last four years of Emperor Wu Ti's life were a time of retreat and regret. His empire could no longer afford an aggressive foreign policy, and he was forced to begin a period of retrenchment. The deeply suspicious Emperor suffered intense personal loss when, in 91 BC, his

Last years

heir apparent was falsely accused by an Imperial confidant of practicing witchcraft against the Emperor. In desperation, the son led an uprising in which thousands of people were killed and in which the heir committed suicide. Shortly before Emperor Wu Ti's death, he designated an eight-year-old son as heir apparent; then, anticipating his death, he had the youth's mother accused of a crime and imprisoned. Reportedly she "died of grief," but Emperor Wu Ti condoned her death, and perhaps caused it, to avoid having the young emperor dominated by relatives as he himself had been. He died on March 29, 87 BC.

Emperor Wu Ti is best remembered for his military conquests; hence, his posthumous title Wu, meaning "martial." His administrative reforms left an enduring mark on the Chinese state, and his exclusive recognition of Confucianism had a permanent effect on subsequent East Asian history.

BIBLIOGRAPHY. There is no Western language study devoted exclusively to the life of Emperor Wu or his reign. For a brief survey of his reign, see EDOUARD CHAVANNES (trans.), *Les Mémoires historiques de Se-ma Ts'ien*, vol. 1 (1895). The "Annals" of his reign have been technically translated with interpretive remarks by HOMER H. DUBS (Pan Ku's *History of the Former Han Dynasty*, vol. 2, 1944). The foreign policies of the Han rulers are surveyed in YING-SHIH YU, *Trade and Expansion in Han China: A Study in the Structure of Sino-Barbarian Relations* (1967). On Emperor Wu's economic policies, see NANCY LEE SWANN (Pan Ku, *Food and Money in Ancient China*, 1950), which is largely a highly technical translation.

(J.L.D.)

Harbin

Harbin (Ha-erh-pin; Ha-er-bin in Pin-yin romanization), a city in Northeast China, is located on the right bank of the Sungari River about 325 miles (520 kilometres) north-northeast of Mukden (Shen-yang), capital of the Chinese province (*sheng*) of Liaoning. Harbin has been the capital of Heilungkiang Province (Hei-lung-chiang *sheng*), the northernmost province of China, since 1954 and is the second largest city in the Northeast, historic homeland of the Manchus. It has a population of about 2,000,000.



Orthodox church in the Russian section of Harbin.
Emil Schulthess—Black Star

History. The city owes its origin to the construction of the Chinese Eastern Railway by the Russians at the end of the 19th and beginning of the 20th century. Before 1896 it was a minor fishing village and market town. Thereafter it became the construction centre for the railway, which by 1904 linked the Trans-Siberian Railroad from a point east of Lake Baikal in Siberia with the Russian port of Vladivostok on the Sea of Japan, and saved some 600 miles in the journey from Moscow over the route that ran entirely through Russian territory. Another railroad was built southward from Harbin to connect it with the Russian-developed city of Port Arthur (Lü-shun) on the Liaotung Peninsula in southern Manchuria. The early city was largely Russian built and in the early 1970s still retained many features of a pre-Revolutionary Russian city. Harbin was a base for Russian military operations in Manchuria during the Russo-Japanese War (1904–05). At the end of that war Harbin for a time came under joint Chinese–Japanese administration. After the Russian Revolution of 1917, it became a haven for refugees from Russia and for a time was the largest Russian city outside the Soviet Union.

Importance as a railroad centre

During the period of the Japanese-dominated state of Manchukuo, Harbin was known as Pinkiang and was the capital of the province of the same name. Soviet troops occupied the city in 1945, and a year later Chinese Communist forces took it over and from it directed their conquest of Manchuria.

In 1911 Harbin had a population of about 40,500; by 1931 it had some 332,000 people; by 1940, 662,000; in 1953, some 1,160,000; and in 1970 was estimated to have nearly 2,000,000. The number of Russians remaining in Harbin is believed to be very small.

The contemporary city. The climate of Harbin is similar to that of Winnipeg, Canada. The January mean temperature is −2° F (−19° C), and that for July is 72° F (22° C). Precipitation is about 21 inches (530 millimetres) annually, but most of it falls in summer, although winter snow is common.

From the first, Harbin acted as the regional capital of the then sparsely settled lowlands of the northern Manchurian (Northeast) Plain. By the early 1970s it was a major railway centre for five railway lines as well as a major port on the Sungari River, which is open to river navigation for six months during the summer season. Before 1917, when the Amur River was regarded as an international waterway, vessels from the Amur, into which the Sungari flows, reached as far as Harbin during the summer months; Harbin then functioned as an international port. A number of major roads also converge on the city.

Long a centre for food-processing industries, and also possessing power plants, railway shops, and boat-building facilities, the city since 1949 has become the chief industrial base of Northeast China. It is a major flour-milling centre. In addition to soybean-processing plants, sugar refineries (for sugar beets), leather works, soap factories, and tobacco factories, the city also has become a major machine-tool producer and the locus of such varied production as that of ball bearings, electric motors, boilers, wire and cable, film projectors, tractors, cement, electric power turbines and generators, plastics, and synthetic as well as traditional fibre products. Since 1966 the city has also been the major outfitting centre for the Ta-ch'ing oil fields, located about 80 miles west-northwest of it.

The city is supplied with electricity from both an older Japanese-built thermal plant and at least one new ther-

mal plant, with combined capacities of about 250,000 kilowatts. In addition, power comes via a 220-kilovolt line from a Japanese-built hydro installation at Feng-man on the Sungari River, 15 miles upstream from Kirin, in Kirin Province.

The surrounding agricultural region is extremely productive despite a short growing season. Spring wheat and soybeans are the principal crops, although sugar beets, flax, sunflowers, corn, and kaoliang (a grain sorghum) also are important. Since the region is a surplus producer, Harbin acts as a major shipping centre for agricultural and forest products which are being sent to the rest of China. A major airfield is located on the southern outskirts of the city.

The city hugs the right (south) bank of the Sungari, but winter port facilities are also located on the left (north) bank near the great railway bridge across the river. New urban development also has been taking place on the left bank. Like most foreign-developed cities in China, Harbin consists of a number of quarters. These include the port and industrial district on the right bank; the adjacent Tao-li commercial district (known in Russian as the Pristan district); Fu-chia-t'ien, the old Chinese residential quarter, an area of shops and wholesale and retailing activities; Man-chou-t'un (Man-chia-k'ou), a Japanese-built residential area; and the relatively newer areas of Harbin away from the river, which include the railway station and city administrative offices and which developed primarily after World War I. These areas have been expanding very rapidly since 1949 as the total population of the city has grown. Large blocks of flats appear to have been constructed in recent years on the outskirts of these areas. In the city's layout, a rectangular grid pattern generally prevails.

The city's districts

The site of the city is generally level to undulating, except near the river itself, where low bluffs lead down to the flood plain in places. Sandy strips of bank have been used as bathing beaches during the summer months. During the winter, skating and sledding on the river ice are popular sports. Bus services connect the various parts of the city, but automobiles and trucks are comparatively scarce; many work animals, including occasional Bactrian camels, may be seen, especially on the outskirts of the city. Theatres, libraries, hospitals, and schools are well distributed. Although the city continues to have a Russian air, many of the Russian-built or Russian-influenced buildings are being replaced with contemporary ferroconcrete structures.

Since large areas of underdeveloped land still exist in the northern region of the Northeast, and since considerable immigration to the region is expected from the rest of northern China, the growth of Harbin may be expected to continue at an accelerating rate for some time to come. (N.S.G.)

Harbours and Sea Works

The construction of harbours and sea works offers some of the most unusual problems and challenges in civil engineering. The continuous and immediate presence of the sea, nature's most restless, temperamental, and most powerful element, provides the engineer with an adversary certain to discover any weakness or fault in the structure built to resist it.

MARITIME ENGINEERING: METHODS, OBJECTIVES, AND ANCIENT WORKS

Objectives of maritime engineering. The principal objectives of such works fall broadly into two classifications: transportation, and reclamation and conservancy. Under the first fall works directed at providing facilities for the safe and economical transfer of cargo and passengers between land vehicles and ships; fishing ports for the landing and distribution of the harvest of the sea; harbours of refuge for ships and small craft; and marinas for the mooring or laying up of small private craft. Under the heading of reclamation and conservancy come works directed to the protection of the land area from encroachment by the sea, to the recovery and conversion to land use of areas occupied by the sea, and to the maintenance of river estuaries as efficient means for the discharge of inland runoff. In many places, without continuous attention to such maintenance, the coincidence of high tides with heavy rainfall would lead to frequent disastrous flooding of inhabited areas.

The civil-engineering techniques used for either of these objectives are broadly similar, and indeed the realization of both objectives at the same time will frequently be a feature of the same project. An operation of maintaining a river estuary at a depth sufficient for navigation, for example, may at the same time greatly improve its capacity for the drainage of upland floodwaters.

By courtesy of the British Transport Docks Board

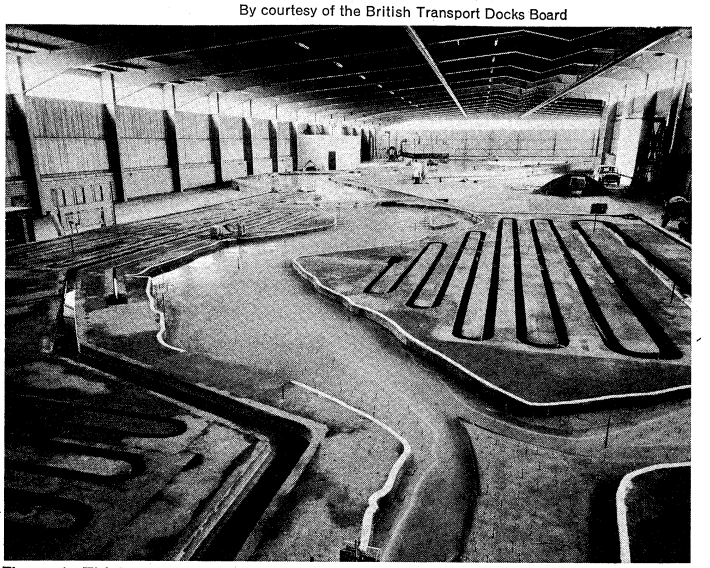


Figure 1: Tidal model of the Humber estuary in England.

Hydraulic models. The planning of maritime civil-engineering works, whether for transportation, reclamation, or conservancy, has been facilitated by the development of the technique of model studies. Once regarded as scientific toys, such studies are now considered an essential preliminary step to any large-scale redevelopment of a port or coastal area and are useful even for minor modifications or additions.

Scale models of the area, harbour, or estuary are made, so that water can be caused to flow in such a way as to reproduce the various tidal and other streams in the same direction and with equivalent velocities to those occurring on the site. A variety of devices, usually electronically controlled, has been developed to produce both wave and tidal effects.

The value of these experiments derives from the reduction in the time scale, which has been found to correspond to the reduction in the dimensional scales of the model. Thus, the large model of the Clyde Estuary of Scotland works on a tidal cycle of about 14 minutes or about 50 times the actual frequency. The effect of three years of tides following any modification of the profile of the harbour can thus be studied on the model in a matter of three weeks, and any tendency to otherwise unanticipated scour (clearing by powerful current) or siltation can probably be detected. The relative values of alternative positions of breakwaters in affording shelter can be similarly studied, using the wave-generating devices available; and the development of secondary, or reflected, waves with undesirable disturbances within the sheltered area may be anticipated and, if possible, forestalled.

Predicting long-range tide effects

Natural and artificial harbours. In certain favoured points on the world's coastlines, nature has provided harbours waiting only to be used, such as New York Bay, which the explorer Giovanni da Verrazano described as "a very agreeable location" for sheltering a ship. Such inlets, bays, and estuaries may require improvement by dredging and of course must be supplied with port structures, but basically they remain as nature made them,

and their existence accounts for many of the world's great cities. Because such natural harbours are not always at hand where port facilities are needed, engineers must create artificial harbours. The basic structure involved in the creation of an artificial harbour is a breakwater, sometimes called a jetty, or mole, the function of which is to provide calm water inshore. Locations for artificial harbours are of course chosen with an eye to the existing potential of the coast; an indentation, however slight, is favoured. Yet it has often been found justifiable on economic or strategic grounds to construct a complete harbour on a relatively unsheltered coastline by enclosing an area with breakwaters built from the shore, with openings of minimum width for entry and exit of ships.

Classical harbour works. Improvements to natural harbours and construction of artificial harbours were undertaken in very ancient times. There is no conclusive evidence for the date or locality of the first artificial harbour construction, but it is known that the Phoenecians built harbours at Sidon and Tyre in the 13th century BC.

The engineers of those days either knew or thought little about conservancy even as applied to the ports they constructed. Evidence is to be seen in the once thriving ports around the shores of the Mediterranean that now are not merely silent ruins but seem so far from even sight of the sea that it is difficult to imagine the presence of seagoing ships at the wharves, the alignment of which can occasionally be traced in the fertile alluvial land now occupying the site. Ephesus, Priene, and Miletus, on the Aegean shores of Asia Minor, are examples of this type of harbour disappearance, the destructive agent in each of these cases being the picturesque River Meander (now the Menderes), the efforts of which toward the creation of new land from the sea are readily perceivable from high ground adjacent to the river mouth. The formation of further bars can be seen to be preceding and, as there is, at the moment, no port in the vicinity whose livelihood can be threatened, it is interesting to speculate how far out to sea this process will ultimately continue in the course of the next Millennium or so.

Disappearance of ancient harbours

At Side, facing the island of Cyprus, the remains of an ancient breakwater, built to protect the anchorage, can still be seen, but the area enclosed between it and the advancing shoreline is now not a stone's throw wide. In this case, not only the river in the vicinity but littoral drift, a current that tends to parallel the coast, that produces and maintains extensive beaches to the east and the west, must be held partly responsible for the scale of siltation.

Of many of the ancient port structures no physical trace remains, but knowledge of the fact that they existed and even a measure of technical description has come down through the written word. With these descriptions and the monuments that still remain, some picture may be formed of the work undertaken by the maritime civil engineers of ancient times.

Given the frailty of the craft for which they were providing, shelter from the weather was the prime consideration; and much effort was devoted to the construction of breakwaters, moles, and similar enclosing structures. Cheap labour was abundant and the principal material used was natural stone. Surviving structures built in this way are likely to give an appearance of indestructibility, which occasionally attracts favourable comparison with the lighter, more rapidly depreciating modern structures. It is not, however, necessary to credit the engineers of antiquity with a conscious intention to build forever. Given the materials they had to use and the purposes they were implementing, they could do little else; moreover, because there was no rapid pace of advance in the development of ships or land transport, they were undisturbed by the shadow of obsolescence. In the 20th century, far from wanting to build forever, the port engineer has to be careful to avoid saddling posterity with structures that may long outlast their usefulness and turn into liabilities. The modern balance between excessive durability and dangerous frailty is one that the ancients never had to strike.

The durability of ancient sea works

Aided by the characteristics of the material they employed, the ancients constructed maritime works on a scale that is certainly remarkable to this day, which they emphasized by the addition of embellishment, such as statues and triumphal arches.

Interesting technical practices included the use by the Romans of the semicircular arch in constructing moles or breakwaters, an arrangement that allowed a measure of ingress and egress by the sea to produce a beneficial scouring action in the harbour. The Romans underpinned their structures with timber piling and frequently resorted to the construction of cofferdams (watertight enclosures) that they could de-water by the employment of Archimedean screws and waterwheels. This practice enabled them to carry out much of their foundation work in the dry; and the use of their famous hydraulic cement, pozzolana, gave their structures a durability far exceeding that afforded by the lime cement available to their predecessors.

Among the more interesting harbours of the ancient world are Alexandria, which had on the island of Pharos the first lighthouse in the world; Piraeus, the port of Athens; Ostia, the port of Rome; Syracuse; Carthage, destroyed and rebuilt by the Romans; Rhodes; and Tyre and Sidon, ports of the earliest important navigators, the Phoenicians.

BREAKWATERS

Because the function of breakwaters is to absorb or throw back as completely as possible the energy content of the maximum sea waves assailing the coast, they must be structures of considerable substance. The skill of the designer of a breakwater lies in achieving the minimum initial capital cost without incurring excessive future commitments for maintenance. Some degree of maintenance is of course unavoidable.

Breakwater design. A common breakwater design is based on an inner mound of small rocks or rubble, to provide the basic stability, with an outer covering of larger boulders, or armouring, to protect it from removal by the sea. The design of this outer armouring has fostered considerable ingenuity. The larger the blocks, the less likely they are to be disturbed, but the greater the cost of placing them in position and of restoring them after displacement by sea action. Probably the least satisfactory type of armour block, frequently used because of its relative ease of construction, is the simple concrete cubic, or rectangular, block. Even the densest concrete seldom weighs more than 60 percent of its weight in air when fully immersed in seawater; consequently, such blocks may have to be as much as 30 tons (27,000 kilograms) in weight to resist excessive movement.

Boulders of suitably dense natural rock are generally much more satisfactory and, in a project completed in the United Kingdom in the 1960s, it was found by experiment, and subsequently confirmed in experience, that armouring of this type could be composed of blocks of as little as six to eight tons (5,400 to 7,300 kilograms) to resist the action of waves up to 18 feet (five metres) in height. The same experiments showed that to afford the same protection in the same circumstances, concrete blocks of 22 tons (20,000 kilograms) would have been necessary.

In such cases, an intermediate layer of smaller blocks or boulders is inserted between the armouring and the inner core to prevent the finer material in the core from being dragged out by sea action between the interstices of the armour—a process that leads to ultimate settlement and possible breaching by overtopping of the breakwater.

The increasing cost and frequent unavailability within economic distance of suitable natural rock has provoked considerable thought to the design of concrete armour units that can, by reason of their shape, overcome the disadvantages of the simple cubic, or rectangular, block. One of the most successful has been the tetrapod, a four-legged design, each leg projecting from the centre at an angle of 109½° from each of the other three. Legs are bulbous, or pear-shaped, with the slightly larger diameters at the outer end. These units have the property, when

Tetrapods

placed, of knitting into each other in such a way that the removal of a single unit without the displacement of several others is almost impossible, while the interstices between them act as an absorbent of wave energy. Weights substantially less than those needed for cubic blocks are adequate in the case of tetrapods in similar storm conditions. The tetrapods can be mass-produced adjacent to the site through the employment of re-usable steel forms.

It is usual to construct some form of roadway along the crest of a breakwater, even when this is not required for any other dockside purposes, to facilitate inspection and access for labour, materials, and equipment for damage repairs.

Solid breakwaters. In certain circumstances, particularly in parts of the world where clear water facilitates operations by divers, vertical breakwaters of solid concrete or masonry construction are sometimes employed. Some preparation of the seabed by the depositing and levelling of a rubble mound to receive the structure is necessary, but it is usual to keep the crest of such a mound sufficiently below the surface of the water to ensure its not becoming exposed to destructive action by breaking waves. Repulsion of the waves by vertical reflection rather than their absorption is the philosophy of protection in all such cases, but it is not possible to state categorically which arrangement produces the most economic structure.

This type of breakwater can be conveniently constructed through the use of prefabricated concrete caissons, built on shore and floated out, sunk into position on the prepared bed and filled either with concrete or, less frequently, simple rubble or rock filling. A historical example of this arrangement was the Mulberry Harbour, built by the Allies and floated into position for the invasion of Normandy in 1944. No previous preparation of the seabed was possible, and only partial filling of the caissons had been carried out when the progress of the war rendered further operations unnecessary. Nevertheless, the fact that several of the caissons remained in position basically undamaged for nearly a decade after the invasion on this notoriously stormy coast demonstrated the possibilities of the method.

Floating breakwaters. Because of the large quantities of material required and the consequent high cost of breakwaters of normal construction, the possibility of floating breakwaters has received considerable study. The lee of calm water to be found behind a large ship at anchor in the open sea illustrates the principle. The difficulty is that to resist being torn away in extremes of weather, the moorings for a floating breakwater must be very massive. They are therefore difficult to install and subject to such constant chafing and movement as to require substantial maintenance. Another problem arises, especially in areas of large tidal range. The unavoidable, indeed, essential, slack in the moorings may allow the breakwater to ride large waves, so that they pass underneath it carrying a considerable proportion of their energy into the area to be sheltered.

Despite the difficulties, floating breakwaters still held out promise in the early 1970s with such projected ones based on the concept of causing the waves to expend their energy at the line of defense by breaking on a large, floating horizontal platform.

Pneumatic breakwaters. Finally, the pneumatic, or diffusion, breakwater has been widely discussed. Experiment and limited experience have shown that a curtain of air bubbles blown up from the seabed through a row of perforated nozzles acts as a barrier to the movement of waves over the surface. The mechanics of the arrangement appear to be that the rising bubbles generate streams flowing on the surface, outward in both directions, and the flow meeting the oncoming waves can be made sufficient to hold them up. There is reason to believe that jets of water would be almost as effective as air. Although the volume of air or water necessary completely to restrain the waves generated in severe weather over a wide front would require installation of a plant of uneconomical size, the device can be useful for the temporary protection of a short length of shore to allow the execution of specific works. The air or water pipes can be laid on the seabed at the perimeter of the area to be protected and fed from a mobile plant on shore, and the whole body of equipment can be removed after the operations have been completed.

DOCKS AND QUAYS

Because the principal operation to which harbour works are dedicated is transfer of goods from one transportation form to another (*e.g.*, from ships to trucks), it follows that docks, wharves, and quays are the most important assets of a port.

Facilities required in a dock

Ships must lie afloat, in complete shelter, within reach of mechanical devices for discharging their cargoes. Although in emergencies ships have been beached for unloading purposes, modern vessels, particularly the larger ones, can rarely afford contact with the seabed without risking serious structural strain. The implications of cargo handling, as far as civil-engineering works are concerned, do not differ much whether the loading and discharge are effected by shore-based cranes or by the ship's own equipment. In either case, large areas of firm, dry land immediately alongside the ship are required; the engineer must find a way to support this land, plus any superimposed loading it may be required to carry, immediately adjacent to water deep enough to float the largest ship.

The capital cost of such works probably increases roughly in proportion to the cube of the deepest draft of ship capable of being accommodated; thus the economic challenge posed by the increase in the size of modern ships is considerable. The advent of containerization—the packaging of small units of cargo into a single larger one—has not fundamentally altered this problem except perhaps in time to reduce the number of separate individual berths required and to increase greatly the area of land associated with each berth. A figure of 20 acres (eight hectares) per berth is freely mentioned as a reasonable requirement. The problem of land support at the waterline remains the same.

Gravity walls. The solution initially favoured, and, indeed, predominant for many years, was that of the simple gravity retaining wall, capable of holding land and water apart, so to speak, through a combination of its own mass with the passive resistance of the ground forming the seabed immediately in front of it. Both to ensure adequate support without detrimental settlement of the wall, to insure its lateral stability, and to prevent problems of scour, it is necessary to carry the foundations of the wall below the seabed level, in some cases a considerable distance below. In earlier constructions, the only guide to this depth, in the planning stage, was previous knowledge of the ground and the acumen of the engineer in recognizing the characteristics of the ground when he saw it. Many projects were carried out in open excavation, using temporary cofferdams to keep out the sea. In particularly unfavourable or unstable soils, accidents caused by collapse of the excavation were not unknown.

In modern practice, no such project is initiated without exhaustive exploration of the soil conditions by means of borings and laboratory tests on the samples. Continuous monitoring of the soil conditions during construction is also considered essential. Even so, accidents caused by soil instability still occasionally occur.

Materials for gravity walls

The material composing the walls is today almost universally concrete, plain or reinforced, according to the requirements of the design. This material has entirely superseded the heavy ashlar (natural rock) masonry at one time used for such construction, when the techniques for the large-scale production of concrete were not so well developed as they are today.

In some circumstances, particularly those in which the water is reasonably clear or the design and soil conditions do not require very deep excavation into the seabed, the construction of quay walls is adopted by means of large blocks, sometimes of stone but generally of concrete, placed underwater by divers. The economics of this method of construction are influenced by the high cost

of skilled divers and by the cumbersome nature of diving equipment. The development of lightweight, self-contained equipment, which leaves the diver considerably more mobile, may relieve this problem.

Concrete monoliths. The risks and difficulties attendant on the construction of gravity walls have, in suitable conditions, been avoided through the use of concrete monoliths sunk to the required foundation depth, either from the existing ground surface or, where the natural surface slopes, from fill added and dredged from the front of the quay wall on completion. This technique amounts to the construction above the ground of quite large sections of the intended wall, usually about 50 feet (15 metres) square in plan, which are then caused to sink by the removal, through vertical shafts, of the underlying soil. Another lift of wall is then constructed on top of the section that has sunk, more soil is removed, and the process is repeated until the bottom has reached a foundation level appropriate to the required stability. Considerable skill is sometimes necessary in the sinking process to keep the monoliths (usually provided with a tapered-steel cutting edge to the lowest lift) sinking uniformly and not listing, an eventuality that can occur if any part of the periphery encounters material particularly difficult to penetrate. Differential loading of the high side and special measures to undercut the material composing the obstruction may be necessary.

The shafts through which the excavated material is removed are generally flooded throughout the operation simply from the intrusion of the groundwater; if necessary, this water can be expelled by the use of compressed air. The excavation of difficult material in detail and in the dry can then be undertaken. It is an operation of some delicacy because the flotation effect of the compressed air adds a further element of instability to the monolith, and a blow (sudden leakage of air) under the cutting edge may result in flooding of the working chamber. When the bottom edge of the monolith has reached the designed level, the excavation shafts are sealed by concrete plugs. The shafts themselves can then be filled, either with concrete or with dry filling to give the final wall the required mass for stability.

Success in this form of construction cannot be guaranteed. In the case of the Western Docks at Southampton, Hampshire, constructed between World War I and World War II, it was found impossible except at inordinate cost to get the monoliths to sink through the opposing strata to the depth required for stability as a retaining wall. It was therefore necessary to reduce the thrust involved in this function by cutting the retained material back to a natural slope and spanning the gap between the back of the monoliths and the top of this slope by means of a reinforced concrete relieving platform, supported along its other edge on reinforced concrete piles. This arrangement has served well enough as far as the quay wall itself is concerned, but the maintenance of the natural slope, stone-pitched as a protection against erosion, has been a continuing liability. In addition, the presence behind the quay of the relieving platform constitutes a formidable obstacle to further construction work; *e.g.*, warehouses or multistory transit sheds.

The problem of Southampton quay

Concrete caisson walls. In situations in which the depth from ground level to the final dredged bottom is not excessive and the material available for retention as reclamation is of good self-supporting qualities, quay walls can be constructed of precast concrete caissons floated into position and sunk onto a prepared bed in the same manner as that described for breakwaters. Care is taken to design caissons able to withstand the thrust of the retained material, which is carefully selected for the areas immediately behind the quay wall. The conditions suitable for this form of construction are generally typical of the Mediterranean, where the slightness of the tidal variation keeps the depth required to a minimum. An outstanding example of this kind of construction is the extension to the area of the Principality of Monaco, which is being increased by as much as 22 percent by reclamation retained by this technique. Similarly constructed installations for transportation and ship-repair purposes exist elsewhere in the Mediterranean, in parts of which the earthquake factor is an additional influence on the retaining-wall design.

In all cases of dock-wall construction by concrete monolith or caisson, it is the basic structure of the wall that is provided by these means; the final superstructure, above highest tide level, will depend for its detail on the requirements for dockside services, crane tracks, and other elements.

The piled jetty. The high cost, difficulties, and possible dangers of providing dock and quay walls of the kind just described have always encouraged a search for alternative solutions that would eliminate the need for operations on or below the seabed. Of these, the earliest and most obvious is the piled jetty—its piles can be driven from floating craft and the deck and superstructure added thereto, working wholly above water. In regions in which there is a large tidal range, it may sometimes be both advantageous and necessary to take the opportunity of extremely low tides to make attachments to the piles for bracing and stiffening purposes. With a reasonable programming of the work, this operation can usually be done without particular difficulty, assuming that the seabed is of a composition reasonably amenable to penetration by piles to a sufficient depth to secure the lateral stability of the structure. A hard rock is not suitable, although some of the more friable rocks can be pierced by steel piles.

Timber piles

Piles may be of timber, reinforced concrete, or steel. Timber is a popular choice if there is a large natural supply. Lateral stiffness and stability can be achieved by using a sufficiently close spacing of the piles in both directions and adequate rigid bracing between the tops, timber being a material readily amenable to the workmanship required. Its chief drawback is lack of durability, particularly in the area between wind and water, although a timber jetty with reasonable maintenance can often outlast normal operational obsolescence. There are examples of construction in which the piles are connected together by casting around the heads a reinforced concrete slab, its soffit (underside) just below lowest water level. By this means, the timber is kept continually submerged, a condition under which its durability is prolonged. On the other hand, in tropical or semitropical waters or waters accidentally kept warm by industrial effluents, the use of timber may be inhibited by the presence of marine borers. Timber jetties have a considerable advantage in the comparative ease with which repairs to accident damage or deterioration can be effected.

Reinforced concrete piled piers and jetties, soundly constructed, exhibit great durability. Attachment to the piles for bracing and similar purposes tends, however, to be more complicated than in the case of timber. This is a disadvantage that applies also to subsequent maintenance and repairs.

Sheet-piled quay. An extension of the piled jetty concept is a quay design based on steel sheetpiling, the design becoming increasingly popular with improvements in the detail and manufacture of the material. Steel sheetpiling consists in essence of a series of rolled trough sections with interlocking grooves, or guides, known as clutches, along each edge of the section. Each pile is engaged, clutch to clutch, with a pile previously driven and then driven itself as nearly as possible to the same depth. In this way a continuous, impervious membrane is inserted into the ground. In most designs the convexity of the trough sections is arranged to face alternately to one side and the other of the line along which the membrane is driven, so that a structure of considerable lateral stiffness is built up. At the same time, a measure of flexibility in the clutches allows some angular deviation so that a membrane curved in overall plan is obtainable, a feature of considerable convenience in developing the layout of a series of wharves or quays.

The development of steel sheetpiling over the years has largely been characterized by the increasing weight and stiffness of the sections becoming available from the rolling mills. In one design, the clutch is a separate unit from the main structural element, generally of broad-flanged

or universal beam section. In this case, the clutch unit appears in a profile of two grooves, or channels, back to back, each capable of embracing the flanges of adjacent beams, which are thus locked together in a continuous sheet, or membrane, of considerable strength. Each universal section is entered, when pitched for driving, into the clutch on the previously driven section and usually carries the clutch for the next section with it. In another design, made economically possible by the advances in the technique of automatic continuous welding, rolled universal beam sections are welded by one flange into the troughs, or pans, of conventional sheet piles, the composite construction producing a unit of unique strength and stiffness.

The development of steel sheetpiling has kept ahead of the development of hammers capable of driving it, probably because the stiffer the section is, the greater the length of pile that can be incorporated in a design. The combination of heavier section and greater length demands a greater proportion of the energy delivered by the hammer being unproductively absorbed in the temporary elastic compression of the pile, leaving less energy to drive the pile further into the ground. Simply increasing the amount of energy delivered, by using a heavier hammer or a higher drop, does not necessarily provide the solution; it may only result in damage to the head of the pile without achieving greater penetration. This difficulty has been in part overcome by the use of high-strength steel piles. Nevertheless, it is not unknown for a pile to appear to be going down with little or no head damage when it is, in fact, sustaining extensive damage below seabed level that gravely compromises its efficiency as a retaining quay wall. This situation, usually the buckling of a pile, can occur particularly where the material of the seabed contains boulders or similar obstacles to penetration.

The problem has obvious major implications for the construction of quay walls and has provoked much debate among engineers. The skill of the quay designer and the advice of the soil mechanics specialist both contribute to the satisfactory reconciliation of the various conflicting factors outlined in order to achieve the most effective and economical solution.

Protecting the quay

In the normal design of sheet-piled quay or wharf wall, the sheetpiling itself forms the quay face, although it is generally found advisable to protect the piles from the impact of ships berthing by timber fenders. Vertical timbers at intervals are generally used. Horizontal walings (wooden ridges) between these timbers can also be employed, but they have a disadvantage, particularly at small wharves and with ships having their own protective belting: on a rising tide the beltings become entangled with the walings, occasioning damage or even minor disaster.

The upper part of the sheetpiling, being laterally unsupported on the sea side, is generally anchored back to resist the thrust of the retained soil. This resistance is commonly effected by using tie rods secured to anchors buried in the retained soil itself to a depth that, for reasons of overall stability, is beyond the natural slope line of the soil. As often as not, these anchors are themselves composed of lengths of sheetpiling driven, if possible, below the retained soil into the strata beneath. The mild, or alloy, steel tie rods, coated and wrapped against corrosion, can be carried through the exposed sheetpiling of the quay wall with large retaining nuts on the outside or can be secured to welded attachments at the back of the piling. The latter practice is the more commonly favoured arrangement, largely on account of its more finished appearance. The sheet-piled quay just described is completed by casting a reinforced concrete cope beam to cover as well as contain the exposed heads of the sheetpiling.

The advantage of this type of quay wall is that the space behind the wall is not occupied, as in the case of the suspended-pile supported deck, by a monolithic, fully structural element, the arrangements of which can only be disturbed for subsequent modification of the services layout at some cost and usually by a potentially complicated design operation. As in the case of a gravity wall, the space can be filled with suitable material that can subsequently be treated, to all intents and purposes, as natural ground in which service ducts can be buried if required. This arrangement is often an advantage in the case of freshwater mains for fire fighting or watering ships because they can thus be protected from frost. Alternatively, it is possible to place concrete-lined service and cable trenches in this material, sometimes conveniently by the use of precast sections, because the ground loads imposed are seldom sufficient to give rise to serious settlement problems.

Structural reinforcement. Identifiable structural loading, arising, for example, from crane tracks, can be supported on reinforced concrete beams on piles driven through the filling to the strata beneath. Dockside railways, a decreasing requirement because of the transfer of much shore-to-ship delivery to road vehicles, need not necessarily have piled support because the loading from these can be spread to remain within the bearing capacity of the filling. Some settlement is bound to take place, and the need for compensating by packing up and relevelling of the track has the incidental disadvantage of breaking up the surfacing of the quay, which is almost always provided to facilitate quayside access by road vehicles.

Sheet-pile quay walls are readily applicable to sites at which only relatively shallow or medium-depth water alongside is needed. As the required depth increases, a sheet-pile section of sufficient strength and stiffness to hold the retained material without further assistance becomes impractical from the point of view of handling and driving. A solution increasingly favoured is the so-called Dutch quay. In this design, after the line of sheetpiling, using one of the heavier and stiffer sections, has been driven, the ground behind is excavated for a distance back determined by the natural slope of the material to be used as filling and taken down as far as possible to lowest water level. At this level, a reinforced concrete relieving platform is constructed up against the sheetpiling but with independent vertical support from bearing piles driven through the bottom of the excavation to an appropriate depth. Piles for crane tracks are driven at the same time as these; that is, before the construction of the relieving platform.

The Dutch quay

Filling material is returned above the relieving platform, and although the latter now prevents further pile driving in the area, the probability of this being required is remote, whereas the retained load against the sheetpiling is much reduced. The advantages of having filled material behind the sheetpiling for installing services remain. In addition, the relieving platform affords the sheetpiling considerable help in resisting horizontal blows from the impact of berthing ships, and to increase this resistance, some of the piles supporting the platform are often driven toward the quay face. Reinforced concrete counterforts between the platform and the sheeting can be an additional help.

Durability. A question mark that hung over the use of steel sheetpiling in salt water in its early years concerned its durability in potentially hostile conditions. Experience of the rate of corrosion, particularly at the waterline or within the tidal range, varied from one locality to another according to the state of the water and the effect of such factors as the degree of salinity and the presence of industrial effluents. Precoating of the pile with a protective film such as tar or a bituminous paint is only of transient value, requiring regular renewal, and is effective only down to lowest water level.

The inclusion in the composition of the sheet-pile steel of a very small percentage of copper was tried as a means of increasing its durability, but the effectiveness is doubtful.

Cathodic protection

The confirmation of the electrochemical basis of much of the corrosion affecting steel sheetpiling led to the development of cathodic protection, a process that has wide application in many other fields, especially shipbuilding. Electrolytic corrosion arises from the passage through the piling of electric currents, causing the pile, or part of it, to become the anode, or positive pole, in what amounts

to a galvanic cell, or battery. In such a cell, metal is normally removed from the anode and may reappear on the cathode, or negative pole, which remains unaffected. These currents in sheetpiling may arise from stray leakages from adjacent electrical installations or be generated within the pile itself by differences in the electrolytic conditions at differing levels on the pile.

Cathodic protection is a means whereby cathodic polarity is imposed upon the whole pile, and its operation as an anode (with consequent deterioration) is prevented. This can be done either by the supply from a suitable source—*e.g.*, a battery—of an electric current that will overcome and reverse the direction of that naturally generated or by connecting the piling at intervals to sacrificial anodes of an element—generally aluminum or magnesium—the atomic relationship of which to the steel in the piling is such as to generate such a current without external assistance. These anodes are buried in the surrounding ground and care must be taken to ensure full electrolytic continuity between them and the piling to complete the circuit. It is sometimes necessary, in order to ensure electrical continuity in the piling itself between the anode connections, to weld adjacent piles together after driving.

By whatever means cathodic protection is applied, a small liability for operational maintenance arises, either for the continuous supply of the imposed current or the periodic renewal of the sacrificial anodes. The considerably increased durability of the structure usually justifies this.

Enclosed docks. Whenever possible, commercial quays are built open to the tide range to provide maximum freedom for shipping. There are, however, some parts of the world in which the range between low water and high water is so great that the resulting variations in the level of the ship's decks and hatches impose unacceptable disabilities on the handling of cargo. In such circumstances the quay walls, the net clear height of which, disregarding depth of foundations, must span the distance from the lowest seabed level acceptable for navigation at low tide to an adequate freeboard for the coping of the quay wall above the level of the highest high tide, may become of such dimensions as to be uneconomic. This condition is equally applicable in cases in which only the berths themselves are made to be usable no matter what the stage of the tide.

Locks

The problem can be met by constructing the quays as enclosed docks in which the water level is kept constant and access to the tidal areas is by means of a lock or locks. An obvious condition for the success of such an arrangement is that the strata of the bed under the enclosed dock area be sufficiently impervious to preclude any significant loss of water through the bottom during low-tide conditions. In this way the tidal range, as a limit on the height of the quay walls, can be eliminated.

Apart from the fact that they have gates at each end, the structure of maritime navigation locks and the problems involved in their design are very similar to those of dry docks. Although, in normal usage, a lock is never completely dry, it is essential that it should be designed to be capable of withstanding the stresses imposed by this condition, so that it may be possible to de-water the lock completely for inspection and maintenance.

It is common practice to design enclosed docks so that the normal water level maintained is not below the highest likely high tide because the invasion of an enclosed dock by a high tide significantly above the normal water level can be disastrous.

Although enclosed docks are frequently of such an area that they can supply the lockage water lost when a ship passes through the lock without any drop in level that cannot be made up on the next high tide, it is normal to provide a measure of impounding capacity in the form of pumps for lifting additional water from outside into the dock. Such a provision is essential for situations in which it is required to keep the enclosed dock water level above the highest tide.

It has sometimes been possible to accommodate ships of larger draft than originally planned for in large, but relatively old, enclosed docks. This is done by installing impounding pumps for topping up the water level to give an increased depth.

Enclosed docks generally suffer the operational disadvantage of restricted times of entry and exit because they are subject to a fairly rigid tidal schedule. First of all, the lower the tide level outside, the greater the amount of water lost in the locking operation; and, second, it is seldom economically feasible to maintain full navigation depths at all states of the tide in the approach channel to the lock entrance. This situation is particularly the case in which enclosed docks are sited adjacent to and operating from a tidal river estuary. The tidal lock at Dunkirk, France, opening to allow the passage of the night channel ferry, which runs on a timetable, is an example of a tidal lock operated whatever the state of the tide.

If possible, the access locks are usually duplicated, lest an accident involving the gates or the structure of the lock put the whole dock area out of action. Stability calculations of the quay walls within an enclosed dock are important; in older installations such calculations may have been based on the continuing presence of water at the designed normal level, and in the event of a serious failure at the lock—resulting in a considerable drop in the water level—the stability of the quay walls could come into question.

ROLL-ON, ROLL-OFF FACILITIES

An enormous increase in the use of the roll-on, roll-off technique of loading and unloading developed in the late 1960s. The principle of embarking whole vehicles under their own power was not new. The report of Hannibal ferrying his elephants over the Rhône in the 3rd century BC might be regarded as the earliest example from which the vast amphibious operations of the invasion of Normandy in 1944 were descended. Since the 1960s, however, the spectacular increase in the use of road transport for heavy freight and the increase in handling charges at ports for the loading and discharge of cargo by conventional means have combined to provide the impetus for the rapid commercial development of the roll-on, roll-off technique. In addition, the tendency to assemble machinery at its place of manufacture in larger and larger units has encouraged the development of special transport vehicles, and the economies of moving load and vehicle together from origin to destination are valuable.

Berthing and access problems

The principal problem for the port engineer is the provision of the special berthing for the ferry vessels and the means of access for the vehicles from the shore to the ship's decks. Rail-car ferries, involving somewhat similar problems, have been known for some time, but because of the severer limits on gradients for such vehicles there has been a tendency to limit the operation of these services to terminals at places where the tidal range is inconsiderable. For the Dover–Dunkirk ferry, opened shortly before World War II, a special enclosed dock was constructed at Dover in which the water level could be kept constant for loading and unloading, while at Dunkirk the entire dock system is totally enclosed, accessible through sea locks.

Many of the new roll-on, roll-off terminals for road services are, by contrast, in tidal water; and where the tide range is of considerable extent, access bridges of considerable length are often needed to keep the change of gradient between low and high tide within acceptable limits. The change in the ship's trim between conditions of light loading and full loading creates yet another problem.

Disadvantages of pontoons

At first sight, the solution might appear to be to support the outer end of the link span on a float, or pontoon, so that it would automatically follow the rise and fall of the tide. Several disadvantages of structural detail arise, however, and the system is vulnerable to damage caused by the movement of the pontoon under adverse weather conditions. A means of adjustment in the height between the span and the supporting pontoon to accommodate changes in ship's trim is still required; and, therefore, the overall economies of a pontoon are less than might at first be imagined.

It is, thus, almost universal practice to support the outer end of the link span from an overhead structure, either through conventional wire-rope hoisting gear or by means of hydraulic rams. The level of the end of the span can thus be continually adjusted, either automatically or by manual control, to match changes in the level of the ship's deck, whether caused by the tide, by the trim of the ship, or by differences in deck levels between one ship and another. Maximum flexibility of access has become increasingly important with the appearance, on some services, of ships with two independent car decks, both of which must be equally accessible to the link span. This situation has sometimes been achieved by the use of two-decker link spans, which has the effect of keeping the length and, unless the span is intended to carry loads on both decks simultaneously, the weight of the span to a minimum.

The sudden proliferation of roll-on, roll-off services simultaneously has led to the rather unfortunate development: a number of the terminals have tended to be tailor-made to suit particular ships and to be unable to accept different ships without, in some cases, quite major structural alteration. This feature clearly reduces the otherwise great flexibility of this technique, and an International Commission to examine the question was appointed in 1970 by the Permanent International Association of Navigation Congresses.

Maximum advantage of roll-on, roll-off is gained in relatively short sea passages. On longer voyages, the idle-road vehicles make the economies questionable. This problem can be overcome to some extent by embarking only semitrailers and leaving the tractive units ashore; the practice has no effect on the terminal details.

BULK TERMINALS

The enormous increase in the marine transit of materials in bulk, with petroleum leading the way, has given rise to the development of special terminals for the loading and discharge of such materials. The principal factor influencing the design of these installations is the still-increasing size of the ships. A single example of the effect of this change on design limits will be sufficient. The "Queen" liners, long the world's largest ships, in conditions of maximum load never drew more than 42 feet (13 metres) of water. Supertankers, on the other hand, when fully loaded, draw up to 72 feet (22 metres). If these ships required berthing structures of the type provided for conventional cargo and passenger liners and if the formula relating the capital costs of such structures to the deepest draft were applied, the cost of building an appropriate berth for such a tanker would reach a figure over six times that of the "Queen Mary's" old berth. Fortunately, the high mobility of the cargo renders such drastic and expensive measures unnecessary. Heavy capacity access for individual shore-based vehicles to carry away the cargo is not required, nor does the provision of services for the relatively small crews who man these great ships present any problem. The berthing positions can therefore be sited well out from the shore in deep water, and the structure itself can be limited to that required to provide a small island with mooring devices.

Unloading oil from tankers

In the case of oil terminals, the link to shore can be a relatively light pier or jetty structure carrying the pipelines through which the cargo is pumped ashore, with a roadway for access by no more than average-sized road vehicles, which will probably be in small numbers or even only one at a time (see Figure 2). As the ship herself carries the pumping machinery for delivering the cargo ashore, heavy mechanical gear for cargo handling is not required.

In the case of bulk carriers bringing more solid commodities, such as iron ore, the problem is more complicated. Hoisting grabs for lifting the ore out of the holds are necessary, even though transit between ship and shore can still be effected by continuous conveyors, corresponding to pipelines. Heavier foundation work is probably necessary at the berthing point to carry this machinery, and, for this reason, ore terminals have not, up to now, sought sites as far out in deep water as the oil terminals. It seems unlikely that the size of ore carriers will reach anything like the dimensions already attained by supertankers.

The employment of piled structures to meet these requirements is almost universal, and a variety of techniques has been evolved for handling and sinking into the seabed the long heavy piles required. At the sites likely to be chosen, penetration by piles may not be easy, particularly in places where most of the reasonably accessible deepwater sites tend to be located on the rockier shores.

One problem that arises is that of shelter in adverse weather conditions. While the ships themselves are reasonably robust, the relatively fragile berthing structures might break up, setting the ship loose, possibly without power immediately available, threatening disaster. As the cost of building breakwaters to protect sites in the depth of water required is likely to be prohibitive, the search has been for natural shelter. In the British Isles, the sheltered creeks of the western shores, such as Milford Haven, Pembrokeshire, have become valuable. Milford Haven had known little shipping other than fishing fleets since the early 19th century, but in the early 1970s boasted four bulk oil terminals. Two supply refineries were built on the spot; the third pumps to a refinery 60 miles (100 kilometres) away.

By courtesy of the British Petroleum Co.

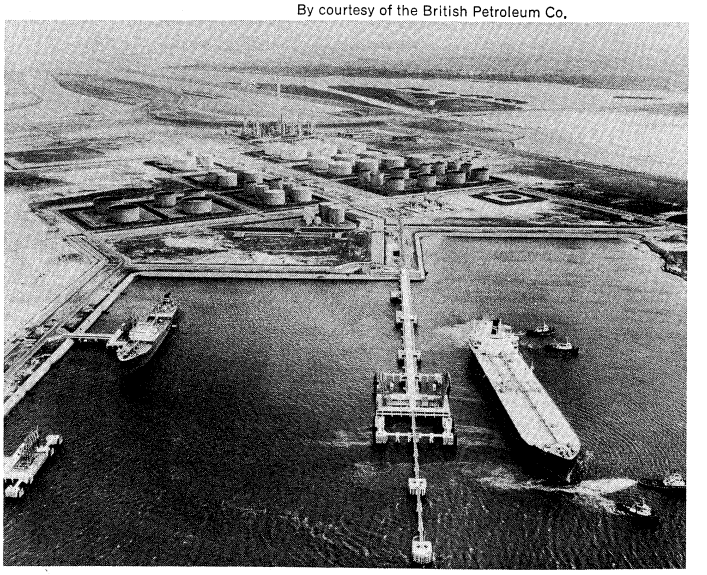


Figure 2: The oil jetty at Rotterdam, The Netherlands.

Another aspect of the terminals is the need for protection against the effects of unavoidable collision impacts. A slight impact from a vessel of these dimensions, by reason of the large kinetic energy of such a mass, can cause considerable damage to the light berthing structure. Much ingenuity and theoretical analysis have gone into devising fendering systems that will absorb this energy. Some use the displacement against gravity of large masses of material disposed pendulumwise in the berthing structure as the energy absorbent; others use the distortion by direct compression, shear, or torsion of heavy rubber shapes or sections; still others rely on the displacement of metal pistons against hydraulic or pneumatic pressure. The common feature of all the devices is that at least part of the energy absorbed is not dissipated but is used immediately to return the ship to its correct berthing position. This feature is not exhibited by the older forms of fenders, which relied on the compression and, in extreme cases, on the ultimate destruction of coiled rope or timber to absorb the impact. A major question still is the exact ship velocity to be allowed for, the determination of which is primarily an exercise in probability, balancing the economics of designing to a specified velocity against the cost of repairs after impacts at greater velocities. The key factor is the frequency of such impacts, which can only be determined by experi-

ence. In the early 1970s the development of suitable observing and recording equipment was under way.

DRY DOCKS

The largest single-purpose structure to be built by the maritime civil engineer is not directly connected with loading, unloading, or berthing but is indispensable to prolonging the life of ships. This is the dry dock, which permits giving necessary maintenance to the underwater parts of ships. The problem of dry-docking is aggravated by the tendency of ships to grow in size by increases in beam (width) and draft (depth below waterline) rather than in length, a process that rapidly renders many of the world's largest dry docks useless for servicing an increasing proportion of the traffic.

Danger of obsolescence

A classic example is the King George V Drydock at Southampton. Opened in 1933, it was 1,200 feet (370 metres) long and 135 feet (41 metres) wide and was capable of accommodating the largest vessels afloat, namely, the two Cunard liners "Queen Mary" and "Queen Elizabeth," each over 80,000 tons (73,000,000 kilograms) deadweight. The later supertankers have deadweight tonnages of 135,000 tons (122,000,000 kilograms) and more, within a length of about 1,150 feet (350 metres) but with a beam of about 175 feet (50 metres), which precludes them from entering the King George V dock. The lengthening of a dry dock would be a comparatively simple and economical operation; widening, on the other hand, involves at least the complete demolition of one sidewall and its rebuilding to give the increased clear width to the other wall, assuming space can be made available. Increasing the depth would mean a new dock altogether, but because tankers generally dry-dock in the unloaded condition in which their draft can be considerably less than that of a conventional ship, this problem has not so far been a practical one.

Structural requirements. Moreover, in a great many cases, the maximum state of stress in a dry dock occurs not when it is carrying the weight of the ship (always considerably less than the weight of the water occupying the dock when flooded) but when it is completely empty and subject to the pressures generated by water in the surrounding ground, particularly under the floor, the support of which may lie at a considerable depth below the level of the adjacent water table. To ensure against any tendency to lift under this pressure, the floor must either have sufficient weight in itself (one foot, or 300 millimetres, depth of concrete will resist a little less than 2½ feet, or 750 millimetres, head [depth] of water) or be designed as a structural element capable of transmitting this pressure laterally to the walls of the dry dock, which can then be designed to contribute the additional extra weight required. Obviously an operation involving both the complete rebuilding of one wall of a dry dock and the strengthening of the floor to cover an increase in its span as an inverted arch or beam is almost tantamount to the construction of a complete new dock.

This problem received somewhat tardy recognition, so that although several large new dry docks were built around the world in the 1960s, only a minority were capable of allowing the entry of tankers of more than 200,000 tons (180,000,000 kilograms), a considerable number of which were expected to be in commission by the middle of the 1970s.

Design. The design of a dry dock probably depends more on the ground conditions than any other engineering structure, with the possible exception of large dams. Mention has been made of the need in many cases to resist upward pressures under the floor. Apart from the simple solution of using the weight of the dock structure itself for this purpose, which is not economical, such devices have been tried as "pegging" the floor to the underlying strata by means of piles or pre-stressed anchors, or extending the floor slab itself beyond the sidewalls and so gaining assistance from the weight of the material filling behind the walls, which are designed to act as retaining walls to this filling. Venting of the floor to relieve water pressure can sometimes be of help provided the volume of water so released is not excessive. If it is, continuous pumping to keep the dock dry will be necessary. On sites in which water pressures do not have to be resisted, the design is generally simpler, and sufficient strength and stiffness to spread the loads from the ships' keels over the underlying ground so as not to exceed the bearing resistance of the latter is the controlling floor-design factor.

Dry docks in shipbuilding

The use of dry docks for the building rather than the maintenance of ships is a practice that has been increasingly adopted. Both the building and the launching of a ship in these circumstances can be considerably simplified. The designs of such dry docks are no different from those hitherto described; what is possibly the largest dry dock in the world was completed in Belfast, Northern Ireland, in 1970. This dock, built along the site of a former channel between two open basins, is capable of accommodating the three Cunard liners "Queen Mary," "Queen Elizabeth," and "Queen Elizabeth 2" simultaneously and is to be used for the building of large tankers. It is spanned by a crane of 400 tons (360,000 kilograms) lifting capacity to handle large prefabricated ship sections.

Entrances. Dry dock entrances are closed by gates of different designs, of which the sliding caisson and the flap gate, or box gate, are perhaps the most popular. The sliding caisson is usually housed in a recess, or camber, at the side of the entrance and can be drawn aside or hauled across with winch and wire rope gear to open and close the entrance. The flap gate is hinged horizontally across the entrance and lies on the bottom, when in the open position, to be hauled up into the vertical position to close the dock—a process occasionally facilitated by rendering the gate semibuoyant through the use of compressed air.

The ship type of caisson gate, a quite separate vessel, floated and sunk into its final position across the entrance, is largely out of favour. Although it was comparatively easy to remove for maintenance and had the further advantage that a spare caisson could be kept in reserve in case of damage, the tie-up of capital is usually found unnecessarily expensive merely as an insurance premium.

The maximum degree of watertightness obtainable between the gate and its seating is essential if continuing and expensive operational commitments for pumping out leakage water are to be avoided. The pressure of the water outside the gate is available to provide a powerful sealing force, but special treatment of the actual contact faces is necessary to make this force fully effective. For a long time it has been held that the only satisfactory arrangement was by the use of a timber lining (generally greenheart) around the contact face on the gate, bearing against stops in the dock structure composed of granite dressed and polished to a high degree of accuracy. The increased expense of such methods and the diminishing supply of skilled labour capable of dressing the granite has led to a search for alternatives. These include such devices as the use of stainless facing bars set in concrete, in place of the dressed granite, and rubber linings on the gates themselves. While these have generally proved effective when first installed, more experience is needed to determine their durability as compared with older methods.

Keel and bilge blocks. Keel and bilge blocks, on which the ship actually rests when dry-docked, are of a sufficient height above the floor of the dock to give reasonable access to the bottom plates. Such blocks are generally made of cast steel with renewable timber caps at the contact surfaces. Individual blocks can generally be dismantled under the ship to allow access to that part of the plates, if required, and can be reassembled to take their appropriate share of the weight after the operation required has been completed. Most modern ships, particularly tankers, are of nearly square section over a large part of their middle length and can be kept upright in dry dock by the support of the bilge blocks under their bilge keels. In the most up-to-date dry docks, the bilge blocks are provided with mechanical means for traversing them across the dock and altering their height by remote con-

trol while the dock is still flooded. This arrangement permits them to be adjusted in their correct position according to the shape of the ship while the latter is still just afloat but in contact with the centre-line keel blocks. The economic advantage of this arrangement is considerable because it allows one ship to be removed and another put into the dry dock on the same opening of the gate, whereas under previous practice it would have been necessary to close the dock and pump it out to reset the bilge blocks to the known profile of the next ship. Apart from the time needed, the power consumed in pumping out a large dry dock is a considerable factor.

As a consequence of the increasing number of ships suitable for bilge docking, the use of side shores to keep hulls upright in dry dock is a rapidly dying process, and indeed the altars provided for this purpose in dry docks of more old-fashioned design are often an embarrassment to the accommodation of a modern square-sectioned ship. Frequently this situation is remedied by cutting away some altars, an operation that must be conducted with discrimination because the removal of any quantity of material from the sidewalls may have a damaging effect on their stability.

Construction. *Basic technique.* Dry docks are usually constructed in open excavation in the dry, shutting out the sea by means of a cofferdam. Sometimes it is found convenient to construct the sidewalls first, in trench, and next to remove the loose material between them, then to lay the floor in stages so as not to endanger the stability of the walls before the floor is in position to give them toe support. Extensive pumping, to keep the excavations from filling with water during construction, is generally necessary.

Keeping out the sea

In one rather unusual case, a dry dock for 240,000-ton (128,000,000-kilogram) tankers was constructed almost wholly under water because large fissures in the rock running through to the sea flooded the site beyond the capacity of any reasonable assembly of pumping equipment. The entire space required for the structure was therefore excavated to formation level by dredging, and the sidewalls were constructed first, using prefabricated concrete caissons, sunk into place and filled with concrete. The spaces between adjacent caissons were sealed by filling with concrete in the same way. Stone aggregate, to a depth of seven metres (23 feet), was then deposited between these walls and consolidated into a concrete floor by a process of grouting in which colloidal cement grout was forced under pressure between the interstices of the aggregate, subsequently setting to form the whole into concrete. A similar process across the floor at the entrance incorporated a cofferdam of interlocking steel sheetpiling, which allowed the sill and gate hinge to be constructed in the dry. The gate, of the flap variety already mentioned, was floated and stepped into place by divers after the removal of the cofferdam. Only then was it possible to pump out the main body of the dock, which was completed by laying a reinforced concrete topping over the floor in order to provide a satisfactory working surface.

Floating dry docks. Floating dry docks have the initial advantage that they can be built and fully equipped in shipyard and factory conditions, in which their construction is not subject to unforeseen hazards arising from weather and variations in the ground conditions from those anticipated during design. The floating dock can be towed to the site, moored, and made ready for operation in a comparatively short time. Expenditure on temporary works, often a large fraction of the cost of a fixed dry dock, is also avoided.

Floating dry docks are usually fully self-contained (Figure 3). The sidewalls provide much of the residual buoyancy and stability required to keep the dock afloat when it has been so far submerged to allow the entry of a ship into the docking space over the main deck. Most of the machine tools and workshop equipment required for all the normal operations of ship repair and maintenance are also housed in the walls as well as the generating plant (usually diesel driven) to supply power for the operation of the dock and its equipment. Travelling cranes, for handling material off and onto the ship, run on the tops of the sidewalls.

A floating dry dock can be moved at relatively short notice to another site, should a long-term change in shipping-traffic patterns dictate a change. This advantage may be more apparent than real because the large work force required to man it may not be so readily transferable.

From A. Amirikian in "Transactions of the Society of Naval Architects and Marine Engineers," 1957

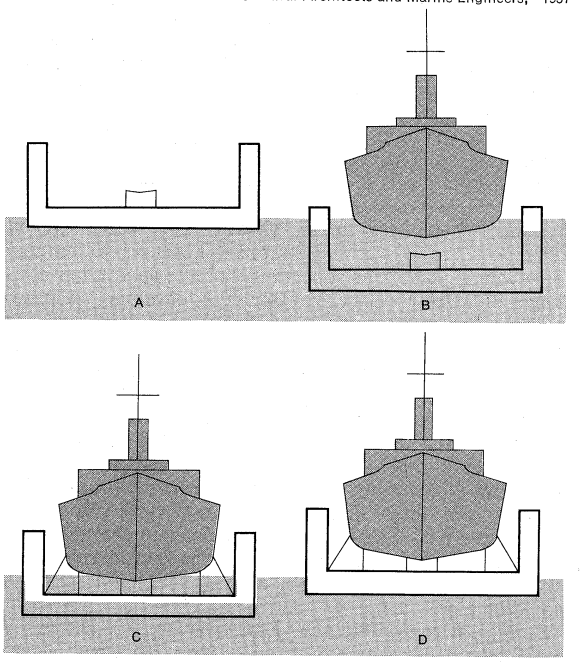


Figure 3: *Operation of a floating dry dock.*
(A) Dock afloat in its light-draft condition, with little or no water in its ballast tanks. (B) Dock submerged to its maximum draft by filling the ballast chambers; a vessel floats freely in the flooded channel. (C) Dock floor emerges from the water, and the vessel comes to rest on the support blocking. (D) All ballast is pumped out, raising the dock floor completely out of the water and providing a dry working area.

Moreover, floating dry docks tend to have large maintenance costs because the steel structure, being continually afloat, requires regular chipping and painting, as the hull of a ship does. The above-water structure presents no particular problem and can generally be given maintenance care without putting the dock out of use. The most vulnerable areas, those immediately adjacent to the waterline, can be reached by careening, a process that involves filling the water ballast tanks along one side to induce a list that lifts those on the other side part of the way out of the water. On completion, the process can be reversed for the other side.

Maintenance. Methods of underwater scaling and painting, or the use of limpet dams with which small areas can be covered with watertight enclosures inside of which men can work under compressed air, allow a limited measure of attention to be given to the bottom plating outside. Occasionally it is necessary to detach one of the sections of the dock, which is usually constructed in separate sections for this reason, and dry-docking it in the remainder, repeating the process until the whole dock has been renovated. This costly and tedious process is only resorted to for compelling reasons.

Underwater scaling and plating

To give a floating dock sufficient depth of water for submerging the docking blocks below the keel of the ship to be docked, it may be necessary to dredge a berth for it. In areas subject to heavy siltation, this dredged area will almost certainly act as a silt trap. Periodic removal of the dock from the berth to allow the latter to be redredged is an additional source of expenditure in such cases. Finally, in places where the tide range is of consequence, special mooring arrangements are necessary to restrain excessive lateral drift of the dock as the mooring chains become slack on low water.

The arrangement of keel and bilge blocks is generally similar to those described for fixed dry docks.

DREDGES, DREDGING, AND RECLAMATION

An indispensable item of equipment over a wide range of the maritime civil engineer's activities is the dredge with its ancillary units, such as hopper barges, tugs, reclamation units, and servicing craft. There are few navigable harbours or harbour approaches that do not require, at varying intervals of time, removal of deposits of unwanted material, the continuing accumulation of which can ultimately obstruct navigation. With the current trend toward larger ships, dredging is especially important.

Extensive research has been devoted to the development of dredging equipment. Through more sophisticated techniques, including, in some cases, permanent profile modification of the harbours and waterways, efforts are made to keep the need for dredging to a minimum. Model studies, mentioned earlier, can be of the greatest assistance.

Dredging. The material to be removed by dredging operations is usually derived from one of two sources or from a combination of both. In harbours at the mouths of rivers, quantities of silt are carried down in suspension and tend, partly because of the deceleration of the flow in the increased waterway available and partly because of the effects of increasing salinity, to be deposited at the mouth, usually the site of harbour works.

Estuarial silt

This process has produced areas of marked agricultural fertility, such as the Nile Delta in Egypt. While over a large time span the action is one of great benefit, in the short term it is generally a considerable inconvenience. The skillful employment of modern dredging equipment, however, has indicated possibilities of getting the best of both worlds. The other source of deposited material likely to obstruct navigation is littoral (coastal) drift, especially in areas where there is a sizable tidal range. The incoming tide frequently brings suspended material, some proportion of which settles to the bottom around the turn of the tide when the movement of water is at a minimum. In the absence of any countervailing tendency, an accumulation takes place, which again requires dredging.

For many years the workhorse of many of the world's dredging fleets has been the bucket-ladder dredge, operating a continually moving chain of open-ended shovels or scoops. At the bottom of the ladder the scoops are pushed into the face of the material, and empty themselves as they turn over at the top, the material falling into chutes that divert it into hopper barges for removal. A four-point mooring system enables the craft, and with it the bucket ladder, to be held up to the working face and, at the same time, swung sideways across it in either direction. By this means, an often remarkably level bed to the sea bottom can be closely controlled by adjusting the position of the ladder under the dredge's bottom. The positive action in filling the buckets enables such a dredge to tackle material of considerable stiffness, thereby extending its use to works of dredging and harbour development in which soils other than recently deposited silt or sand have to be excavated. Even some of the softer rocks can be removed in this way if the buckets are provided with hardened and stiffened edges and ripping teeth.

The principal disadvantage of the bucket-ladder dredge is the need for an elaborate system of fixed moorings. The area that can be covered by one placing of the moorings is limited. Continuous lifting and replacing of the moorings are not only time-consuming, but must be carried out in such a way as to offer minimum obstruction to navigation, a requirement that sometimes involves a great number of interruptions in dredging operations.

In areas in which the deposited silt is highly mobile and accumulates in considerable quantities, it can be economically removed by a suction dredge, which pumps water mixed with silt into open hoppers. By adjustment of the capacity of the hopper to the rate of flow from the pump, the water can be made to remain in the hopper long enough to deposit most of the silt. Careful design of the pumping machinery is required to assume a continuous mixture of maximum silt with minimum water.

The suction dredge

The first suction dredges generally operated from moored positions in the same way as bucket-ladder dredges, but a less elaborate system of moorings generally sufficed because the levelling of the seabed could be left to occur naturally through the mobility of the material. A marked advance was achieved by the elimination of much of the lifting and laying of moorings through the development of the trailer suction dredge. This craft has the capacity to dredge while on the move and cruises up and down the waterway or other area, sucking up silt as it goes. This operation does not eliminate all interference to navigation because a working trailer suction dredge moves more slowly than a ship under normal steerage way, but the obstruction is markedly less. The dredge's turn at the end of each sweep is usually facilitated by the incorporation of a bow side thrust propeller.

The growing tendency to use dredged material for reclamation purposes and the suitable condition for such purposes of the spoil as delivered by a suction dredge has encouraged its development. The seabeds and river bottoms in their natural state are often largely composed of relatively soft material and can be deepened by the use of suction dredges operating normally. Where rock or other hard material must be handled, conditions are favourable to the use of the suction-cutter dredge, which incorporates at the suction head a powerful rotating screw cutter that fragments the hard material. The increased dredging stresses arising from the use of a cutter require that a craft so equipped should be operated as a stationary dredge with moorings. Because such operations seldom take place in areas already under use by traffic, the obstruction problem is not often critical. Additionally, in modern equipment, the incorporation of heavy spud legs in the craft to anchor in the seabed reduces the number of separately laid moorings required.

A useful ancillary piece of equipment to all the above is the grab dredge, either self-propelled or towed to the site. Grab dredges are especially suitable for dredging close up to existing quay walls or other structures with minimum risk of damage, and the grab equipment is often capable of lifting individual boulders. Not infrequently grab dredges have value for maintenance dredging, particularly in restricted areas and with silt of sufficient mobility to level out the individual holes almost inevitably left. Although the return fall of the grab takes place with the bucket empty and is, to that extent, nonproductive, with skillful operators this element can be reduced to a minimum and, with some large craft operating four grabs simultaneously, considerable outputs can be achieved.

Dredges are characteristically designed to deliver their output either overside into attendant hopper barges or, in the case of self-propelled dredges, into hopper compartments incorporated in their own structure. These compartments are essential in the case of trailing suction dredges, but their value in other cases depends on the circumstances and the method of disposal of the spoil. When a long journey to the depositing area is involved, it is obviously more economical to leave the dredge continuously at work and remove the spoil in separate barges.

When the journey is short and the spoil is to be simply dumped, for which purpose the hoppers are provided with bottoms that fall open, then an economical work cycle between dredging area and spoiling ground, using one craft only, can frequently be established.

A special case is the side-boom dredge, which discharges straight back overside; by making the work coincide with an appropriate state of the tidal current, this arrangement secures the removal of the dredged silt by the tide's operation.

The side-boom dredge

Dredged spoil is less and less often disposed of by dumping out at sea, a practice once almost universal; instead it is used for the reclamation of land from the sea and foreshore. This process has been stimulated by the rise in the value of the land so created and by the dis-

covery that, in many instances, spoil taken out to sea frequently returns. This phenomenon has been investigated, both on hydraulic models and by mixing radioactive tracers with the dumped spoil in small quantities, permitting its subsequent movements to be followed with Geiger counters.

A variety of procedures have been developed for the combined operation of dredging and reclamation. Where the area to be dredged and the area to be reclaimed are in close proximity, as sometimes happens, the whole operation can be carried out by a single suction dredge pumping ashore through a floating pipeline. When, as is more often the case, there is a considerable distance between the two sites, transport in hopper barges is more economical. At the reclamation site, the barges can either be pumped out by a suction reclamation unit, or occasionally can dump their loads on the bottom; from there the material can be pumped ashore by the unit acting as a stationary suction dredge.

The layout of reclamation areas is a matter to which adequate scientific investigation should be devoted, covering such aspects as the adequacy and subsequent maintenance of any navigable waterways it is intended to provide through them, the design of the banks required to contain the pump spoil while the solids settle, and the relative positions of delivery and runoff points to obtain the maximum recovery of solid matter. Such schemes for reclamation, carried out in this way, can simultaneously ensure more valuable new land and improve navigation facilities.

The Delta Plan. It was noted at the beginning of this article that maritime engineering has two large objectives: improvement of transportation and reclamation and conservancy of land. Outstanding among examples of human ingenuity in the second category has been the long effort of the people of The Netherlands to keep their country, large areas of which are below sea level, habitable and productive.

The purpose of these efforts has generally been twofold, first to recover, reclaim, and retain more land for occupation; and second, to prevent the percolation of seawater into the water table of both the recovered and the original ground, which, if not prevented, would seriously reduce or even altogether destroy the value of the land for agricultural purposes. This second purpose has sometimes been described as "pushing back the salt line."

A prime example of the first purpose was the enclosure, by means of a dike some 28 kilometres (17 miles) in length, in 1926–32, of the large inlet formerly known as the Zuiderzee and, after its enclosure, renamed the Ijsselmeer. Considerable areas of this body of water have since been reclaimed by the pumping ashore of dredged sand, and the reclamation of further areas is either in hand or planned for the future. A large proportion of the area will, nevertheless, be maintained as a freshwater lake by the flow of the river Ijssel, which takes off from one of the outfalls of the Rhine, known as the Lek, or Neder Rhine, just south of Arnhem. In the 1960s it was found necessary to place a dam across the Lek just below the takeoff of the Ijssel, to divert an increased quantity of Rhine water down the Ijssel to the Ijsselmeer. The growth of shipping traffic on the canal, which connects Amsterdam with the North Sea, the locking operations of which necessarily discharge quantities of salt water into the Ijsselmeer, would otherwise tend to nullify the effects of the freshwater flow of the Ijssel.

By courtesy of The Netherlands Sluice and Tunnel Construction Co.

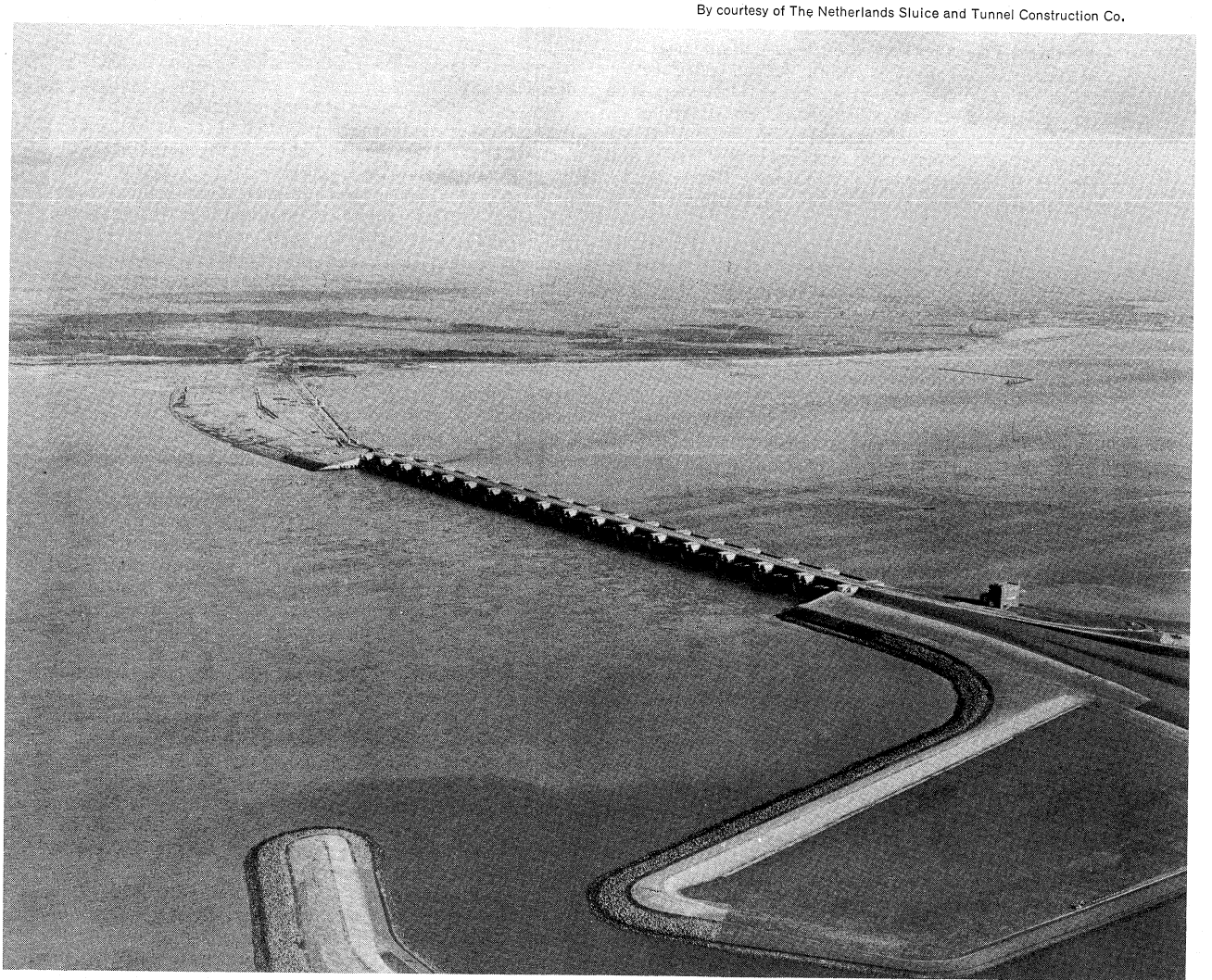


Figure 4: The giant sluices across the Haringvliet, part of the Dutch Delta Plan.

To maintain navigation in the Lek, in spite of the reduction in water flow, two further dams are provided downstream toward Rotterdam, and all three dams are capable of being opened, in the event of excessive floodwater coming down the Rhine.

Desalinization. The second purpose, that of desalinization or "pushing back the salt line," has been at the heart of the Delta Plan, one of the most imaginative civil-engineering projects ever undertaken. The incident that triggered the Delta Plan was the disastrous flooding of February 1, 1953, when the notorious North Sea surge brought tide levels higher than ever previously recorded, overtopping many of the existing dikes and causing untold damage and salt contamination of vast areas of agricultural land. The surge also caused considerable flooding and damage on the other side of the English Channel, along the east and southeast coasts of England. Statistical research suggests that tides of this level are to be expected at a frequency of at least once in 300 years.

The weak point in The Netherlands' defenses against flooding from the sea is the several deep inlets formed at the mouths of the Rhine and Maas rivers, through which the greater part of the water coming down these rivers discharges into the North Sea. Around the shores of these inlets run many miles of dikes, the maintenance of which is a constant burden and the strengthening and heightening of which to prevent a repetition of the disastrous 1953 floods represented a project of considerable magnitude.

It was considered that the most economical result would be obtained by a major operation of shutting out the sea, more or less at the main coastline, by a series of dams across the mouths of the inlets. By this means some 700 kilometres (435 miles) of dikes would be cut off from direct sea attack and reduced to a secondary function, whereas the total of the new dams that might still require a measure of maintenance is only 30 kilometres (19 miles). By conserving and controlling the vital flows from the Rhine and the Maas, the inlets themselves would be gradually transformed into freshwater lakes, thus greatly contributing to "pushing back the salt line."

A secondary effect in this direction will be an increase in the flow of freshwater toward Rotterdam, as a result of the raising of the levels in the estuarial inlets, particularly in the most northerly, the Haringvliet. This result should greatly assist desalinization in the Rotterdam area, where the penetration inland of the salt line had reached alarming proportions, as a result of the improvement in the navigational approaches to the port, effected by the construction of the channel known as the New Waterway from the Hook of Holland.

A further benefit to be gained from the structures of the Delta Plan is the great improvement in communications between the mainland and the hitherto somewhat isolated communities on the islands lying between the inlets; the new dams across the inlets will provide foundations for motor roads.

The Delta Plan construction was scheduled to take nearly a quarter of a century and the total cost represents a significant percentage of The Netherlands' national budget.

Although the authors of the plan stress that it is not properly a land-reclamation scheme (little or no extra land will be created by it), there is no doubt that many of the techniques developed for reclamation work are of the utmost value in carrying out the work, and, conversely, lessons learned in the course of the project will no doubt find useful application in future reclamation work the world over.

Thus, for the construction of the sluices through the dam across the Haringvliet, necessary to provide for escape of river water in times of flood, a working island was created in what was almost open sea, by the continuous depositing of sand on the seabed until the level rose above that of the water (Figure 4). Procedures for the rapid waterproofing of the banks so created have been brought to a high pitch of efficiency. This has been accomplished through the use of nylon carpets or asphalting by special high-speed placing machines.

The former take the place of the previously well-tried practice of using fascine mattresses weighted down with stones for which labour on the scale required to cover large areas with sufficient speed is no longer available.

The closure of the final gaps in the dams, a hazardous operation because of the large volume of water rushing through the narrow remaining gap at this stage, is effected at the Delta by the use of concrete caissons floated into the gap and scuttled in position. The technique has progressed there from the use of solid-walled caissons that had the disadvantage of closing the gap suddenly, with consequent hazard, to caissons incorporating their own sluices, thus allowing the flow of water to continue until all were in place and the sluices could be safely closed.

BIBLIOGRAPHY. The Nautical Charts and the Sailing Directions issued by the Hydrographic Office of the U.S. Navy and the Sailing Directions published by the Hydrographic Department of the Admiralty give full descriptions of the harbours of the world. For U.S. harbours, see the "Port Series" and the "Lake Series" published by the Corps of Engineers, U.S. Army, and the charts issued by the U.S. Coast and Geodetic Survey.

The Corps of Engineers, U.S. Army, publishes many manuals and articles concerning particular harbours and problems of harbour design.

The *Proceedings* of the Permanent International Association of Navigation Congresses (PIANC) contain much detailed information on harbours of the world, as well as technical studies on harbour problems.

Books on this subject include: F.M. DU-PLAT TAYLOR, *The Design, Construction and Maintenance of Docks, Wharves and Piers*, 3rd ed. rev. (1949); J.F. BRAHTZ (ed.), *Ocean Engineering: Goals, Environment, Technology* (1968); J. CLAPON, *Travaux maritimes*, 2 vol. (1966–67); C.M. TOWNSEND, *The Hydraulic Principles Governing River and Harbour Construction* (1922); and A.M. MUIR WOOD, *Coastal Hydraulics* (1969).

See also the monthly journal, *Dock and Harbour Authority*, which is devoted to problems of dock and harbour operation and construction.

(J.H.J.)

Hardenberg, Karl von

Karl August, Fürst von Hardenberg, Prussian statesman and administrator, was mainly responsible for preserving the integrity of the Prussian state during the Napoleonic Wars. Domestically he was able to continue the reforms introduced by Karl, Freiherr vom Stein; in foreign affairs he exchanged Prussia's alliance with France for an alliance with Russia in 1813, and in 1814–15 he represented Prussia at the peace negotiations in Paris and Vienna. He vainly fought for the establishment of a constitution but gained lasting fame for his domestic reforms—the liberalization of financial, economic, and agricultural policies—and for his conduct of foreign affairs, which created the political requisites for Prussia's liberation from French rule in 1813–15.

Early years

Hardenberg's father, Christian Ludwig, a member of an

By courtesy of the Staatsbibliothek, Berlin



Hardenberg, engraving by an unknown artist, 1820.

aristocratic family with estates in the southern part of the electorate of Hanover in Germany, was a general. Karl August was born on his mother's estate near Brunswick, on May 31, 1750, the oldest of seven children. During the Seven Years' War (1756–63), in which Austria and Prussia fought for supremacy in Germany, his mother lived with the children in the city of Hanover. While her husband was in the field, his oldest brother, a high Hanoverian civil servant, saw to the education of the children. As was the custom, Hardenberg was tutored at home in languages, history, and geography and attended a prestigious private school in Hanover for a year (1762–63). At the end of the war his father took over his ancestral estate, Hardenberg, about 80 miles (130 kilometres) south of Hanover, where Karl August acquired his much-admired skills in shooting, riding, dancing, and music.

To prepare himself for a career in public administration, Hardenberg enrolled at the University of Göttingen in the fall of 1766. In 1768 he spent a year at the University of Leipzig, where he made the acquaintance of a number of valuable and interesting people, among them Goethe, a contemporary and fellow student; F.A. von Heinitz, the head of the Elector of Saxony's mining industry; and Christian Fürchtegott Gellert, professor of philosophy and literature, an outstanding representative of the German-Protestant version of the French Enlightenment. He had brief contacts with Goethe throughout his life. Heinitz, later Prussian minister of mines, was not only a first-rate expert in mining and metallurgy but also a highly cultured and humane man who later helped both Hardenberg and Stein to enter the Prussian civil service. Hardenberg attended lectures on archaeology, history, literature, mathematics, the natural sciences, and economics. He also took lessons in drawing and music, but his main field was law, in which Göttingen provided the best instruction in Germany—often paving the way for an appointment in the imperial civil service or in that of one of the German states.

In 1770 Hardenberg left Göttingen and entered the Hanoverian ministry of justice. In order to advance his career he set out in the summer of 1772—on the advice of King George III of England, who was also elector of Hanover—on a year's travel throughout the whole of Germany, primarily to widen his political horizons. His voluminous diary, packed with information about the lands, institutions, and people he visited, shows his remarkable powers of observation and his capacity for independent judgment. He stayed in about a dozen capitals of ecclesiastical and secular princes in the middle Rhine region and southern Germany, studied the operation of the imperial administration in a few cities, and, after a visit to Vienna, began his homeward journey via Dresden and Berlin. In Nassau he stayed with the parents of Stein, and in Wetzlar he visited Goethe. His travels afforded him an insight into the political realities of Germany before 1789 and at the same time awakened in him a preference for the North German states. In 1773 he went to England to be presented to King George III, who appointed him Hanoverian councillor.

In 1774 Hardenberg married the 15-year-old Countess Juliane von Reventlow, who bore him a son and a daughter. A wealthy girl, she was, like himself, gay and extravagant, but their similarity of dispositions doomed their marriage and they were divorced in 1788. Because his career had come to a standstill and his wife had involved him in a scandal by her liaison with the Prince of Wales, Hardenberg left the Hanoverian service and entered that of the Duke of Brunswick. There, however, he proved to be unsuccessful as head of the department of education; moreover, his personal life became the subject of public gossip, for immediately after his divorce he had married Sophie von Lenthe, who had been divorced from her husband on Hardenberg's account. He therefore gladly accepted the post of Prussian provincial minister in Ansbach-Bayreuth that was offered him in 1790, a post in which he performed splendidly. He had the knack of selecting highly capable experts and attracting talented junior executives; among the former was the naturalist Alexander von Humboldt, who was in charge of the technical improvement of the mines. All in all Hardenberg made a model Prussian province out of the two former margravates.

Entry into the Prussian service

In 1793 Hardenberg described himself as follows:

> Endowed by nature with robust health and talents, a vivacious temperament that often gave free rein to my passions, enough material means to satisfy them but also fortunate to be by nature kind-hearted and humane, imbued from childhood with the desire to become great and good for the sake of mankind, early aware of my father's wealth and the wealth that awaited me, educated in the old-fashioned way to become a man of integrity and of learning—with these attributes I entered upon the great stage of the world.

When, in 1798, he won the abiding trust of King Frederick William III of Prussia, Berlin became this stage for him. He was entrusted with the most important administrative and diplomatic tasks, (*e.g.*, serving as foreign minister from 1804 to 1806). But, his nature being what it was, "enemies, women, and debts" hounded him throughout his life. In Ansbach his second marriage had come to grief when he took his mistress, a singer of bourgeois origin, into his household. She stayed with him for over 20 years, going with him to Berlin and later to his estate in the province of Brandenburg. He married her in 1807, six years after he had been divorced from his second wife; but shortly before his death he also separated from her. In order to rid himself of some of his chronic debts he sold his ancestral estate. But forced by his family to reinvest the money in land in order not to reduce the total family holdings, he bought the estates of Tempelberg near Berlin, which he developed into a handsome private residence with a splendid library. In 1814 gifts of land from the King enabled him to expand the estate into the demesne of Neuhardenberg, and the celebrated architect Karl Friedrich Schinkel rebuilt the residence into one of the most beautiful of the neoclassic manors in the province of Brandenburg.

Appointment as Premier

When Prussia lost the war of 1806–07 against France, Hardenberg, on Napoleon's orders, had to relinquish his premiership after only a few months in office. In 1810, however, when Prussia was faced with insolvency and could hardly maintain payments to France, he was the appointed Premier with far-reaching powers—this time with Napoleon's approval. Faithful to his motto, "democratic principles within a monarchic government," he continued the economic and governmental reforms that Stein had begun. He also granted full citizenship to the Jews. In foreign affairs he skillfully steered Prussia through the difficult years until 1813. Early in 1812 Prussia had had to sign a military alliance with France. After Napoleon's disastrous Russian campaign, Hardenberg preserved the appearance of the alliance while watching for a favourable moment for liberation. He advised the King to break away only when Prussia secured an alliance with Russia, which was finally concluded in 1813.

After Napoleon's fall in 1815, however, Hardenberg's power declined. He became entangled in Prussia's constitutional controversy; his advocacy of a constitution based on his essentially liberal beliefs eventually brought him into opposition to the plans of the King. Politically, Hardenberg's last years were marked by a mood of resignation, but even in old age he remained unusually well-read and maintained a many-sided interest in science and literature. Technical progress was as fascinating to him as the arts, of which he became a well-known patron. He helped artists and architects, took part in the reconstruction of Marienburg Castle near Danzig, and unstintingly provided the funds for restoring the Berlin theatre and opera house. When the first steamboat travelled down the Elbe River, Hardenberg was one of the passengers. On his 70th birthday Goethe honoured him with a poem. He died on November 26, 1822, in Genoa.

Assessment

As Hardenburg grew older, respect for his ideas increasingly declined in political circles. While patriots and reformers found him too accommodating and conciliatory, in the eyes of advocates of the return to absolutist

rule he was too liberal. By 1822 his great diplomatic achievements and notable domestic reforms of 1810–13 had been largely forgotten. Later in the 19th century the great German historian Leopold von Ranke was to emphasize Hardenberg's achievements as a statesman, pointing out that he had preserved the Prussian state when it was on the verge of destruction at Napoleon's hands. Since then Hardenberg has been primarily remembered in that role. While the social progress achieved by his reform legislation has always been acknowledged, it has only been truly appreciated in the 20th century. The unfavourable opinions of some of his important contemporaries, such as Stein and Humboldt, cast a shadow on his personality and diminished his stature when compared to the imposing figure of Stein. Only his more recent biographers have done justice to the unique and valuable aspects of his personality.

BIBLIOGRAPHY. K.A. VON HARDENBERG, *Denkwürdigkeiten des Staatskanzlers*, ed. by L. VON RANKE, 5 vol. (1877), contains an introduction by Ranke in vol. 1 and 4; H. HAUSSHERR, *Hardenberg: Eine politische Biographie*, ed. by K.E. BORN (1963–), vol. 1 of this excellent work is the best presentation of Hardenberg's youth and political career up to 1798; P.G. THIELEN, *Karl August von Hardenberg 1750–1822* (1967), the best complete biography in existence, including an outline of Hardenberg's private life using recently discovered sources; W.M. SIMON, *The Failure of the Prussian Reform Movement, 1807–1819* (1955), an attempt at a critical general assessment, with a somewhat biassed thesis; H. HAUSSHERR and W. BUSSMANN in *Neue Deutsche Biographie*, vol. 7, pp. 658–663 (1966), a brief biography with bibliography.

(E.W.Z.)

Hardy, Thomas

Thomas Hardy is the foremost—and perhaps the only important—English regional novelist. His most impressive novels are set in what he called "Wessex" (in fact, southwest England), chiefly in "South Wessex" (his native county of Dorset) and derive much of their strength from Hardy's intimate knowledge of the speech, customs, and way of life of people in that part of England. They are also a conscious attempt to translate some of the themes of Greek tragedy into terms of the English novel. As a poet—and Hardy attached more importance to his poetry than to his novels—he produced several poems of concentrated, intensely realized personal feeling that are perhaps his finest literary achievement, though these comprise a relatively small part of his poetic output.

EB Inc.



Hardy.

Early years. Hardy was born on June 2, 1840, at Upper Bockhampton, Stinsford ("Mellstock" in his stories), in Dorset. He was the son of Thomas Hardy, builder and master mason, and Jemima, born Hann. Both his parents were of long-established Dorset families. The writer was the eldest of four children. Though rather delicate in childhood, by the age of eight he was able to enter the new village school at Bockhampton. After one year he left to enter a Nonconformist school at Dorchester, and when the headmaster opened the Academy, a more ambitious school in the town, Hardy went with him. This master was a talented Latinist and aroused and encouraged in Hardy a love of classical writers that was to have a marked influence on his work. Leaving school in 1856, Hardy became a pupil of an architect and church-restorer named John Hicks, who practiced in Dorchester. At this time and for some years afterward Hardy read widely with a view to taking holy orders.

In 1862 he left Dorchester for London and soon obtained a post with a well-known architect, Arthur Blomfield (later Sir Arthur), as a "Gothic" draftsman, capable of designing and restoring churches and rectories. He remained with Blomfield until 1867, when a deterioration in health compelled him to return to Dorset, where he again worked for Hicks. During his years in London, Hardy had begun seriously to write poetry: some of the poems of this period—for example, "Neutral Tones"—are among his finest and most characteristic work; it is notable that there is no line of development in Hardy's poetry from immaturity to maturity, followed by a falling-off; at any period, his best and inferior work is found mixed, and his style undergoes no significant change.

First novel

During 1867–68 he wrote his first novel, *The Poor Man and the Lady*, which was rejected, though with some favourable comment, by two publishers, chiefly on the grounds that it was too satirical and Socialistic. The novelist George Meredith, who worked as reader to Chapman and Hall, one of the publishers, advised him to attempt a novel with a purely artistic purpose and a more complicated plot—advice that he followed too faithfully in his first published novel, *Desperate Remedies* (1871).

In March 1870, Hardy was asked by a Weymouth architect, for whom he was then working, to go to St. Juliot, near Boscastle, in Cornwall, in connection with the restoration of its church. There he met his future wife, Emma Lavinia Gifford, the rector's sister-in-law; this meeting and its setting were to be re-created over 40 years later in Hardy's most poignant poems, a group known as *Veteris Vestigiae Flammae* ("Vestiges of an Old Flame").

Desperate Remedies had been published anonymously and had met with a mixed reception. In 1872 Hardy returned to London and architectural work, having meanwhile written *Under the Greenwood Tree*, which was published in May of that year. Although slight compared with the later tragic novels, it has much humour and sympathetic observation, both in the rather burlesque account of the Mellstock choir and in a delicately handled courtship. The first of the Wessex novels, it shows already how closely connected these stories are with the rhythms of the rural year.

Hardy's next novel, *A Pair of Blue Eyes*, published serially (1872–73) in *Tinsley's Magazine*, owes its setting and certain minor details to St. Juliot and its surroundings. In 1873 he began *Far from the Madding Crowd*, first published serially (1874) and anonymously in Leslie Stephen's *Cornhill Magazine*. It was his first popular success, and he was encouraged by it to devote himself entirely to writing. It is also the first "typical" Hardy novel, for, although it has humour—not only in the treatment of the rustic characters—and what may pass for a happy ending, its scheme and general tone belong more to tragedy than to comedy.

First popular success

That Hardy was obliged to publish most of his novels in serial form may well account for his somewhat detached attitude toward them; it also explains their strong melodramatic content, almost inevitable if each installment was to conclude forcefully enough to sustain the reader's interest until the next installment appeared one month later. He wrote to Leslie Stephen, during the serial appearance of *Far from the Madding Crowd*, "I may have higher aims some day . . . but, for the present, circumstances lead me to wish merely to be considered a good hand at a serial." Toward the end of his life he said that his only ambition was to have some poem in a good anthology such as Francis Palgrave's *Golden Treasury*.

Hardy married Emma Gifford in August 1874. For the first years of their married life they lived at various addresses in London and Dorset; in 1885, they settled at Max Gate, a house that Hardy had had built on the outskirts of Dorchester, where he lived for the rest of his life.

Writings of maturity. From 1878 to 1895 is the period of Hardy's greatest achievement as a novelist. During this time he published *The Return of the Native*, *The Trumpet-Major*, *The Mayor of Casterbridge*, *The Woodlanders*, *Tess of the D'Urbervilles*, and *Jude the Obscure*. In these books Hardy's stoical pessimism—based on his conception of the "Immanent Will" (to be developed most plainly in his prose-poetry drama *The Dynasts*)—and his sense of the inevitable tragedy of human life are continually apparent.

Sometimes, by his use of disastrous coincidence, Hardy gives the impression of wrenching his material to fit his outlook (for instance, the events leading to the death of Mrs. Yeobright in *The Return of the Native* and the fate of Tess's letter of confession to Angel Clare); and sometimes the tragedy topples over into melodrama (the whole episode of little "Father Time" in *Jude the Obscure*). Yet there is real tragic dignity in the story of Tess, as well as of Henchard in *The Mayor of Casterbridge*. The latter, with the gradual, relentless revelation of Henchard's past, must be more than an unconscious recollection of Sophocles' *Oedipus*, while the function of the villagers or Casterbridge townsfolk in these books is clearly to provide a commentary akin to that of the chorus in Greek tragedy. But Hardy's intimate knowledge of the Wessex countryside and of rural life and speech (somewhat stylized though the latter may be) prevent the novels from being regarded only as attempts to transplant Greek tragedy into the 19th-century English countryside.

Conflict with the conventions of Victorian morality

With *Tess*, Hardy began to come into conflict with the conventions of Victorian morality. Certain scenes had been omitted from the serial publication, and others were altered; the subtitle, "A Pure Woman Faithfully Presented," aroused resentment. That a girl who had an illegitimate baby and who was eventually hanged for the murder of the man she was living with should be treated with compassion and understanding seemed an affront to accepted moral standards. *Jude the Obscure* aroused even greater indignation: here a married man and a married woman left their respective partners, lived together, and had children—and yet the author appeared to have sympathy for them. The Bishop of Wakefield announced that he had thrown the book in the fire, and he certainly got it withdrawn from circulation by one important firm of booksellers. Yet there was little to take exception to in the ethics of these novels: all too clearly in either instance, "the wages of sin is death." Hardy's crime had been to present the sinners as unhappy human beings, rather than as monsters of depravity.

The reception given to *Jude* so disgusted Hardy that he wrote no more novels, henceforth devoting his energies to poetry, which he had always regarded as far more important than his fiction. In 1898 *Wessex Poems*, including a good deal of poetry written earlier, was published; *Poems of the Past and the Present* followed in 1901, and several more volumes appeared during the remainder of his life.

From 1903 to 1908 appeared, in three installments, *The Dynasts*, a huge drama (unperformable and, indeed, not intended for performance) of the Napoleonic Wars, written mostly in blank verse; its lighter scenes are in prose, being chiefly concerned with the attitudes of the Wessex peasantry. Impressive though the endeavour is, the achievement is not altogether successful because of the pedestrian quality of much of the blank verse. It is in *The Dynasts* that Hardy's conception of the Immanent Will, implicit in the tragic novels, is most clearly stated. Put simply, this Will is an indifferent and unconscious force "that neither good nor evil knows" and is the motive power of the universe. The results of its impulses are almost invariably disastrous. In *The Dynasts* there is an implication that it may be growing into self-consciousness: perhaps the most striking expression of this concept is in "The Convergence of the Twain," a poem written on the sinking of the "Titanic" in 1912.

Hardy's conception of the Immanent Will

In 1910 Hardy was awarded the Order of Merit. On November 27, 1912, Mrs. Hardy died. Although the marriage does not seem to have been a very happy one, Hardy's grief was deep and found expression in the summit of his poetic achievement, the *Veteris Vestigiae Flammae* group mentioned above, which includes such moving poems as "The Voice" and "After a Journey."

In 1914 Hardy married his secretary, Florence Emily Dugdale, who survived him and wrote his biography. He continued to write poetry almost up to his death on January 11, 1928. His ashes were placed in Westminster Abbey, and his heart was buried, as he had wished, in the church at Stinsford, near his birthplace.

Hardy remains—owing in some measure to interest stimulated by cinema and television adaptations of his work—one of the most widely read Victorian novelists, and his work has been the occasion of a number of critical essays and monographs. What might have pleased him more is the general recognition of his genius and importance as a poet.

MAJOR WORKS

NOVELS: *Desperate Remedies* (1871); *Under the Greenwood Tree* (1872); *A Pair of Blue Eyes* (1873); *Far from the Madding Crowd* (1874); *The Hand of Ethelberta* (1876); *The Return of the Native* (1878); *The Trumpet-Major* (1880); *A Laodicean* (1881); *Two on a Tower* (1882); *The Mayor of Casterbridge* (1886); *The Woodlanders* (1887); *Tess of the D'Urbervilles* (1891); *Jude the Obscure* (1895, dated 1896); *The Well-Beloved* (1897); *Our Exploits at West Poley* (1952).

SHORT STORIES: *Wessex Tales: Strange, Lively and Commonplace* (1888), contained "The Distracted Young Preacher," "Fellow-Townsmen," "The Three Strangers," "Interlopers at the Knap," and "The Withered Arm," to which in the 1896 edition Hardy added "An Imaginative Woman"; *Life's Little Ironies: A Set of Tales with Some Colloquial Sketches Entitled a Few Crusted Characters* (1894), included "The Fiddler of the Reels," "On the Western Circuit," "To Please His Wife," and "Wessex People"—in 1927 Hardy transferred "An Imaginative Woman" to *Life's Little Ironies*, and added "The Melancholy Hussar of the German Legion" and "A Tradition of Eighteen Hundred and Four" to *Wessex Tales; A Group of Noble Dames* (1891), including "Barbara of the House of Grebe," "The First Countess of Wessex," "The Duchess of Hamptonshire," and "The Honourable Laura"; *A Changed Man, The Waiting Supper, and Other Tales* (1913), containing "The Romantic Adventures of a Milkmaid"; *An Indiscretion in the Life of an Heiress* (1934), a reworking as a long short story of Hardy's lost, unpublished first novel, *The Poor Man and the Lady* (written 1867).

POEMS: *Wessex Poems and Other Verses* (1898); *Poems of the Past and the Present* (1901, dated 1902); *Time's Laughingstocks and Other Verses* (1909); *Satires of Circumstances: Lyrics and Reveries with Miscellaneous Pieces* (1914), later combined with *Moments of Vision, and Miscellaneous Verses* (1917); *Late Lyrics and Earlier* (1922); *Human Shows, Far Phantasies: Songs and Trifles* (1925); *Winter Words in Various Moods and Metres* (1928).

POETIC EPIC DRAMA: *The Dynasts: A Drama of the Napoleonic Wars*, pt. 1–3 (1903–08, the three parts together, 1910).

BIBLIOGRAPHY. FLORENCE HARDY, *The Life of Thomas Hardy, 1840–1928* (1962), previously issued as two volumes, *The Early Life of Thomas Hardy, 1840–1891* and *The Later Years of Thomas Hardy, 1892–1928*, more an editing of Hardy's letters and diaries than a biography, is indispensable for Hardy's life, opinions, and sources of his work; LIONEL P. JOHNSON, *The Art of Thomas Hardy* (1894), one of the first attempts at an appreciation of Hardy's art and definition of his ideas; VIRGINIA WOOLF, "The Novels of Thomas Hardy," in *The Second Common Reader* (1932), a perceptive essay on the power and excitement of the novels; D.H. LAWRENCE, "A Study of Thomas Hardy," in *Selected Literary Criticism*, ed. by ANTHONY BEAL (1961), a brilliant, though highly subjective study of the psychology of Hardy's characters; WILLIAM R. RUTLAND, *Thomas Hardy: A Study of His Writings and Their Background* (1938), valuable for the background and origin of Hardy's beliefs; LORD DAVID CECIL, *Hardy, the Novelist* (1943), attempts to relate Hardy's fiction to Elizabethan drama and 18th-century fiction; DOUGLAS BROWN, *Thomas Hardy*, 2nd ed. (1961), perhaps the best critical account of Hardy; RICHARD L. PURDY, *Thomas Hardy: A Bibliographical Study* (1954), the definitive bibliography.

(F.Ch.)

Harmony

In music, harmony can be broadly defined as the sound of two or more notes heard simultaneously. In practice, this broad definition can also include some instances of notes sounded one after the other. If the consecutively sounded notes call to mind the notes of a familiar chord (a group of notes sounded together), the ear creates its own simultaneity in the same way that the eye perceives movement in a motion picture. In such cases the ear perceives the harmony that would result if the notes had sounded together. In a narrower sense, harmony refers to the extensively developed system of chords and relations between chords that characterizes Western music.

Musical sound may be regarded as being both horizontal and vertical. The horizontal aspects are those that proceed in time such as melody, counterpoint (or the interweaving of simultaneous melodies), and rhythm. The vertical aspect is the sum total of what is happening at any given moment: the result either of notes that sound against each other in counterpoint, or, as in the case of a melody and accompaniment, of the underpinning of chords that the composer gives the principal notes of his melody. In this analogy, harmony is primarily vertical. It also has a horizontal aspect, however, since the composer not only creates a harmonic sound at any given moment but joins these sounds in a succession of harmonies that give the music its distinctive personality.

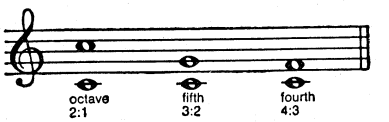
Non-harmonic music

Melody and rhythm can exist without harmony. By far the greatest part of the world's music is nonharmonic. Many highly sophisticated musical styles, such as those of India and China, consist basically of unharmonized melodic lines and their rhythmic organization. In only a few instances of folk and primitive music are simple chords specifically cultivated. Harmony in the Western sense is a fairly recent invention of limited geographic spread. It arose less than a millennium ago in the music of western Europe and is embraced today only in those musical cultures that trace their origins to that area.

This article will discuss the basic concepts of Western harmony, such as consonance and dissonance, the feelings of relative repose or tension sensed in chords; functional harmony, the theories regarding the relationships of specific chords to each other; keys, the sets of interrelated notes and chords used in harmony; modulation, the change of key; and tonality, the organization of notes, chords, and contrasting keys around a centrally important note, the tonic. This discussion will examine the development of Western harmony from its roots in the melodic music of antiquity and the Middle Ages, through the evolution of its classical form in the "common practice period" (the period of traditional Western harmony; *c.* 1650–*c.* 1900) and its disintegration in the 20th century. Finally, the article will discuss the relationship of harmony to chromaticism (use of notes not belonging to the key of a composition), dissonance, melody, and musical form, and the role of harmony in avant-garde music.

HARMONY AND TONALITY

The concept of harmony and harmonic relationships is not an arbitrary creation. It is based on certain relationships among musical tones that the human ear accepts almost reflexively, and that are also expressible through elementary scientific investigation. These relationships were first demonstrated by the Greek philosopher Pythagoras in the sixth century BC. In one of his most famous experiments a stretched string was divided by simple arithemetical ratios (½, ⅔, ¾, . . .) and plucked. By this means he demonstrated that the intervals, or distances between tones, that the string sounded before and after it was divided are the most fundamental intervals the ear perceives. These intervals, which occur in the music of nearly all cultures, either in melody or in harmony, are the octave, the fifth, and the fourth. (An octave, as from C to the C above it, encompasses eight white notes on a piano keyboard, or a comparable mixture of white and black notes. A fifth, as from C to G, encompasses five white notes; a fourth, as from C to F, four white notes.) In Pythagoras's experiment, for example, a string sounding C when cut in half sounds C, or the note an octave above it. In other words, a string divided in the ratio 1:2 yields the octave (C′) of its fundamental note (C). Likewise the ratio 2:3 (or the entire length to two thirds of its length) yields the fifth, and the ratio 3:4, the fourth. These notes—the fundamental and the notes a



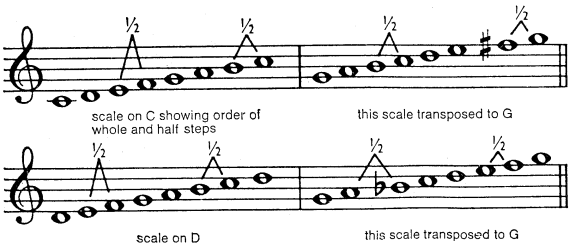
fourth, fifth, and octave above it—form the primary musical intervals, the cornerstones on which Western harmony is built.

The roots of harmony. The organized system of Western harmony as practiced from *c.* 1650 to *c.* 1900 evolved from earlier musical practices: from the polyphony—music in several voices, or parts—of the late Middle Ages and the Renaissance, and ultimately, from the strictly melodic music of the Middle Ages that gave rise to polyphony. The organization of medieval music, in turn, derives from the medieval theorists' fragmented knowledge of ancient Greek music.

Although the music of ancient Greece consisted entirely of melodies sung in unison or, in the case of voices of unequal range, at the octave, the term harmony occurs frequently in the writings on music at the time. Leading theorists such as Aristoxenus (fl. 4th century BC) provide a clear picture of a musical style consisting of a wide choice of "harmonies," and Plato and Aristotle discuss the ethical and moral value of one "harmony" over another.

Greek scales

In Greek music a "harmony" was the succession of tones within an octave—in modern usage, a scale. The Greek system embraced seven "harmonies," or scale-types, distinguished from one another by their particular order of succession of tones and semitones (*i.e.*, whole steps and half steps). These "harmonies" were later erroneously called modes, a broader term involving the characteristic contours of a melody, as well as the scale it used. The distinctions between the Greek scale-types can be approximated on the piano by playing a series of scales beginning on seven successive white keys and using *only* white keys (as from C to C′, from D to D′, etc. It will be seen that the scale on C begins with two whole tones (C–D, D–E) followed by a semitone (E–F), while that on D begins with one whole tone (D–E), a semitone (E–F) and a whole tone (F–G), etc. Because the particular shape of a Greek "harmony" depends on its arrangement of whole and half spaces, any harmony can be transposed; *i.e.*, duplicated at a higher or lower pitch, by using the black keys of the piano to provide the proper sequence of steps. When a melody is built using one of these "harmonies," or scale patterns, the arrangement of tones and semitones of the scale helps form the distinctive musical character of the melody.



In the Greek view, each of the "harmonies" embodied a distinctive ethos, or moral characteristic. Plato refers to the Lydian and Mixolydian scales (those that the foregoing experiment would produce beginning, respectively, on C and B) as "dirgelike," while the Iastian (beginning on G) is "too relaxed." Both he and Aristotle called for an emphasis on certain scales, those of an uplifting and heroic nature, in the education of young men. A faint echo of this complex ethical view persists in contemporary scale-forms, major and minor, in which the first is often thought of as "happy" and the second "sad."

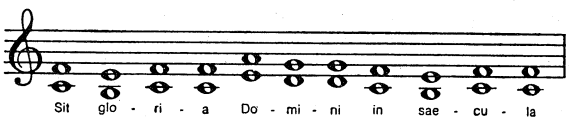
Medieval modes

Early Christian music demonstrates a similar diversity of scale patterns. As early as 560 AD, the post-Roman musician and monk Flavius Cassiodorus in his treatise "Institutiones musicae" discerned fifteen scale patterns normally used as a basis for the modes, or melody patterns, of medieval music. Cassiodorus also identifies three basic harmonies of notes sounded simultaneously: the fourth (notes in the 4:3 Pythagorean ratio), the fifth (3:2), and the octave (2:1), plus compounds of these intervals (*e.g.*, the octave plus a fifth, a ratio of 3:1). Thus, by the 6th century both scale patterns and the primary musical intervals were recognized.

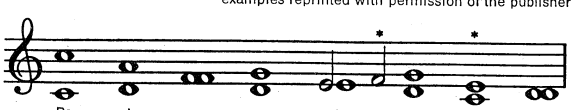
By the end of the eighth century, the Christian liturgy had been codified, and a large body of chants designated for specific religious occasions had been collected by a succession of monks beginning with Pope Gregory I, whose name the collection of Gregorian chant bears. In this body of chant six principal modes can be discerned using scale patterns beginning, respectively, on the notes D, E, F, G, A, and C. Each mode also had a variant form, yielding a total of 12 modes. Lydian, Mixolydian, and other Greek names were later applied to these modes, although they began on notes different from the Greek "harmonies." The modes consisted of more than mere scale patterns; in addition, one note of the mode was the proper final note for a melody; another was the "reciting note," characteristically acted as a point around which melodies centred. The modal melodies were sung unharmonized and in a rhythmically free manner.

Harmony before the common practice period. By the ninth century the practice arose in many churches of performing portions of chant melodies with an added, harmonizing voice—possibly as a means of greater emphasis, or of reinforcing the sound to carry through the larger churches that were being built at the time. This harmonizing technique, called organum, is the first true example of harmony. The first instances were extremely simple, consisting of adding a voice that exactly paralleled the original melody at the interval of a fourth or fifth (parallel organum). Within a short time the new technique was explored in far greater diversity. Added harmonic lines took on melodic independence, often moving in opposite, or contrary, motion to the given melody. This style was called free organum. In such cases it was impossible to maintain at all times the accepted harmonies of fourth, fifth, and octave. These intervals were considered consonances—*i.e.*, intervals that because of their clear sonority, implied repose or resolution of tension. In free organum they were used at the principal points of articulation: the beginnings and ends of phrases and at key words in the text. In between occurred other intervals that were relatively dissonant; *i.e.*, they implied less repose and more tension. In the following example of free organum, dissonances are marked by asterisks. Free organum is an early example of harmonic motion from repose to tension to repose, basic to Western harmony. The emphasis on consonances at the end of compositions set the final points of arrival in strong relief and reinforced the idea of the cadence, or the finality of the keynote of a mode (on which pieces normally ended).

From the booklet edited by Gerald Abraham accompanying *The History of Music in Sound* (booklet published by Oxford University Press); examples reprinted with permission of the publisher



From the booklet edited by Gerald Abraham accompanying *The History of Music in Sound* (booklet published by Oxford University Press); examples reprinted with permission of the publisher



Until the late 14th century the attitude toward consonance, especially among continental composers, adhered largely to the Pythagorean ideal, which accepted as consonances only intervals expressible in the simplest numerical ratios—fourths, fifths, and octaves. But in England the interval of the third (as from C to E) had been in common use for some time, although it is not expressible as such a simple ratio. A kind of English organum known as gymel, in which the voices move parallel to each other at the interval of a third, existed in the late 12th century; and in the famous *Sumer is icumen in* canon of the 13th century, a remarkably elaborate piece for the time, the harmonic style is almost entirely centred on thirds. The sixth (as from E to C′), an interval closely related to the third, was also common in English music. These two intervals sounded much sweeter than did the hollow-sounding fourths, fifths, and octaves.

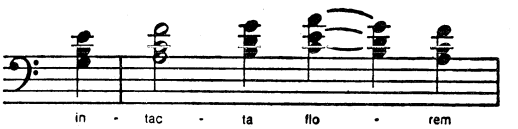
Rise of the intervals of the third and the sixth

By the early 15th century, in part because of the visits of the illustrious English composer John Dunstable (*c.* 1385–1453) to the courts of northern France, the third and sixth became accepted in European music as consonant intervals (prior to this time they were considered mildly dissonant). The result was an enrichment of the harmony in musical compositions.

This was a time, too, of a developing awareness of tonality, the concept of developing a composition with a definite keynote used as a point of departure at the beginning and as a point of arrival at the final cadence.

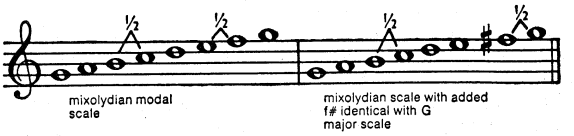
At this time there also began the tendency by composers to think of harmony as a "vertical" phenomenon, to regard the sound of notes heard simultaneously as a definite entity. Although the basic style of composition was primarily linear—*i.e.*, concerned with counterpoint—the chords that emerged from the coincidences of notes in contrapuntal lines took on a personality of their own. One phenomenon that bears out this development is fauxbourdon (French: "false bass"), or, in England, faburden. The following example illustrates English faburden of about 1300. This was a musical style in which three voices move parallel to each other; the middle voice consisted of a succession of notes in parallel organum a fourth below the top voice; the lowest voice paralleled the sequence a third below the middle voice, producing a chord such as G–B–E also known as a $\frac{6}{3}$ or first inversion chord. This was originally an English development adopted in the 15th century by continental composers seeking to enrich their harmonies. It combined the continental fondness for "pure" intervals such as the fourth (here, B–E) with the English taste for parallel thirds (here, G–B) and sixths (here, G–E).

From the booklet edited by Gerald Abraham accompanying *The History of Music in Sound* (booklet published by Oxford University Press); examples reprinted with permission of the publisher



The weakening of the modes

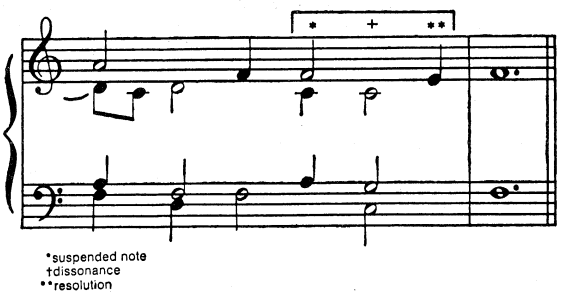
A final phenomenon in early 15th-century harmonic practice clearly foreshadowed the end of the ancient modal system in favour of the major and minor modes of the later common practice period. The old modes were used by composers of the time and persisted to some extent until the end of the 16th century. But their purity became undermined by a growing tendency to introduce additional notes outside the mode. This was achieved either by writing a flat or sharp sign into the manuscript, or by leaving the performer to understand that he was expected to improvise accordingly. The effect of this musical ficta (Latin: "invented music"), as the technique of introducing nonmodal notes was called, was to break down the distinction between modes. A mode, it will be remembered, owes its character to its specific pattern of whole and half steps. Introducing sharps and flats upsets the mode's normal pattern by placing half steps at unusual points. In many cases the resulting change made one mode resemble another. For example, adding an F♯ to the medieval Mixolydian mode (from G to G′ on the white keys of the piano) made that mode's intervals identical with those of the Ionian mode (from C to C′ on the white keys), which in turn is identical with the modern major scale. Likewise, adding a B♭ to the Dorian mode



(from D to D′) made its intervals equivalent to those of the Aeolian (A to A′) mode, which is identical with the modern minor scale. In this way the major and minor modes gradually became predominant over the medieval church modes. The process is especially observable in the music of the late Renaissance.

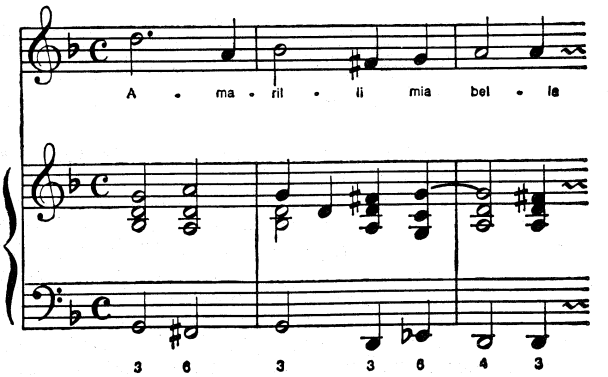
New uses of dissonance

At the same time there emerged a more sophisticated attitude toward dissonance, favouring its use for expressive purposes. By the time of the Fleming Josquin des Prez (*c.* 1445–1521), the leading composer of the Renaissance, contrapuntal music had assumed a more resonant texture through the use of four-, five-, and six-part writing instead of the older three-part scoring. The increased number of voices led to further enrichment of the harmony. A typical Josquin device using harmony for expressive purposes was the suspension, a type of dissonant harmony that resolved to a consonance. Suspensions arose from the chords occurring in contrapuntal music. In a suspension one note of a chord is sustained while the other voices change to a new chord. In the new chord the sustained, or "suspended," note is dissonant. One or two beats later the suspended voice changes pitch so that it resolves into, or becomes consonant with, the chord of the remaining voices. The following illustration from Jean d'Okeghem's *Missa prolationum* shows a suspension at the cadence. The suspension, which became a standard



musical device, creates tension because the expected harmony is delayed until the suspended voice resolves. Its use as the next to last chord of a cadence, or stopping point, was favoured by composers as a way to enhance, through dissonance resolving to consonance, the sense of completeness of the final chord. The use of suspensions indicates a growing awareness both of chords as entities, rather than coincidences, that have expressive potential, and of the concept that harmony moves through individual chords toward a goal. This concept was developed in the harmony of the common practice period.

At the end of the 16th century there was an upheaval in musical style. Contrapuntal writing was frequently abandoned, and composers sought out a style that placed greater emphasis on an expressive melodic line accompanied, or supported, by harmonies. This style, called monody, brought about no marked changes in the harmonic language (the particular chords used), although such composers as the Italian Claudio Monteverdi (1567–1643) did experiment with a heightened use of dissonance toward expressive ends. The major change at this time was in the conception of harmony. The bass line became the generating force upon which harmonies were built. It was often written out with figures below it to represent the harmonies to be built upon it. From this single line—plus figures, known variously as "figured bass," "basso continuo," or "thorough bass"—the accompanying instrumentalists were expected to improvise, or "realize," a full harmonic underpinning for the melody of the topmost voice or voices. In the example below, from the continuo madrigal *Amarilli* by Giulio Caccini, the second line shows the harmonies supplied by the keyboard player. There was, thus, a polarization between the melodic and bass lines, with everything in the middle regarded as



From W.J. Starr and G.F. Devine, *Music Scores Omnibus*, Part 1 (© 1964); Prentice-Hall, Inc.

harmonic filling-in. This contrasts markedly with the older concept, in which all voices were regarded as of equal importance, with the harmony resulting from the interweaving of all parts.

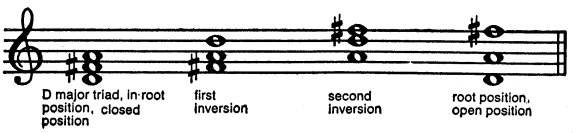
Classical Western harmony: its growth and dissolution. The approach to harmony according to which chords are purposely built up from their bass note marked the beginning of the common practice period of Western harmony. The transition began around 1600 and was nearly complete by 1650. Certain new concepts became important. These had their roots in the harmonic practices of the late Middle Ages and Renaissance and in the medieval modal system. They include the concepts of key, of functional harmony, and of modulation.

A key is a group of related notes belonging to either a major or minor scale, plus the chords formed from those notes, and the hierarchy of relationships among those chords. In a key the tonic, or keynote, such as C in the key of C—and thus the chord built on the keynote—is a focal point toward which all chords and notes in the key gravitate. This is a further development of the idea of a harmonic goal that appeared in the music of the late Renaissance, and that ultimately developed from the medieval idea that modes have characteristic final notes.

In the new system keys further assumed relationships to one another. The larger organizational system embracing keys, key relationships, chord relationships, and harmonic goals was called tonality, or the major-minor system of tonality, because the keys were built on major and minor scales. In the tonal system, given chords assumed specific functions in moving toward or away from harmonic goals, and the system assigning goals to all chords was called functional harmony. The main goal was the keynote, or tonic, of the principal, or tonic, key. Modulation, or change of key, became an important factor in composition because it allowed the composer to exploit the listener's ability to sense the relations between keys.

Rameau's theories of chords

The approach to harmony that emerged about 1650 (the bass-note approach) was soon formalized in one of the most important musical treatises of the common practice period, *Traité de l'harmonie* (1722), by the French composer Jean-Philippe Rameau (1683–1764). The crux of Rameau's theory is the argument that all harmony is based on the "root" or fundamental note of a chord; for example, D. Other notes are placed a third (as D–F or D–F♯) and a fifth (as D–A) above the root. A chord formed in this way is a triad (as D–F–A or D–F♯–A), the basic chord type of the common practice period. The third and fifth above the triad can be placed within the same octave as the root (close position) or can be spread out over several octaves (open position) in compound intervals such as an octave plus a third or two octaves plus a fifth. A triad can exist in its basic, or root position, with the root as the lowest, or bass, note (as D–F♯–A). It can also exist in inversions or rearrangements of its notes placing the third or fifth in the bass, as F♯–A–D (first inversion) and A–D–F♯′ (second inversion). Theorists after Rameau observed that inverted chords are less stable than chords in root position; at the end of a composition, for example, they do not have sufficient finality. Although Rameau's monumental work contains certain elements that later



practices tended to disprove, his writing remains the basis for the study of common-practice harmony.

By Rameau's time no vestige remained of the ancient modal system, which was replaced by 12 major and 12 minor keys beginning on each of the 12 notes of the piano keyboard (C, C♯, D, . . . A♯, B). The invention in the late 17th century of equal temperament (a tuning system) made it possible to play keyboard and other instrumental music in all 24 keys of the chromatic system, the system embracing all possible notes of the 24 scales. (Previous tuning systems made several keys sound out of tune on any one instrument.) Such a work as J.S. Bach's *Well-Tempered Clavier* was, among many things, a set of exercises to acquaint keyboard players with this newfound freedom. Equal temperament also made it possible for a composer to modulate freely from one key to another to obtain contrast in works of an extended nature. Modulation was no new invention, but it now became of prime importance.

In normal, or functional, harmony, the succession of chords is analyzed by the distance, or interval, between their roots. The most common movement from chord to chord is through "strong" intervals: fourths (as C to F), fifths (as C to G), and seconds (as C to D). A movement from one chord to another having this root relation is strong because the two chords have the fewest notes in common and therefore contrast more with each other. Movement by "weak" intervals—thirds (as C to E) and sixths (as C to the A above it)—is weaker, or less pronounced, because the two chords in this case usually share two out of their three notes; for example, C–E–G and E–G–B, or C–E–G and A–C′–E′. Similarly, modulation from one key to another in the course of a piece was most characteristically from one key to another whose keynote is a strong interval apart from that of the first key, as from C to G. Usually the modulation was to the key built on the fifth note, or dominant, of the original scale. A work in C major, for example, tended to move toward the area of G. In works in a minor key, the modulation might be to the dominant minor key (A minor to E minor, for example); or it might be to the relative major key (the key that shares the same scale notes as the minor scale but arranging them in major scale order rather than minor scale order [A minor and C major, for example]). In the second case the contrast of major and minor mode appeared to compensate for the weak modulation (A and C are a third apart).

Harmony and modulation in the 18th century

By the early 18th century these modulatory principles were well established and were made use of in musical form. In the keyboard sonatas of Domenico Scarlatti, for example, or the instrumental dance movements in Bach's partitas, the opening key is well established at the beginning of the piece. There then begins a movement to a new key, normally the dominant key. This is characteristically achieved by an emphasis on chords common to both keys (known as "pivots"), plus a strong musical statement in the new key leading to a cadence in that key. After the modulation there is a process of return to the initial key. During this process the harmonic motion tends to be more rapid, passing quickly through many chords and often including momentary diversions into many new keys, thus lending greater impact to the eventual return to the original key. Such a composition is said to be in "binary form." In binary form compositions in a minor key, there occasionally occurred an exception to the rule of return to the home key. The composer could at his option return to the tonic major, the major key built on the same keynote, or tonic, as the original minor key—A major from A minor, for example. But even in this case the harmonic goal toward the tonic note (A in this case) remained the same.

This basic modulatory scheme from tonic key to dominant key back to tonic key formed the basis of the large-scale musical forms that developed during the 18th century and persisted well into the 19th. The sonata forms of Mozart and Haydn, with their exposition, development, and recapitulation, adhere closely to this plan, often greatly expanded. Here the movement from the tonic to the dominant key or to the relative major key made up the exposition; the rapid harmonic movement en route back to the tonic made up the development; and the return to the tonic key—usually reinforced by a return of the initial thematic (melodic) material—signalled the start of the recapitulation. An optional final coda, or concluding section, further strengthened the sense of the tonal journey's having come to an end. In the large, multi-movement works from this period, there was usually a further contrast achieved by having one of the inner movements in another key, but the final movement almost invariably was once again in the same key as the first movement.

Romantic changes in classical harmony

This clear and logical system of organization seemed highly consistent with an age that took its cues from the clarity and balance of ancient classical architecture. It was not so consistent, however, with the ideals of the ensuing era of Romanticism. Already in the mature works of Beethoven, there is the beginnings of a breaking-down of the classic modulatory scheme; the opening movement of the *Waldstein Sonata*, Opus 53 (completed, 1804) for example, is built on a modulation from the tonic, C major, to the sharply contrasting key of E major, instead of the expected key of G. Much of the individual harmonic language of Franz Schubert (1797–1828) is based on his purposeful disavowal of modulation via the smooth succession of pivot chords, and his fondness, instead, for dropping suddenly into unrelated, and therefore unexpected, keys, as C major to E flat major in the opening movement of the *String Quintet in C Major*, Opus 163 (1828); C major to E minor in the opening movement of the *Symphony No. 9 in C Major* (1828), known as the *Great Symphony*.

Throughout the 19th century there was also a great increase in the use of chromatic tones—tones not belonging to the scale of a given key and that formed "foreign," sometimes dissonant, harmonies with the notes of that key. In addition to the triad, the typical chord of functional harmony, other more complex chords were used, the harmonic functions of which were extremely ambiguous to the listener. As a result the sense of clearly established tonality created by traditional harmonies began to vanish from the musical language—doubtless in line with composers' greater obsession with music and all arts as something mysterious and personalized.

By the time of the German composer Richard Wagner (1813–83), the sense of tonality as the unifying musical force showed definite signs of disintegration. For one thing, Wagner's idea of the "endless melody" led him in his late works to abjure almost completely, except at the end of acts, the full cadence that establishes tonality. A seeming approach to a cadence in *Tristan und Isolde* or the *Ring des Nibelungen* tetralogy is more often than not thwarted by a quick and unprepared switch to a sharply contrasting key and a continuation of the music in that new area. For another, Wagner's passion for complex chords subject to more than one functional interpretation made the tonality of even short passages difficult to assess.

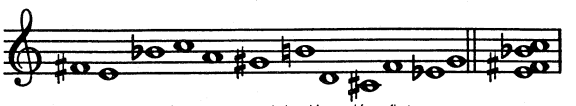
Post-Wagnerian developments

Although Wagner's specific harmonic concepts were not universally accepted, during his time or afterward, the blurring of the tonal sense by one means or another became prevalent throughout Western music by the last decades of the 19th century. Even in the works of the Italian Giuseppe Verdi (1813–1901), whose music was regarded as the opposite pole from Wagnerian techniques, this abandonment of clear tonal outlines may be noted: the sudden changes to unrelated keys, the piling up of dissonances that leave the sense of key obscured for minutes at a time, the emergence in his late works of a continuous melodic style that avoided regular, key defining cadences. In France, the blurring of clear outlines characteristic of Impressionist painters found its musical counterpart in the music of Claude Debussy (1862–

1918), who employed such devices as the scale consisting entirely of whole tones as a means of sidestepping the tonal feeling created by traditional scales. In the music of later French composers, especially the members of the post-World War I group known as "Les Six," a common practice was polytonality, or the sounding of two tonalities simultaneously, each defined with relative clarity, but neither dominating the other. Similar polytonal methods also occur in the works of the Hungarian-born Béla Bartók (1881–1945) and the Russian émigré Igor Stravinsky (1882–1971).

The Wagnerian influence continued most directly, via the music of Gustav Mahler (1860–1911), into the serial techniques developed in the 1920s by Arnold Schoenberg (1874–1951) and his Viennese school. In Schoenberg's serialism the 12 notes of the chromatic scale are arranged into an arbitrary series, or 12-tone row, that becomes the basis for the melodies, counterpoint, and harmonies of the composition. Of these 12 notes no single note is allowed to predominate. This is in complete contrast to the predominance of the tonic, or keynote, in the music of the late Renaissance and the common practice period. Serialism thus completely and systematically obliterated traditional harmonic organization. With no single note serving as a musical goal, tonality—at least as it was known from the 15th century—ceased to be a unifying musical force. Other elements, including serialization of rhythms and tone colours as well as of notes, came to prevail.

Schoenberg's 12-tone row



twelve-tone row and chord formed from first four notes of this row

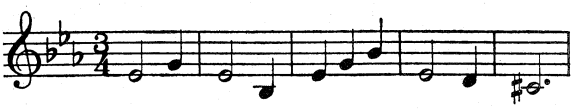
CHROMATICISM IN HARMONY

Although the preceding paragraphs represent a brief outline of composers' attitudes toward harmony and tonality from the late Middle Ages to the present, there is the danger that the broad outlines may be taken as a rigid statement of standard practices by composers at any period in musical history. Actually, although these outlines remained the general framework in which composers worked, they frequently diverged from it to some extent, particularly in their use of chromatic notes (notes outside the scale of the basic key of the composition) and chromatic chords (chords containing chromatic notes).

The capacity of chromatic tones to add harmonic colour, expressiveness, and interest was apparent to composers from the beginnings of standard harmonic practice. J.S. Bach, for example, in a striking passage at the end of the "Crucifixus" of the *Mass in B Minor*, lent poignance to the verbal description of the burial of Christ by the musical device of a sudden modulation from B minor to a sharply contrasting new key, G major, that contained notes chromatic to the basic key. Mozart, too, derived much of the drive of his harmonic style from a constant use of chromaticism. A characteristic device of Mozart, for example, is his frequent use of secondary dominants to intensify harmonic movement. A secondary dominant is a chord related to the dominant; specifically, it is the dominant of the dominant. If the key is C, the dominant is G and the secondary dominant is D. Secondary dominant chords by their nature contain a note that is chromatic to the basic key. In Mozart's music a harmonic progression from tonic chord (I) to dominant chord (V) will often pass through the dominant of the dominant (V-of-V): from I to V-of-V to V. By using the secondary dominant, he expanded the harmonic range of the composition by introducing chromaticism. In his later works Mozart also came to rely more and more on the dissonant value of suspensions to create harmonic interest. The slow introduction of his *String Quartet in C Major*, K. 465 (the *Dissonant Quartet;* 1785) consists of a string of long-delayed suspensions so that the harmonic definition at any given instant is as blurred as anything in Wagner.

Although the harmonic style of the common practice period remained a basic framework, the history of music from Mozart's time to the present shows a constant increase in harmonic density, or the amount of chromaticism and frequent chord changes present. The opening bars of Beethoven's *Eroica Symphony* demonstrate the power of chromaticism to enhance the emotional effect.

Increased use of chromaticism



The first eight notes of the theme are resolutely normal in their outline, the triad of E flat major, the tonic chord of the movement. But the ensuing two notes lead violently away from this harmonic stability, with the tenth note a totally unrelated C sharp. This completely upsets the harmonic structure and gives unmistakable notice that a long, complex movement will be necessary to right the imbalance. Not until the coda of the movement, 15 minutes later, is this opening theme allowed to follow the expected harmonic outline dictated by the style of the times.

Throughout the 19th century, composers remained rooted in the basic concept of tonality, while at the same time doing everything in their power to complicate or obscure the tonal sense for the listener. Even at the present time, the large, varied, and important group of composers who are called conservative—among them, Samuel Barber (1910–), Aaron Copland (1900–), Sir William Walton (1902–), Dmitri Shostakovitch (1906–), Gian Carlo Menotti (1911–), Benjamin Britten (1913–)—adhere to the concept of tonality only as a challenge. Tonality in their works exists, in the sense that there are extended stable areas that give the impression of being in some definable key. But the intense chromaticism of current-day composition, be it conservative or radical, makes it nearly impossible for the listener to grasp the unity of a work in terms of its adherence to a clear tonal plan. Unity is achieved, rather, by melodic means, the organization of rhythms, or even of tone colour. For all practical purposes, the function of tonality as the prime unifying force in musical structures, known from the 15th through the 19th centuries, is a thing of the past.

THE ROLE OF DISSONANCE IN HARMONY

The very foundation of harmonic music has been the interplay of consonance and dissonance. Consonance can be defined as the normal range of tone combinations accepted by theorists and composers of any given time as implying repose; dissonance, therefore, refers to any sounds outside that range. From the 19th century on, as writers increasingly explored the exact effect of music on the emotions, these two terms took on the aspect of value judgments. There is a tendency to confuse consonance with concord, or sweet sound, dissonance with discord, or clashing sound. This has led to a certain amount of confusion.

Dissonance is in fact the prime element in the harmony that creates movement, and this has been recognized by composers from the dawn of the harmonic millennium. When the human ear recognizes a certain harmony as unstable within the context of a composition, it demands that this instability be rectified by the resolution to a stable harmony. Dissonance, therefore, has never been forbidden in music, for without it music would be hopelessly static. What has been clearly defined in each era has been, rather, the treatment of dissonance, the approach toward it and away from it in a smooth and logical manner so that the musical flow is of a continual tension and relaxation.

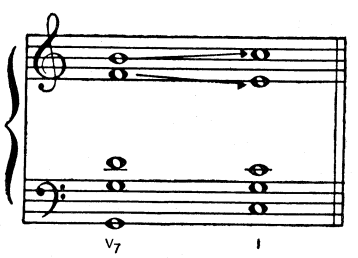
The regulation of dissonance. The notion of which specific chords and intervals constitute consonance and dissonance has altered violently from the beginning of harmony. In the earliest harmonic writing, the parallel organum of the ninth century, the accepted intervals were the perfect consonances, or those of the simplest harmonic ratios: the fourth, the fifth, and the octave. As contrapuntal movement among voices became freer, certain other combinations necessarily occurred: thirds

and sixths and, in some cases, seconds (as C–D) and sevenths (as C–B). These combinations were regarded as dissonances and were to be confined to weak beats of the musical metre; they were to be resolved, for the most part, by stepwise movement downward to the adjacent consonance. Another interval that the musicians of the modal era took great pains to avoid was the augmented fourth (the tritone, or "devil in music"), an interval containing three whole steps, as between F and B—the whole steps F–G, G–A, and A–B. This interval was considered intolerably dissonant. Primarily to avoid the forbidden, unstable harmonic relationship of the tritone, the use of accidentals (sharps, flats, natural signs) entered music and introduced chromatic tones into a mode.

Chords with a pull to cadence

By the time of Rameau, the concept of the dissonance had altered markedly. The basis of the harmony had changed, as noted above, from the perfect intervals (unison, fourth, fifth, octave) to the triad, or chord such as C–E–G, built of thirds above a root, or bass note. The tonic, or keynote, triad became the point of departure and of arrival for an entire composition and also for melodic phrases and larger sections within that composition. The harmonic movement to the cadence, a prime means of establishing points of articulation, became by the mid-18th century a more or less standard progression of harmonies subject to variation according to the composer's own powers of imagination. Preceding the tonic chord in these cadences, and pushing toward it, was the chord built on the dominant, or fifth note of the scale. This convention developed because of the nature of the dominant chord. In a dominant chord, the note a third above the root (as B in the chord G–B–D′—considering the G chord the dominant and the basic key C) is the seventh note of the scale (C, D, E . . . B). This note has a strong leading tendency toward the tonic, or keynote (here, C), because it is only a half step away from the tonic, and is thus called the leading note. Because the leading note is a member of the dominant chord, this chord also has a strong pull toward the tonic chord.

By Rameau's time it was also a common practice to enhance the pull of the dominant chord to the cadence by adding to it the note a seventh above the root of the chord (as F′, in the chord G–B–D′–F′), that note being the fourth note of the scale (C′, D′, E′, F′). Such a chord, a dominant seventh chord (V_7) contains two leading notes: the seventh of the scale, here B, with its strong pull toward the tonic and the fourth of the scale, here F′, which has a strong pull toward another of the notes of the tonic chord (in this case toward E′ in the chord C′–E′–G′), being a half step away from that note. In this way two notes of the dominant seventh chord pulled strongly toward two notes of the tonic chord. Another reason for the strong pull of the seventh chord toward the tonic is that that chord contains a tritone (in this case B′–F′). Although the tritone was less intolerable by that time than it was to medieval ears, it was still considered a particularly strong dissonance that demanded resolution. This resolution occurred when the dominant seventh chord moved to the tonic chord. In the example below a dominant seventh chord (V_7) moves to a tonic chord (I) in the key of C major. Arrows show the resolution of the tritone dissonance. The dominant seventh chord thus became one



of the basic chords in functional harmony. In addition, because it contained two dissonances (a seventh, as G–F′ in the chord G–B–D′–F′, and a tritone, as B–F′ in the same chord), it was the first instance of incorporating dissonance into a system built on the basically consonant triad. Throughout the common practice period dissonances were continually added to the basic harmonic language, so that the range of harmony and use of dissonance in late 19th-century music had expanded considerably beyond that of the early 17th century.

Music using the system of functional harmony has a flow of harmonic movement through contrasting chords and through passages from consonant to dissonant to consonant chords. If the change of chords is frequent in relation to the musical rhythm, there is said to be a rapid harmonic rhythm. Similarly, a leisurely pace of chord change is a slow harmonic rhythm. The slow or fast harmonic rhythm of a composition helps define its musical character, and by varying the harmonic rhythm within a piece a composer can create contrast, thereby defining sections of musical form.

Modulation and dissonance in the common practice period

Modulation, or change of key, was, like dissonance, increasingly explored during the common practice period. In the sonata forms that emerged as the primary musical forms of the mid-18th century, modulation from the tonic to other keys as a means of obtaining contrast became of prime importance. This musical esthetic involved not only the necessity of modulation itself but also drew much of its strength from the varying rate of modulation. Thus, the exposition, or first section, of the "normal" sonata form involves a modulation from the tonic to a nearby related key—usually the dominant, or in works in a minor key, the relative major. The development, or second section, on the other hand, depended on a rapid series of modulations, the purpose being to cast the return to the tonic in as strong a dramatic light as possible by having the stability of the tonic contrast with the instability of rapid modulation that preceded it.

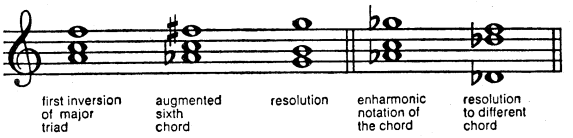
The process of modulation to many keys involved the addition of dissonant, often chromatic, notes to the basic harmonic outline of a composition. A common way of preparing for the appearance of the dominant key area in a sonata exposition was for the composer to overshoot his mark, moving temporarily to the dominant of the dominant, thereby using chromatic chords. Thus, in the transition from tonic key (C major) to dominant key (G major) in the first movement of *Symphony No. 1 in C Major*, Opus 21 (1800) of Beethoven, there is considerable emphasis on the chords of D, both major and minor, establishing D as a dominant leading to a cadence on G, the point of arrival. Much of the dissonance in music of the late classic composers is traceable to this use of secondary dominants. The tendency to move quickly through extended sequences (musical patterns repeated at higher or lower pitch) based on secondary dominant chords became a highly sophisticated technique in the mature works of Haydn and Mozart (as for example, the extraordinary sequence in the slow movement of Haydn's *Symphony No. 104 in D Major*, the *London Symphony*).

Functional harmony was based on chords built from the diatonic (seven-note) major and minor scales; chromatic notes and chords were integrated into the functional system. Although composers of this period proved remarkably adventurous in straying beyond the limits of purely diatonic harmony, their use of dissonance and chromatism was at all times both rational and functional. Chords, even though complex, normally resolved sooner or later into the chords toward which they tended, even when the composer, as in the Haydn passage cited, added unstable elements to the chord of resolution, and therefore occasioned further resolution.

Use of dissonance for harmonic colour. By the early 19th century, composers became aware that harmony could also serve another purpose: it could exist outside of a purely functional context as a means of enhancing the pure harmonic colour of a composition. The opening of the *Quintet in C Major* of Schubert provides a simple and quite early example of chords used for the sheer effect of their sound. The C major triad of the first two bars seems to swell out in the ensuing two bars into a diminished seventh chord, a chord functioning much like a dominant chord in its pull to its tonic, but built instead with a leading note as its root, as, for example, F♯, the leading note of G, on which is built a chord such as F♯–A–C–E♭. (The top and bottom notes of such a chord, here

F♯–E♭, encompass the interval of a diminished seventh, giving the chord its name.) In Schubert's quintet the particular diminished seventh chord used would normally resolve to a chord on G. Instead it simply subsides back to the C major triad of the preceding bars, so that there occurs no real harmonic movement in the opening six bars.

Nineteenth-century harmonic usage, therefore, tended to expand not only the chordal vocabulary itself but also the function of chords. In the former respect there was an increase in the use of chords the particular type of dissonance of which lent them an unstable and a functionally ambiguous quality; for example, a chord that became of prime importance as a means of thickening the harmonic sound and of blurring the exact tonality of a musical passage was the augmented sixth chord. This is an altered chord, or one built by taking a chord normally occurring in its key and chromatically altering it. In this case, two of its notes are changed by a half step. Specifically, an augmented sixth chord is built on the first inversion of a triad, as, for example, A–C′–F′, the first inversion of the triad F–A–C′. Taking the first inversion (A–C′–F′), the A is flatted and the F′ is sharpened, resulting in a chord (A♭–C′–F♯′) that is both dissonant and ambiguous in harmonic function. The ambiguity of sound is partly due to the nature of enharmonic chords, chords that sound identical but in musical notation use different notes (as G♭, identical in sound with F♯). Thus the chord A♭–C′–F♯′ may move smoothly to a chord built on G, but the identical sounding chord A♭–C′–G♭′ will progress to a vastly different chord, on D♭. Composers can thus

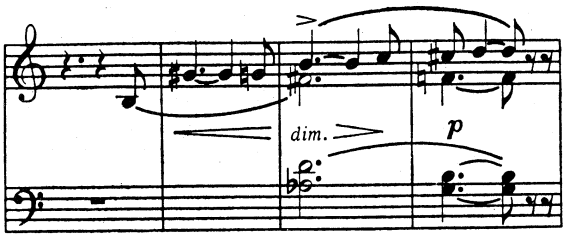


use such ambiguous chords to achieve unusual or expressive harmonies that blur the listener's expectations and therefore his ability to perceive key and tonality.

Dissonance and emotion in Wagner's music

The opening of Wagner's musical drama *Tristan und Isolde*, famous for its ambiguous sense of tonality, is an augmented sixth chord that resolves by way of a second dissonance to the dominant seventh chord of the key of A. This sequence is repeated at a higher pitch, here resolving to the dominant seventh chord of the key of C.





Although this passage can be explained in terms of normal harmonic analysis, it was in itself strikingly abnormal for its time. The passage occurs at the beginning of the composition, the point where a composer normally would be expected to establish his basic tonality. In addition, there is considerable doubt as to the exact nature of the resolution. The dominant seventh chord (here the chord of resolution) is itself dissonant, although less so than the augmented sixth chord. The tonality of the passage is obscured, for it is impossible to tell whether the passage is in A minor or A major. Since the notes of the scale that would give this information to the listener are missing from the passage, it is clear that Wagner does not want the listener to be sure. He wants the passage, rather, to stand for the substance of the opera itself: unrequited passion is equal to unresolved harmonies.

Other composers, too, sought out harmonic as well as melodic and rhythmic means to underscore the content of passion, restlessness, mystery, or tragedy in their scores. The unstable, ambiguous chord of the diminished seventh accompanies the appearances of the evil Samiel and his seven supernatural bullets in the opera *Der Freischütz* (*The Freeshooter*) by Carl Maria von Weber (1786–1826). Long strings of this chord, moving rapidly up and down the scale for purely colouristic purposes, also appear in climactic passages of the tone poem *Les Préludes*, by Franz Liszt (1811–86), expressing the struggle of the soul against supernatural forces. The highly embroidered piano style of Frédéric Chopin (1810–49) touched, in passing, on showers of dissonant, often chromatic tones—again used not for any exploitation of their functional value but as a spray of colour used as an overlay for a basically diatonic (nonchromatic) style, well-hidden underneath and recognizable only at the cadence.

Until the genuinely revolutionary *Tristan und Isolde* of 1865, the increase in the amount of chromaticism in the musical language of the Romantic composers was largely an enhancement of expressive detail. The diatonic (nonchromatic) basis of 18th-century functional harmony was in the main respected, as was the orderly process of modulation as a means for giving structure to large musical forms. With *Tristan* and even more markedly with Wagner's music drama *Parsifal*, one can discern the beginnings of a gradual but unmistakable dissolution of the diatonic system on which traditional harmony was based. The analysis of *Tristan*'s harmony by Rameau's principles, although possible, is simply unimportant. What matters more is the constant flow of chromaticism, of Wagner's wide variety of means—altered chords, chains of secondary dominants, and resolutions to chords that themselves prove unstable—for blurring any sense of functional harmony. Doubtless impelled by the dramatic substance of this music drama, he succeeded in evading the cadence, or coming to rest, that traditionally defined harmonic direction.

The impact of this step became apparent in the directions taken by harmony by the end of the 19th century. After Wagner, dissonance, particularly dissonance caused by chromaticism, largely ceased to function as it had in traditional harmony, and composers created their own individual, often experimental, usage of dissonance. No composer, whether he accepted *Tristan* as a masterpiece or dismissed it as madness, was left untouched by its implications.

Dissonance after Wagner. In France, where musical culture stood in some ways the direct antithesis to Wagnerism, Claude Debussy evolved his own style that succeeded, as Wagner's had, in beclouding the harmonic basis of a work either altogether or for extended periods.

Debussy's sources

Debussy was influenced by a number of sources: the Impressionist painters, who were involved with the renunciation of clear perspectives and outlines in favour of the play of light across surfaces and the effect of images only half seen; exotic music, particularly that of Indonesia; and folk music, especially the modal scales of Russia. All of these led him to a partial abandonment of functional tonality. Among the devices he used toward this end is a scale composed entirely of whole tones (as C–D–E–F♯–G♯–A♯–C′). Such a scale lacks the distribution of whole and half steps that define the character of the major and minor scales of the common practice period. Chords built from the whole tone scale are by normal harmonic analysis unstable: all possible triads are augmented (the top note is altered by being sharpened; for example, C–E–G♯ instead of C–E–G) and as a result are dissonant. The perfect fourth and fifth, the ancient cornerstones of harmony, do not exist. Because the chords to which dissonances traditionally resolve are impossible with this scale,

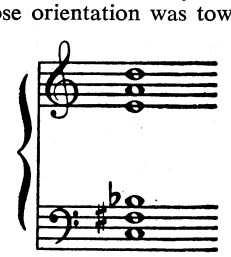
a work built upon it—*e.g.*, "Voiles" ("Sails"), from the first book of preludes for piano—can be said to exist without harmonic resolution and, therefore, without traditional tonality. Other Debussian devices include the regarding of the seventh chord (*e.g.*, dominant seventh, diminished seventh) as a self-sufficient harmony instead of as a dissonance that must resolve; sequences of sevenths moving parallel to each other giving the effect, in his music, of lines of harmony plus a dissonant descant (a countermelody in the highest part, or voice) blurring any real sense of traditional harmonic movement. This use of self-sufficient seventh chords was also much exploited by Maurice Ravel (1875–1937) and came, through his great appeal, into a great deal of the popular music of western Europe and America from the 1920s to the present time.

Again, as with Wagner, Debussy's methods cast their shadow over composers both influenced by and hostile to his musical style. Igor Stravinsky, who was a little of both, first mirrored some of Debussy's harmonic usage in *Le Sacre du printemps* (*The Rite of Spring;* 1913). In *Le Sacre*, chords appear, as they often do in Debussy, purely for their colouristic value, related to each other only by virtue of the rhythmic insistence in the music's patterns. Much of Stravinsky's harmonic style, however, is actually derived from much simpler elements than Debussy's. His complex chord structures often break apart to reveal two unrelated and dissonant diatonic chords sounded simultaneously. In the works of his Neoclassical period, Stravinsky reverts to a clear harmonic language reminiscent, at least as regards individual chords, of the 18th century; but in harmonic movement from chord to chord there is a noticeable difference from earlier styles. Stravinsky, even in this clear compositional style that occupied him in the 1920s and into the 1930s, tends to use these classical harmonies in isolation, for the chords move freely one to the other without their classical function.

Polytonality of "Les Six"

Similar in a sense to Stravinsky's pandiatonicism, or use of diatonic chords without the limitations of classical harmonic function, is the tendency toward polytonality in the works of the post-World War I group of French composers known as "Les Six." These composers, notably Darius Milhaud (1892–), worked for a time with simple, diatonic chords piled upon each other in a way that suggested a clash between simultaneous tonal areas, almost a kind of counterpoint of tonalities—again leading to the dissolution of any sense of a single, central key area. Some traces of polytonality also occur in the early works of Bartók, who was much taken with French influences early in his career. But Bartók did not pursue this device to any great extent later on. He turned instead to an exploration of the folk styles of eastern Europe—Hungarian and Rumanian, predominantly. His music, though harmonically dense and complex, remained rooted in tonality, with an admixture of harmonies gleaned from the modal scales of folk music.

Certain other composers, similarly obsessed with the desire to expand the harmonic vocabulary but loath to abandon the tonal system entirely, experimented with some success with synthetic scales of their own devising, and with chords built of intervals other than the third. The Russian mystic Aleksandr Scriabin (1872–1915) and the German Paul Hindemith (1895–1963) both worked extensively with chords built out of fourths (as C–F–B♭). Scriabin employed these sounds primarily in a quasi-Impressionist way, using their unusual sounds as sonorous self-sufficient units. His "mystic" chord, shown below, formed the entire basis for many of his later works. Hindemith, whose orientation was toward the Neoclassical, dealt with these chords by devising his own system of harmonic function, creating a quite successful reincarnation of the dissonance-consonance tensions of earlier composers.



The direct impact of Wagner's methods, however, was felt within the German-Austrian orbit. The restless, nonresolved chromaticism of *Tristan* was directly reflected in the late works of Gustav Mahler. In such a work as the long, slow movement that ends Mahler's *Ninth Symphony*, one feels the *Tristan* influence quite directly: the long, lyric lines move freely through a systematic evasion of cadences and through a widening range of tonalities, often reaching tonal regions far removed from the starting point. Yet Mahler, too, remained a tonal composer. So did Richard Strauss, whose overlays of dissonance in such works as *Elektra* are easily separable from a basically tonal harmonic movement.

Schoenberg's relation to Wagner

And so, in his early years, did the Viennese Arnold Schoenberg. Such scores as the string sextet *Verklärte Nacht* (*Transfigured Night;* 1899), the *Chamber Symphony in E Major*, Opus 9 (1906), and the first two string quartets are direct outgrowths of *Tristan*'s chromaticism, masking but not obliterating the tonal basis. But by 1912 Schoenberg began actively to question the importance of tonality as a musical inevitability and to accept the broader implications of Wagner's style. From then on the pileup of dissonance in Schoenberg's music became so pronounced as to make the concept of dissonance itself meaningless. In such a seminal work as the chamber cantata *Pierrot Lunaire* (1912), tonality has been put aside. And since this is so, it is no longer possible to discuss consonance and dissonance, for these concepts relate to the structure of a composition according to the harmonic principles of tonality.

Schoenberg's far-reaching musical philosophies, which were epitomized in his invention of the technique of serialism, have had a potent impact on the music of the decades following his own writing. They have also been resisted by large numbers of composers who are conveniently, if not always accurately, described as conservative. It seems likely that the conflict between tonality and atonality (*i.e.*, nontonality) will provide the dynamic dualism for musical styles in the foreseeable future.

HARMONY AND MELODY

As noted above, melody and harmony were synonymous in classical Greek theory; the term harmony referred, not to notes sounded simultaneously, but to the succession of notes, or the scale, out of which melody was formed. During classical antiquity and the European Middle Ages melodies were written that had an inner logic in terms of their scale, or mode, its important notes, and the melodic patterns associated with it. This is also true of many non-Western melodies. After the gradual evolution in Europe, through the polyphony of the late middle ages and the Renaissance, of the common practice, or classical, system of Western harmony, the inner logic of melodies was strongly affected by harmony. Because the ear can perceive harmonic patterns in certain groups of notes, even when sounded successively rather than simultaneously, melodies began to carry a strong implication of underlying harmonies. During this period there arose the conception that melody was the surface of harmony. Thus, for example, the partitas for unaccompanied violin by J.S. Bach, despite their melodic basis and lack of outright harmonic underpinning, clearly set forth their basic tonality and harmonic direction. This is achieved by a melodic style that includes frequent scale passages and arpeggiated chords (chord notes played successively, in melodic fashion, rather than simultaneously, as in a chord) that make clear to the listener the scales, harmonies, and keys belonging to the tonality of the composition. Through the 18th century and well into the 19th, melodies tended to be the bearers of their own harmonic implications. The above noted opening of Beethoven's *Eroica Symphony* represents this practice at both its height and at the beginnings of its dissolution. The opening eight notes outline an unmistakable E flat triad, and

Melody as the harmonic surface

would do so even if they were sounded unharmonized; the ensuing plunge to the unexpected C sharp likewise indicates by its mere melodic shape the harmonic unrest arising at this juncture.

Nevertheless, melody in the hands of a composer seeking a genuine expressiveness must function with some degree of independence from its harmonic underpinning. The long, expressive dissonances in the vocal lines of Romantic composers not only heighten the passion sought after in the music they also specifically represent a seeking for a heightened independence of melody from harmony.

The shift of harmonic usage in the 20th century can be viewed partly as a marked change in relationships between melody and harmony. In Schoenberg's techniques, the generating force is the 12-tone row, which is primarily a melodic sequence out of which harmonies, as well as themes, are generated. Thus, it is possible to detect a reversal of the traditional relationship, whereby harmony has become the surface—or at least the final result—of melody.

HARMONY IN MUSICAL FORM

The chief problem of composition, in any style from ancient times to the present, is the creation of a form, or structure in which the principles of unity and contrast operate in some kind of equilibrium. The listener himself enters into this process by the use of his powers of recognition and of memory.

In purely melodic, modal music the form often derives from the inner logic of the melody in terms of the important notes and melodic patterns of the mode. In polyphonic music before the common practice period musical form depended partly on the unity achieved by setting a piece in a given mode, partly on the use of musical themes, and partly on creating harmonic movement and tension toward stopping points, or cadences. During the common practice period—from Bach to Debussy—much of the creation of musical form took place through the organization of harmonies into keys and relationships between keys. Thus, the sonata forms of the 18th and early 19th centuries depended as much on the statement of a key, the movement to other key areas and the eventual return to the same key, as they did on themes and other melodic devices. The composer was likely, of course, to employ the two principles of melody and harmony simultaneously; return to the tonic key late in the course of a movement was usually reinforced by a restatement of the initial themes. In certain works of Haydn and Mozart, the listener was often thrown off course purposely by the premature return of initial themes in an unexpected key; such devices served further to enhance the drama of the genuine recapitulation or return to the main key.

By the 19th century, however, the power of harmony to suggest clear formal structures was greatly undermined by freer use of dissonance, which broke down the clarity with which a key was defined. Other customary procedures were also abandoned. Many of Schumann's songs do not return to the tonic, or home key, for the final cadence. The extended length of Wagner's music dramas, and their wide-ranging modulation, make it impossible to regard key as a unifying force. Mahler's *Ninth Symphony*, the first movement of which is in D major, ends with a movement in D flat major, and since the symphony lasts nearly 90 minutes, there seemed to Mahler to be no reason to pay any closer lip service to classical practices of unity of key. Such necessities, by the time of this symphony, had vanished from the musical language.

AVANT-GARDE CONCEPTIONS OF HARMONY

The course of harmony after Wagner followed three distinct paths. (1) Within the broad outlines of tonality, composers explored the potential of chords of far greater complexity than the traditional triad. In doing so they often allowed unstable chords such as dominant sevenths to stand as self-sufficient entities, and they greatly increased the use of ambiguous chords such as augmented sixths and diminished sevenths to thicken or occasionally to blur the sense of a stable tonality. (2) Composers broke away from the classical system of tonality by using chords that, although clearly recognizable as derived from earlier harmonic practices, resolved in other than the expected direction. Some also went further afield by substituting for the major and minor scales unusual scales such as whole tone and Gypsy folk music scales; by using chords built out of fourths; and by utilizing polytonality. (3) Composers systematically abandoned tonality through Schoenberg's technique of granting equal importance to all twelve chromatic tones, rather than allowing one tone to predominate as tonic. When this was done, the concept of a single, predominant key centre vanished entirely in favour of atonality. In such cases, the traditional duality of consonance and dissonance also disappeared.

New ways of creating harmony

Among most "progressive" composers of the 20th century, atonality has been extensively explored. By far the greatest concern among avant-garde composers has been to revive contrapuntal writing, or composition stressing the combination of independent melodic lines. This was partly a reaction against the lush harmonies and lyricism of the Romantic period. During the common practice period any counterpoint that occurred was subordinated to the principles of traditional harmony. The 20th-century obsession with counterpoint tended to sweep aside concern with harmonic relationships beyond the incidental fact that clusters of notes in counterpoint are indeed heard simultaneously. In the music of the American Charles Ives (1874–1954), for example, many skeins of fully developed atonal, contrapuntal writing pass by simultaneously, producing momentary sonorities. Such sonorities may occasionally, and quite accidentally, be identical with recognizable harmonies; but these accidental sonorities have little to do with traditional harmonic organization. Similarly, the "tone-cluster" writing of another elder American innovator, Henry Cowell (1887–1965), whereby a pianist's forearm sounds every note it can depress at once, can hardly be analyzed as functional harmony in any sense.

Other developments, too, point to the dissolution of traditional attitudes toward harmony. The aleatoric, or indeterminacy, experiments of John Cage (1912–), Earle Brown (1926–), and others, assign part of the composer's melodic, harmonic, and rhythmic events to a specific performer at a specific instance. In such music any discussion of harmonic direction is irrelevant. Most importantly, the rise of electronic music, which breaks away from any traditional scales such as might be produced on "normal" instruments, can only with the greatest stretch of the imagination lend itself to considerations of harmony.

Yet, there is a possible analogy with traditional harmony in electronic music, as its musical styles and languages tend to take shape. In the works of Karlheinz Stockhausen (1928–), one of the electronic pioneers and a composer of enormous influence among his younger colleagues, there are organizational systems emerging that point to a clear control and regulation of musical elements. The strict control of musical factors such as densities of sonority, rates of rhythmic change and of change in phrase structure, rates of change in the spread of sound in an auditorium through the use of carefully positioned and modulated loudspeakers point toward a new musical system that may possibly be analyzed in terms of a new, fundamentally different harmony.

Search for a new order

The dissolution of harmony in the "progressive" music of the 20th century was not a matter of anarchy replacing order. Actually, the common practice period is of relatively short duration against the entire history of harmony. Before Bach other rules existed. Such rules, in contrast to the later system of traditional harmony, depended not on the contrast of keys but on harmonic unity brought about by the use of a given mode. Since Debussy, similarly, harmonic styles have been dictated by new rules, or by the desire of many composers to seek out new rules. And as both the modal and the common practice systems of harmony evolved only after centuries, so is it also safe to predict that the seeming anarchy of much

of today's music represents a state of movement toward new harmonic precepts. The question at hand, moreover, is not one of the dissolution of harmony itself, for any notes sounded simultaneously produce a harmony—whether the notes be from traditional scales or from the infinity of musical pitches produceable through electronic means. The matter is, rather, the question of the uses to which these harmonies are put, and the changing relations of harmony to musical structure.

An awareness of the value of harmony as pure, expressive sound persists among all composers of the present time. Some have pursued the atonal principles toward the point where harmonic sounds are totally dissonant (which is the same as saying that they are all consonant, because the contrast between consonance and dissonance disappears). Others have written works that consist of almost nothing but static, unadorned harmony—not necessarily harmoniousness. Such a work as Terry Riley's *In C*, for example, consists basically of a sustained triad on C (lasting, at the performer's option, anywhere from 30 minutes to several hours), over which fleeting dissonances are occasionally sounded, seldom more revolutionary than an F sharp or B flat. Here again, although one is more conscious of harmony in this work than of any other musical element, the harmony itself does not move or progress in the traditional sense; the sound exists, but not its function. Here one can discuss the work as pure consonance—which is the same, for lack of harmonic contrast, as pure dissonance.

Thus, in the 20th century the concepts basic to traditional harmony began to lose their importance. In counterpoint harmonies became the incidental result of the combination of melodic lines. New experiments with unusual harmonies, such as tone clusters, functionless in the traditional sense, the lessening of the tension between consonance and dissonance, the creation of unprecedented harmonies by the use of computers, have been the result of a search for new methods of musical organization. This in turn was the natural outgrowth of the blurring and final dissolution of the harmonic system that had prevailed for over two centuries in Western music.

BIBLIOGRAPHY

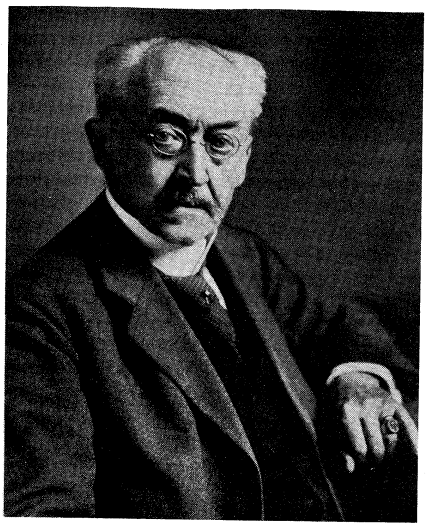
General historical sources: D.J. GROUT, *A History of Western Music* (1964), is an important outline of stylistic development from ancient music to the present. Other specialized histories of specific periods include CURT SACHS, *The Rise of Music in the Ancient World, East and West* (1943); GUSTAVE REESE, *Music in the Middle Ages* (1940), *Music in the Renaissance* (1959); M.F. BUKOFZER, *Music in the Baroque Era: From Monteverdi to Bach* (1947); R.G. PAULY, *Music in the Classic Period* (1965); and ERIC SALZMAN, *Twentieth-Century Music: An Introduction* (1967). W.S. NEWMAN, *A History of the Sonata Idea*, 3 vol. (1965–69), is a comprehensive survey of large-scale musical form from the Baroque through Mahler. Selections from the actual writings of musical theorists and aestheticians from Plato through Wagner may be found in W.O. STRUNK (ed.), *Source Readings in Music History* (1950), an invaluable work.

Textbooks and theoretical works: Two works by 20th-century composers that give considerable insight into the role of all musical elements in composition are PAUL HINDEMITH, *Unterweisung im Tonsatz* (1937–39; Eng. trans., *The Craft of Musical Composition*, 2 bks., 1941–42); and ARNOLD SCHOENBERG, *Harmonielehre*, 3rd ed. (1922; abridged Eng. translation, *Theory of Harmony*, 1948). Schoenberg's formulation of his 12-tone theories may be found in his *Style and Idea* (1950). Other theoretical works that trace the fluid state of harmony since the later 19th century are ELLIOTT ZUCKERMAN, *The First Hundred Years of Wagner's "Tristan"* (1964); HENRY COWELL, *New Musical Resources* (1930, reprinted 1968); and GEORGE PERLE, *Serial Composition and Atonality*, 2nd. ed. (1968). Standard textbooks based on the theories of Rameau are WALTER PISTON, *Harmony*, 3rd ed. (1962); and ROGER SESSIONS, *Harmonic Practice* (1951).

(A.Ri.)

Harnack, Adolf von

Adolf von Harnack, the leading historian of the Christian Church in the late 19th and early 20th centuries, exerted a profound and long-recognized influence in modern theological and historical scholarship and in scientific endeavours.



Harnack, photographed during the 1920s.
The Bettmann Archive

Early life and career

Harnack was born on May 7, 1851, in Dorpat, Livonia (Estonia), where his ancestors had moved from Germany. His father, Theodosius Harnack (1817–89), was a professor of practical and systematic theology, first in Dorpat, then for 13 years in Erlangen, Germany, and again, until his death, in Dorpat. His chief work, dealing with the theology of Luther, is still widely read.

Adolf von Harnack was educated at the universities of Erlangen, Dorpat, and Leipzig. After obtaining the doctorate with a dissertation on a text of an early Christian heresy (Gnosticism), he became a lecturer at the University of Leipzig in 1874. Two years later, he was promoted to a professorship in church history. In 1879 he moved to Giessen and in 1886 to Marburg. From there he was called in 1888 to a professorship at the University of Berlin. Because of his liberal theological views, especially with respect to the validity of the historical Christian creeds, his appointment to the post at Berlin was opposed by the supreme council of the Evangelical Church of Prussia, but the opposition was overruled by Chancellor Otto von Bismarck and, on his advice, by the emperor William II; the latter had become emperor in 1888, the year of Harnack's appointment at Berlin.

Throughout his life, though maintaining academic appointments in theology and church history, Harnack was denied ecclesiastical posts. Nevertheless, he exercised broad influence in Protestant churches, because, through his masterful teaching and his solid learning, he won an enthusiastic following among his students, many of whom rose to positions of ecclesiastical leadership.

He also was highly effective as a historian, not only because he completely mastered the primary sources but also because he regarded himself as called upon to assume the royal function of a judge. "We study history," he said,

> in order to intervene in the course of history, and it is our right and duty that we do this, for if we lack historical insight we either permit ourselves to be mere objects placed in the historical process or we shall tend to lead people down the wrong way. To intervene in history—this means that we must reject the past when it reaches into the present only in order to block us. It means also that we must do the right thing in the present, that is, anticipate the future and be prepared for it in a circumspect manner.

Major works

In his voluminous writings, he brought to a culmination the interpretation of the Christian religion as a historical "development" as it had been taught by the earlier German biblical and historical theologians Johann Salomo Semler, Ferdinand Christian Baur, and Albrecht Ritschl. His purpose was to replace theological dogmatism by historical understanding. His chief works were *The History of Dogma* (3 vol., 1886–89; 4th ed. 1909); *The Mission and Expansion of Christianity in the First Three Centuries* (2 vol., 1902; 4th ed. 1924); *The History of Ancient Christian Literature* (3 vol.,

1893–1904). He was the chief editor of a critical edition of *The Greek-Christian Authors of the First Three Centuries* (1891–). In addition to these writings he published numerous monographs on the New Testament and on the doctrines and institutions of the ancient church. It was a signal honour that, though a theologian, he was asked to write *The History of the Prussian Academy of Sciences* in connection with the celebration of its 200th anniversary (3 vol.; 1900). Harnack revealed himself in this work as a comprehensive historian of modern learning. Indeed, he was a master of many fields of knowledge. Throughout his life, he concerned himself with the life and thought of the German Romantic poet Johann Wolfgang von Goethe (1749–1832) and devoted to him several profound studies.

His most popular book, *Das Wesen des Christentums* (1900; Eng. trans., *What Is Christianity?*, 1901, 15th ed. 1950), is a transcript of a student's stenographic record of a course of lectures, which were delivered in the University of Berlin.

In all these works it was Harnack's purpose to show how it happened that the gospel of Jesus, which in his view has nothing in common with authoritarian ecclesiastical statutes and doctrines, became embodied in the doctrines of the church. He also wanted to offer support for his conviction that if the gospel is to retain power in the modern world, it must be freed from its connection with the dogmas of God and Christ with which it became identified necessarily to survive in the Hellenistic world.

Holding three important positions at the height of his career, each one of which normally would have demanded a man's whole effort, Harnack had a profound impact on historical theological scholarship and was a major influence in the natural sciences. A professor of church history at the University of Berlin, he made important and creative contributions to historical and theological scholarship, in recognition of which he became a member of the Academy of Sciences in Berlin. In addition, he was the general director of the Prussian State Library and, after 1911, president of the Kaiser-Wilhelm-Gesellschaft (now called Max-Planck-Gesellschaft zur Förderung der Wissenschaften). Harnack secured the support of government and industry for this foundation and established research institutes in the natural and medical sciences.

Death and assessment

Harnack never left his position at the University of Berlin until he retired in 1921, at which time he was granted the title professor emeritus. He died on June 10, 1930. At the memorial service held in Berlin after his death (1930), the rector of the university characterized him as follows:

> He was of a noble aristocratic character; his outward distinction was softened by a generous, considerate, and kind disposition. In conversation, he never let one feel his superiority; on the contrary, he enhanced the self-confidence of the one who was speaking with him by rearranging in a most agreeable way whatever was said to him and putting it in such a form that the other took great delight in the thought he had expressed.

BIBLIOGRAPHY. AGNES VON ZAHN-HARNACK, *Adolf von Harnack*, 2nd ed. (1951), a beautiful biography written by his daughter; WILHELM PAUCK, *Harnack and Troeltsch: Two Historical Theologians* (1968); G. WAYNE GLICK, *The Reality of Christianity: A Study of Adolf von Harnack As Historian and Theologian* (1967).

(W.P.)

Harness and Saddlery

Harness and saddlery is the equipment for any animal used for packing, traction, or riding. "Harness" is the general term for the gear used for traction. "Saddlery" applies to articles traditionally produced by a saddler; *i.e.*, a person who makes, repairs, or sells riding saddles and related equipment, such as bridles, bits, and stirrups.

History. With the possible exception of the sail, harness represents man's first attempt to employ nonhuman power for his purposes. In pre-Columbian times, the New World knew no harness except for llamas and dogs carrying packs, dogs pulling travois on the North American plains, and dogsleds among the Eskimo. The Indian failure to develop harness may be due to the scarcity in America of animals anatomically suited to it and to the failure to observe systematically the effects of gelding (castration), which makes male animals sufficiently docile to be used for traction.

The yoke

The earliest Old World form of harness, which presupposes gelding, is the yoke resting on the withers (shoulders) of two animals, first used in the Near East, perhaps by 4000 BC, to attach a pair of oxen to a plow. By about 3000 BC the yoke was employed in Mesopotamia with onagers (Asian asses) to draw light wagons and chariots; and by about 2000 BC it was applied to horses, then newly arrived from Central Asia. Asses were used as pack animals in Egypt by 3000 BC, but there is no evidence that any animal was ridden until about 1300 BC.

The first use of the camel for labour is of very uncertain date. Presumably it is early, since the single-humped camel appears as a pack animal by 1000 BC in Mesopotamia. In the region north of the Altai Mountains, reindeer were used for traction earlier than horses, to judge from the fact that horses drawing the funeral sled of a chieftain buried at Pazyryk about 400 BC, which were sacrificed and buried with him, wore reindeer masks.

Bits. Both oxen and onagers were controlled by rings through the nose. The horse, however, appears from the beginning to have been governed normally by a bridle and bit, sometimes of bone but generally of metal. At first this was a simple bar with a ring at each end for reins, although by 1400 BC the jointed snaffle was known throughout the Near East. While there were many variations of both bar and snaffle bit, no important innovation appeared for 1,000 years; in the 4th century BC, Xenophon speaks of the curb bit. During the European Middle Ages the development of mounted shock combat demanded that the knight have absolute control of his charger, with the result that heavy and very severe curbs were produced. But no basic new idea in bits has appeared for the past 23 centuries.

Spurs. The simple prickspur is found in Celtic graves of about 400 BC. Although its use became habitual in Europe and eventually in the Americas, it seems not to have spread widely in the Orient or Africa. Rowel spurs, which replaced the single prick with a wheel of radiating points, first appeared in late 13th-century Europe.

Saddles. For at least 12 centuries after horses were first ridden, there were no saddles but only saddlecloths attached by surcingles or bellybands. These cloths were gradually supplemented by cushions or rolls that improved the rider's comfort but did not add greatly to his stability. Rigid saddles without stirrups appear in China during the Han dynasty (206 BC–AD 220) and are found in Roman Gaul in the 1st century AD, probably as an introduction from the barbarian world. The development of mounted shock combat required that the saddle become a large seat, giving maximum support to the knight. In early modern times, as shock combat in warfare and the tournament as sport went out of fashion, saddles steadily grew lighter. The decline of the saddle into the so-called English saddle of the late 19th and the 20th centuries is a sign of the technological obsolescence of the horse among those using such saddles. Herdsmen continued to use very substantial saddles.

The sidesaddle

The sidesaddle developed from an ancient packsaddle with a pannier, a board suspended on one side upon which the woman rider could rest her feet for support. By Frankish times the term *astraba* referred specifically to the footrest of such a saddle, and the word conveniently labelled the stirrup when it came from Central Asia.

During the Middle Ages and early modern times, the sidesaddle was varied in many ways from the basic device.

Stirrups. The first stirrups appeared as big-toe stirrups in India in the late 2nd century BC. These were diffused widely in the warm regions, where horsemen were barefoot, from Timor and the southern Philippines on the east to Ethiopia on the west. The wave of Indian culture that took Buddhism to China likewise carried the idea of the stirrup, which the booted Chinese, in their

chillier climate, expanded into the foot stirrup by the 5th century AD.

The foot stirrup reached Korea in the 5th century and Japan and some of the Central Asian countries in the 6th century. In 694 Muslim armies in northern Persia received the stirrup from Turkistan. Shortly afterward it was found in Byzantium, and by the early 730s it had arrived in the Frankish realm.

Before the support of the stirrup supplemented that of pommel and cantle, a mounted warrior could thrust his spear only with the strength of his own arm. With stirrups he could lay the spear at rest under the upper arm while his hand guided the blow that was now delivered by the force of his charging horse. The increase of violence was immense. Because it made possible mounted shock combat, the stirrup was the most significant invention in the history of warfare prior to gunpowder. In the early 730s Charles Martel saw the military potentialities of the stirrup, seized great areas of church lands, distributed them to retainers as endowment on condition that they serve him by fighting in the new manner, and thus instituted the feudal regime.

Invention of modern harness. The Bronze Age transfer of the yoke from oxen to onagers, horses, and similar animals was not successful for anatomical reasons: a horse or ass could be attached to such a yoke only by two straps running from the end of the yoke around the animal's neck and under its belly. The neck strap pressed on the windpipe and jugular vein in proportion to the pull of the animal, while the withers were so high a point of traction as to be mechanically inefficient. The modern harness, with its padded collar supporting rigid hames on the shoulders in a way that does not interfere with breathing or blood circulation, enables the animal to throw all its weight into pulling. It has been shown experimentally that a team of horses equipped with modern harness can pull a load four to five times heavier than it can pull with the ancient yoke harness.

Invention of the horse collar

Modern harness appears to have been introduced into Europe from Asia. China under the Hans developed a type of chariot drawn by a horse harnessed between lateral shafts, and the modern harness is depicted in frescoes of AD 500 in Kansu.

The first modern harness in Europe was probably seen about AD 800.

Animals or teams could not be harnessed tandem, or in sequence, until the 1st or 2nd century AD, when the first evidence of suitable harness is found in Gaul, Italy, and China. When this was later combined with horse collar and traces, horsepower became available for plowing and heavy hauling, although most historians believe that the use of horses in plowing was rare before the 12th century. In the 11th century the whiffletree, a horizontal transverse bar joining the two harnesses, helped to equalize the pull of traces on the load; and the pivoted front axle, known in Roman times, became common. By the 12th century the big four-wheeled wagon drawn by several teams of horses was revolutionizing land transport.

It was the ancestor of all later Western coaches and wagons, including the Conestoga wagons of the American frontier and the railroad coach (see WAGONS AND CARRIAGES; RIDING AND HORSEMANSHIP).

Modern equipment. Modern harnesses for riding and draft horses respectively are shown in Figures 1 and 2.

Bridles. The bridle, a harness of horses and mules, is a set of straps that makes the bit secure in the animal's mouth and thus ensures human control by means of the reins (see Figure 1). The upper portion of the bridle consists of the headpiece passing behind the ears and joining the headband over the forehead; the cheek straps run down the sides of the head to the bit, to which they are fastened; in the blind type of driving bridle the blinkers, rectangular or round leather flaps that prevent the animal from seeing anything except what lies in front, are attached to the cheek straps; the noseband passes around the front of the nose just above the nostrils; and the throatlatch extends from the top of the cheek straps underneath the head.

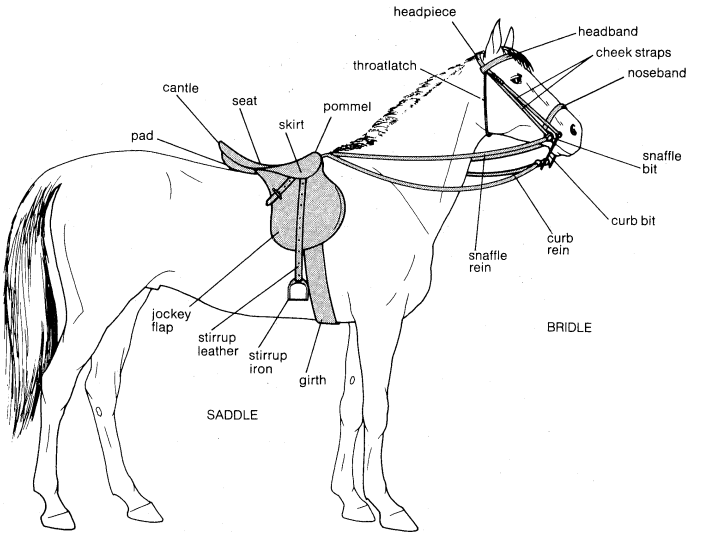


Figure 1: Nomenclature of a modern bridle and English saddle.
By courtesy of the American Saddle Horse Breeders Assoc. and W. W. Rouch & Co.

The martingale, sometimes used on riding horses, but not shown in the Figure, passes between the horse's forelegs with one end fastened to the saddle girth and the other to the head harness. It keeps the horse from throwing back its head.

The bit, the metal contrivance inserted in the mouth to which the reins are attached, is of three general types: the straight bar, snaffle, and curb. The snaffle has a jointed steel mouthpiece with straight cheek bars, the rings for the reins and cheekpieces of the headstall being fixed in the bars at the junction with the mouthpiece. The bars prevent the horse from pulling the bit through its mouth. The snaffle without bars is termed a bridoon. The curb bit is one to which a curb chain or strap is generally attached, fastened to hooks on the upper ends of the cheek bars of the bit and passing under the horse's lower jaw in the chin groove. The reins are attached to rings at the lower ends of the cheek bars, the leverage thus pressing the curb chain against the jaw. The mouthpiece of the curb bit is unjointed and in the centre commonly has a port; *i.e.*, a raised curve, allowing liberty for the tongue and bringing the pressure on the base of the horse's jaw. The curb bit and the bridoon can be used together with separate headstalls and reins, and there are many combination bits.

Saddles. The skeleton of the riding saddle is composed of the tree, or framework, the parts of which are the pommel, or head, the projection that fits over the animal's withers, and the sidebars that curve around to the back to form the cantle, or hindbow. The rigid parts are almost always covered with leather. The stirrup bars are fastened to the tree. On either side of the saddle are a leather skirt and hanging flaps to protect the rider from the mount's sweat. In saddles to be used with unbroken or unruly horses, the front of the flap is often padded to form a knee roll.

The saddles of Asia, Africa, and continental Europe tend to be heavy and high. But the English saddle, removed from all utilitarian purpose, is designed for extreme lightness for sport, racing, and fox hunting; the pommel and cantle are greatly reduced, and provision is made for clinging with the knees in jumps.

The English saddle

The saddles of Latin America and of western North America are diversified but are derived largely from equipment developed in the cattle industry of Spain during the Middle Ages. On the pommel is a sturdy metal horn to which the lariat, or lasso, is tied when cattle are being roped. The cantle is high to prevent the rider from slipping off the seat when the horse sits back on its haunches at the moment of roping. The saddle is made secure by one or two surcingles called cinches or girths; these are held to the tree rings by latigos, long leather straps. The sides and front of the stirrups are covered by

leather taps (Spanish *tapaderas*), a casing that prevents them from snagging in brush.

Harness. Many types of harness are used in different parts of the world for oxen, zebus, asses, reindeer, elephants, camels, water buffalo, yaks, and dogs. In the basic Occidental harness for horses and mules (Figure 2),

By courtesy of the Orrville Leather and Hardware Corp.

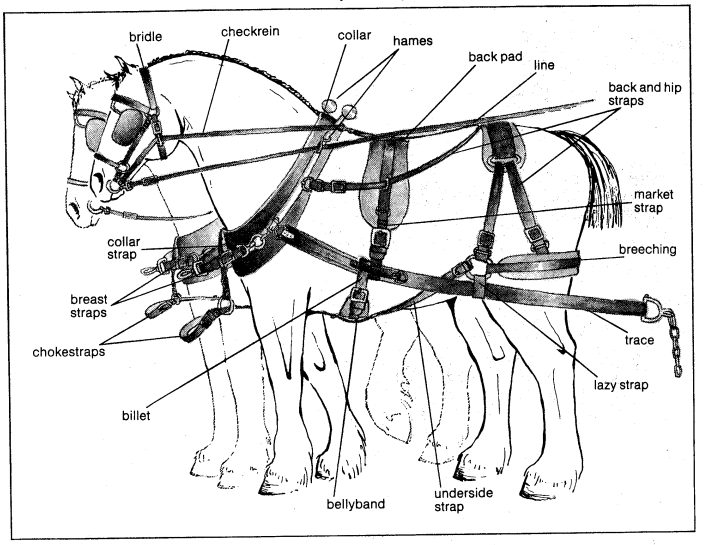


Figure 2: Typical harness for draft horses.

a leather collar, heavily padded, passes over the head and rests firmly on the shoulders; two hames, rigid curved pieces of metal, rest on this collar and are fastened at top and bottom by hame straps. To the hames are attached the traces, straps that pass along the animal's sides and are connected to the end of the whiffletree (also called whippletree, singletree, or swingletree), which is connected to the load by a centre link (see Figure 2).

When the animal is harnessed between shafts, the shafts are usually supported by a back pad; the pad is a narrow leather cushion resting on the back, attached to the shaft by straps and held in position by a girth, or bellyband, as well as by a backband and crupper, a loop strap passing under the tail. Reins pass through terrets, or rings, on the hames and pad. The other harness on the animal's hindquarters consists of the breeching, straps passing around behind the haunches to help in backing, braking, and stopping the vehicle, and the hip strap, fastened to the breeching and passing over the hindquarters.

The checkrein (bearing rein), a device sometimes used to add to the elegant arch of a horse's neck, consists of a separate bridoon bit with the reins passing through rings on the collar and then attached to a hook on the pad.

BIBLIOGRAPHY. CHARLES SINGER *et al.* (eds.), *History of Technology*, 5 vol. (1954–58, reprinted 1965), scattered but substantial references throughout; RICHARD LEFEBVRE DES NOETTES, *L'Attelage: le cheval de selle à travers les âges*, 2 vol. (1931), a classic work that first drew attention to the significance of the padded horse collar and traces; P.A. ROLLINS, *Cowboy: An Unconventional History of Civilization on the Old-Time Cattle Range*, rev. ed. (1936), on western U.S. saddles and bridles; A.D.H. BIVAR, "The Stirrup and Its Origins," *Oriental Art*, New Series, 1:61–65 (1955), a good source for the subject stated.

Hārūn ar-Rashīd

Hārūn ar-Rashīd, the fifth caliph of the ʿAbbāsid dynasty, though he ruled the Islāmic Empire of the caliphate from 786 to 809, when it was at its zenith, is now celebrated mainly for the colourful and romanticized picture of himself and his court in Baghdad as set out in the tales known as *The Thousand and One Nights* (*The Arabian Nights Entertainments*).

Family and early life

His father was al-Mahdī, the third ʿAbbāsid caliph (ruled 775–85). His mother, al-Khayzurān, a former slave girl from Yemen, was a woman of strong personality who greatly influenced affairs of state in the reigns of her husband and sons. The elder of these, al-Hādī, was four when Hārūn was born at Rayy, near modern Tehrān, Iran, in February 766 (or March 763). The princes were brought up in the court at Baghdad and educated in the Qurʾān (the holy book of Islām), poetry, music, anecdotes about the Prophet Muḥammad, early Islāmic history, and current legal practice. Hārūn had as tutor Yaḥyā the Barmakid, a loyal supporter of his mother. In 780 and 782 Hārūn was nominal leader of expeditions against the Byzantine Empire, though the military decisions were doubtless made by the experienced generals accompanying him. The expedition of 782 reached the Bosporus, opposite Constantinople, and peace was concluded on terms favourable to the Muslims. For this success Hārūn received the honorific title of ar-Rashīd, "the one following the right path," and was named second in succession to the throne and appointed governor of Tunisia, Egypt, Syria, Armenia, and Azerbaijan, with his tutor Yaḥyā acting as actual administrator. These moves were presumably engineered by al-Khayzurān and Yaḥyā. The two are even said to have induced al-Mahdī to make Hārūn his immediate successor, but al-Mahdī died in August 785 without officially changing the succession. Al-Hādī became caliph and Hārūn acquiesced. When al-Hādī died mysteriously in September 786, rumour suggested that al-Khayzurān was behind the death, because he had resisted her domination.

Hārūn ar-Rashīd thus became caliph on September 14, 786, succeeding to the rule of an empire reaching from the western Mediterranean to India. He made Yaḥyā the Barmakid his vizier, or chief minister. With Yaḥyā were associated his sons al-Faḍl and Jaʿfar, for the vizier at this period was not only an initiator of policy but also had attached to himself a corps of administrators to carry out his decisions. Al-Khayzurān had a considerable influence over the government until her death in 789. Thereafter until 803 the Barmakids largely controlled the empire, but the caliph was not wholly dependent on them, since certain offices of state were held by other men.

The reign was one of much internal trouble. At various times there were revolts for local reasons in Egypt, Syria, Yemen and several eastern provinces, but the central government was strong enough to quell these and restore order. Ifrīqīyah (or Tunisia), after having had a series of incompetent governors, was given in 800 to Ibrāhīm ibn al-Aghlab, who agreed to make a substantial yearly payment to Baghdad in return for semi-independent status. This was immediately advantageous to Hārūn financially but was the beginning of the loss of power by the caliphs, for the Aghlabid family continued to rule the province for over a century without interference from Baghdad, and similar status was granted to other regional dynasties. Though the revolts fill the pages of the historians, much of the empire was peaceful most of the time. This led to a great development of industry (textiles, metal goods, paper, etc.) and to an expansion of trade. The resulting prosperity made possible the concentration of vast wealth in the hands of the caliph and leading men and women of the empire.

ʿAbbāsid wealth under Hārūn

The fabulous descriptions of Hārūn and his court in *The Thousand and One Nights* are idealized and romanticized, yet they had a considerable basis in fact. Untold wealth had flowed into the new capital of Baghdad since its foundation in 762. The leading men, and still more their wives, vied in conspicuous consumption, and in Hārūn's reign this reached levels unknown before. His wife Zubaydah, herself a member of the ʿAbbāsid family, would have at her table only vessels of gold and silver studded with gems. Hārūn's palace was an enormous institution, with numerous eunuchs, concubines, singing girls, and male and female servants. He himself was a connoisseur of music and poetry and gave lavish gifts to outstanding musicians and poets. The brilliant culture of the court had certain limits, however, since, apart from philology, the intellectual disciplines were in their infancy in the Arabic world. There was also a rougher and

more sombre side. Instead of listening to music, Hārūn might watch cocks and dogs fighting. As caliph he had power of life and death and could order immediate execution. In the stories of his nocturnal wanderings through Baghdad in disguise, he is usually accompanied by Masrūr the executioner as well as friends like Ja'far the Barmakid and Abū Nuwās, the brilliant poet.

The less pleasant aspects of Hārūn's character are highlighted by the fall of the Barmakids, who for over 16 years had been mainly responsible for the administration of the empire, and who had provided the money for the luxury and extravagance of the court. Moreover, Ja'far the Barmakid had become Hārūn's special friend, so that gossip spoke of a homosexual relationship. Gossip also alleged that Hārūn had arranged that Ja'far should secretly marry his sister 'Abbāsah, on condition that he did not consummate the marriage, but that Ja'far fell in love with her, and she had a child. Whether in anger at this or not, Hārūn had Ja'far executed on January 29, 803. The other members of the family were imprisoned and their goods confiscated. Modern historians reject this gossip and instead suggest that Hārūn felt dominated by the Barmakids and may even have coveted their wealth. Moreover, diverse interests within the empire were being attracted to two opposing poles. On the one side were the "secretaries," or civil servants, many Persians, and many men from the eastern provinces; on the other side were the religious scholars (*'ulamā'*), many Arabs, and many from the western provinces. Since the Barmakids favoured the first group of interests and the new vizier, al-Faḍl ibn ar-Rabī', favoured the second, it is likely that this political cleavage was involved in the change of ministry.

The fall of the Barmakids

The struggle between the two groups of interests continued for at least half a century. Hārūn recognized its existence by assigning Irāq and the western provinces to his son al-Amīn, the heir apparent, and the eastern provinces to the second in succession, his son al-Ma'mūn. The former was son of the Arab princess Zubaydah and after 803 had al-Faḍl ibn ar-Rabī' as tutor. Al-Ma'mūn was son of a Persian slave girl and after 803 had as tutor a Barmakid protégé, al-Faḍl ibn Sahl. Hārūn has been criticized for so dividing the empire and contributing to its disintegration, for there was war between his two sons after his death; but it may well be that by making the cleavage manifest, he contributed to its eventual resolution after 850.

As vizier, al-Faḍl ibn ar-Rabī' lacked the efficiency of the Barmakids, and Hārūn's personal decisions may have had more weight. There were further successful operations against the Byzantine Empire, but in the autumn of 808, while on his way to deal personally with a serious two-year-old revolt in Khorāsān (in Iran), Hārūn fell ill at Ṭūs (near modern Meshed) and died there on March 24, 809. Al-Amīn succeeded him as caliph.

Hārūn was neither a great ruler nor a man of prepossessing character, though he was a lavish patron of the arts. He owes his fame to the wealth and luxury of his court, surpassing anything previously known, and to his place in Arabic legend.

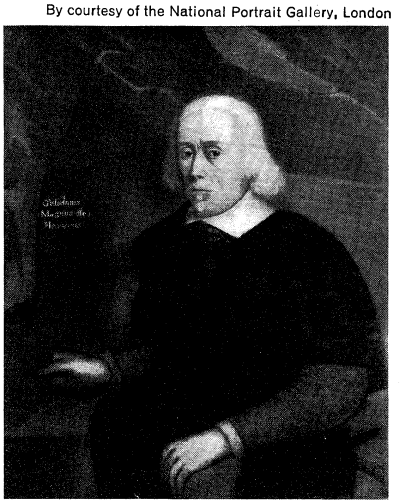
BIBLIOGRAPHY. There is no standard biography. H. ST.J.B. PHILBY, *Harun al-Rashid* (1933), is a popular work of limited merit. More scholarly is NABIA ABBOTT, *Two Queens of Baghdad* (1946), describing the political roles of Hārūn's mother and his wife Zubaydah. The fullest account in general histories is that by P.K. HITTI in his *History of the Arabs*, 8th ed. (1964).

(W.M.W.)

Harvey, William

A leading English physician of the first half of the 17th century, William Harvey achieved fame by his conclusive demonstration of the true nature of the circulation of the blood and the function of the heart as a pump. Functional knowledge of the heart and the circulation had remained almost at a standstill ever since the time of the Greco-Roman physician Galen—1,400 years earlier. Harvey's courage, penetrating intelligence, and precise methods were to set the pattern for research in biology and other sciences for succeeding generations, so that he shares with William Gilbert, investigator of the magnet, the credit for initiating accurate experimental research throughout the world.

By courtesy of the National Portrait Gallery, London



Harvey, oil painting by an unknown artist after an etching attributed to R. Gaywood. In the National Portrait Gallery, London.

William Harvey was born on April 1, 1578, at Folkestone, Kent. His father, Thomas Harvey, was a prosperous businessman and a leading citizen of the small town. William, the eldest of nine children, was the only one to achieve special distinction in his career, but all his brothers were successful in business or at the royal court in London and among them amassed considerable wealth.

Career as physician and scientific innovator. Little is known of William Harvey's boyhood in the Kentish countryside. During the years 1588 to 1593 he was at the King's School attached to the cathedral at Canterbury. In his 16th year Harvey entered Gonville and Caius College, Cambridge, where he was awarded a scholarship in 1593. Although Harvey attended Caius College because of its special interest in educating doctors, his training was grossly inadequate. He was absent from the university for the greater part of his last year (1598–99) because of illness—probably malaria—but had received the B.A. degree in 1597. Determined, in spite of his experience at Cambridge, to continue with medical training, he began a two-and-a-half-year course of study at the University of Padua, reputed to have the best medical school in Europe. His teacher was a celebrated anatomist, Hieronymus Fabricius ab Aquapendente, and it was in the now-famous oval Anatomy Theatre, still to be seen at the university, that Harvey first recognized the problems posed by the function of the beating heart and the properties of the blood passing through it.

Early theories of the heart and circulation

From the time of Aristotle in the fourth century BC it had been widely believed that the blood vessels contained both blood and air. Galen, the Greco-Roman physician, in the second century AD proved that the arteries contained only blood but still believed that air entered the right side of the heart from the lungs. There was a general belief that the movement of the blood was by ebb and flow, an analogy being found in the movement of the sea. Galen's view on this are difficult to assess with exactitude, but it is apparent that he, like everyone else, had no conception of a circular movement of the blood, leaving the heart by one set of vessels, the arteries, and returning to it by another set, the veins. The main propulsive force initiating this oscillatory movement was supposed to be derived from a contracting of the arterial system, rather than by a pumping action of the heart. The blood in the veins was believed to be formed in the liver, passing to the right auricle (*i.e.*, one of the two upper chambers of the heart), and from there to the right ventricle (one of the two lower chambers), to make its way through holes in the septum, or partition, to the left

side, where it met with blood from the arteries, which was mixed with air derived from the lungs. This was the extent of man's knowledge about the movement of the blood until 14 centuries later. Early in the 16th century the idea of a pulmonary circulation—that is, a circular motion of blood through heart and lungs—began to occur to some anatomists. In addition, the presence of a perforated septum was beginning to be questioned. In the middle of the 16th century a great anatomist, Andreas Vesalius, also working at Padua, first established accurate knowledge of human anatomy but was less interested in function. Several other medical investigators refined the anatomical knowledge of the heart. Realdus Columbus of Cremona, working as assistant to Vesalius, developed the idea of a pulmonary circulation, and this was made more definite by his pupil, Andreas Caesalpinus, though they still thought that the blood was distributed to the body by the great veins and their branches. Fabricius had a special interest in the anatomy of the veins and first described the system of valves found in them, but he was quite ignorant of their true function. In brief, there existed no convincing explanation of how the heart worked, and Harvey's logical mind remained unsatisfied.

His 28 months at Padua are only meagrely documented, but it is clear that he was outstanding among the students of his year. After receiving his diploma as doctor of medicine of Padua in April 1602, he returned to England. By the standards of the time he was fully trained in anatomy, the simpler functions of the human body, and in therapeutics based on the writings of Aristotle. He had had some clinical experience in the hospitals of Padua and Venice and was entitled to obtain a fellowship of the College of Physicians in London after passing through the preliminary stage of candidate for the higher qualification. At his first oral examination, in May 1603, he was given limited permission to practice medicine, but only after further examinations in April and August 1604 was he fully licensed to practice within the jurisdiction of the college—that is, in the London area.

Shortly after his return to England, Harvey married Elizabeth Browne, daughter of Lancelot Browne, physician to King James I and his queen, and a senior fellow of the college. The couple set up house in the parish of St. Martin's by Ludgate, not far from the College of Physicians; and, backed by Browne, Harvey then tried to obtain the appointment of physician to the Tower of London, where a number of distinguished men were imprisoned. Though he failed in this attempt, in 1607 he finally obtained a fellowship of the College of Physicians, which entitled him to seek an appointment as physician to one of the two great hospitals then serving London—St. Bartholomew's and St. Thomas's. It may have been through his brother John, who had obtained employment in the King's household, that early in 1609 the King gave Harvey a recommendation for an appointment at St. Bartholomew's, which was conveniently near his house in St. Martin's. He was given the post of assistant physician, and, when the physician died in the summer of that year, Harvey succeeded him. The hospital at that time had about 200 beds for patients in 12 wards, and Harvey's duties consisted in attending in the hall of the hospital on at least one day a week throughout the year to see the patients and prescribe for their treatment and at any other time when specially needed. The physician was usually expected to live within the hospital precincts, but the rule was waived for Harvey since he lived not far away. He received an annual salary of £25 with £2 extra for his livery and a further £8 since he did not use the official residence. His colleagues were three surgeons and an apothecary in charge of the dispensary.

Appointment to St. Bartholomew's

Harvey held this office for 34 years until 1643 when he was displaced for political reasons by Oliver Cromwell's party, then in power in London. These years saw the development and culmination of his active career as physician and scientific innovator. He developed a large private practice, attending many of the most distinguished citizens, including Sir Francis Bacon—and, about 1618, was made physician extraordinary to King James I, thus becoming a colleague of Sir Theodore Turquet de Mayerne, the senior court doctor. There can be no doubt that Harvey was for many years one of the most widely trusted doctors in England, although his unorthodox views on the circulation of the blood did injure his practice after their publication in 1628. Invariably courteous and regarded with affection and respect by his colleagues, he conducted his practice with common sense and honesty. Though advanced in his ideas of anatomy and physiology and scientific in his methods of research, he was inevitably conservative in his use of remedies. Very few potent drugs were known in his time, and accurate diagnosis was, more often than not, impossible, so that he never escaped from the influence of Aristotle, in whose principles he had been trained. He was the great protagonist of experimental biology but did not apply himself to this form of originality in therapeutics.

At the time of the King's last illness in 1625, de Mayerne was out of the country, and Harvey led the team of doctors in attendance. After the King's death it was rumoured that his favourite, the duke of Buckingham, had contributed to the fatal outcome by applying remedies not approved by the doctors. He was actually accused of having poisoned the King, and an inquiry was ordered by Parliament in 1626. Harvey was the most important witness of several who contributed to exonerating the Duke from any direct responsibility. Charles I, the new king, continued Harvey's appointment as his personal physician and gave him a special award for the care he had given the last King. Charles's health remained good until the day of his execution, so that he rarely had need to consult his doctor. Nevertheless, Harvey became his close friend and was always in attendance on his journeys, such as his state visits to Scotland in 1633 and 1638. The King helped Harvey's scientific researches by putting the deer in the royal parks at his disposal, and he delighted in showing the King anything of curiosity or scientific interest. At the same time, Harvey took his full share in the affairs of the College of Physicians, being constantly present at the meetings of the fellows and occupying all the official positions in the college hierarchy except that of president. His duties at court would not have allowed him to fill this position during his active years, and when it was offered to him in 1654 he was too old and ill to be able to accept. Yet it is clear from the college records that Harvey was always the man to whom his colleagues turned for advice in resolving their difficulties. The physicians at this time had precedence over the other branches of the profession, and Harvey had a prominent part in maintaining this ascendancy over the surgeons, obstetricians, and apothecaries whenever they became restive under the authority of the college.

Scientific research

In spite of Harvey's activity in medical practice and college affairs, he spent much time in scientific research from the time of his return to England in 1604 until the beginning of the Civil War in 1642. His interest lay primarily in elucidating the facts of the movement of the heart and its relation to the circulation of the blood. Fabricius at Padua had opened his eyes to the value of comparative anatomy, and he was tireless in dissecting every kind of living thing, from insects, earthworms, reptiles, birds, and mammals up to man himself. He seized every opportunity to increase his knowledge of pathology through postmortem examinations and was an acute clinical observer of his patients, not omitting their psychology. Most of his scientific papers were destroyed by parliamentary soldiers during the Civil War, so that there is now no direct evidence of his methods. On the other hand, his lecture notes used from 1616 onward survive. In 1615 he was appointed to a college lectureship intended to cover all parts of medical knowledge, though each lecturer modified the course to suit his own interests. Harvey's manuscript, now in the British Museum, was entitled "Lectures on the Whole of Anatomy." It is written in a very bad hand in mixed Latin and English, and it is incomplete, lacking any account of the skeleton, the sense organs, and other systems. The systematic anat-

omy is enlivened by many references to comparative anatomy, morbid anatomy, and clinical observations, even naming individuals whom he had treated. It is evident that he wrote these notes before he had come to any conclusions about the circulation of the blood, so that they contain nothing that seriously questioned the authority of Galen. The only reference to his novel views is on a leaf inserted some years later, probably after 1628. Harvey held this lectureship until 1656.

Discovery of circulation of the blood. It is evident from his writings that Harvey reverenced Aristotle, even though he had to dismiss some of his teachings as absurd. He also valued the views of Galen, his predecessor in experimental physiology, and enlisted his support whenever he could do so. Yet Harvey depended essentially on reasoning from his own observations and experiments for proof of his contentions. During the 12 years after 1616, Harvey may have introduced some novelties into his lectures; but, by his own assertions, he demonstrated the results of his researches to his friends privately at the college. Finally he published his book, *De Motu Cordis et Sanguinis in Animalibus* (*On the Motion of the Heart and Blood in Animals*), in 1628, a slender volume that established concisely and conclusively a new and unorthodox demonstration of the facts about the circulation of the blood. With Galen's help, he first disposed finally of the idea that the blood vessels contained air. He then elucidated the function of the valves in the heart in maintaining the flow of blood in one direction only when the ventricles contracted—on the right side to the lungs and on the left to the limbs and viscera. He proved that no blood passed through the septum, separating the two ventricles, and he explained the purpose of the valves in the larger veins in directing the return flow toward the heart. He showed that blood was expelled from the ventricles during contraction, or systole, and flowed into them from the auricles during expansion, or diastole. He proved that the arterial pulse was due to passive filling of the arteries by the systole of the heart and not by active contraction of their walls. He explained the purpose of the pulmonary circulation from the right ventricle through the lungs and back to the left auricle and ventricle. His only failure was in not demonstrating the connection of the arterial and venous systems in the tissues of the limbs by means of the smallest, or capillary, vessels. These he was unable to see, having no adequate form of microscope at his disposal. He was the first scientist to employ measurement of the content of the chambers of the heart and estimation of the total amount of blood in the body—that is, quantification.

Fame and criticism

Harvey's book made him famous throughout Europe, though the overthrow of so many time-hallowed beliefs attracted virulent attacks and much abuse from those who did not wish to believe the plain evidence of their senses. He refused to indulge in controversy and made no reply until he published a small book in 1649 answering the criticisms of a French anatomist, Jean Riolan. In this he reiterated some of his former arguments and utterly demolished Riolan's objections.

In 1636 King Charles dispatched a diplomatic embassy to the Holy Roman emperor Ferdinand II at Regensburg, Germany, in an attempt to establish the claim of his nephew, Prince Charles Louis, as Elector Palatine. Harvey was chosen as doctor to the mission and spent ten adventurous months of travel by land and water through territories ravaged by the Thirty Years' War, extending his journey by visits to Vienna, Prague, Venice, Rome, and Naples. At Nuremberg Harvey had a historical encounter with Caspar Hofmann, professor of medicine at the University of Altdorf, whom he attempted, at a public demonstration, to convince of the truth of his doctrine of the circulation. Though he did not succeed, Harvey behaved with great dignity and good temper in the face of obstinate blindness to demonstration of the facts.

At the start of the Civil War in 1642, Harvey was with the King and was in charge of the two princes, Charles and James, in the early stages of the Battle of Edgehill. When the King established his headquarters soon afterward at Oxford, Harvey remained with him and was given the position of warden of Merton College in 1645. Here he resumed his work on the development of the chick in hens' eggs and first met John Aubrey, antiquary and gossip, who afterward left a revealing account of Harvey in his *Brief Lives*. When the defeated King fled from Oxford to surrender himself to the Scots, Harvey joined him for a time at Newcastle but was forced to leave the King when he was handed over to the parliamentary army and was not allowed to go to him when he was imprisoned in the Isle of Wight. Harvey had never been much interested in politics but felt a deep personal regard for the King and after his execution in 1649 was a broken and unhappy man.

Yet two years later he published his second great book. After the publication of *De Motu Cordis*, the main achievement of Harvey's life, he had continued active research into the difficult subject of reproduction in animals. This led in 1651 to the publication of *De Generatione Animalium* (*Anatomical Excitations, Concerning the Generation of Living Creatures*, 1653) through the persuasions of his younger friend Sir George Ent, a fellow of the college. The book contains much of historical and scientific interest, but Harvey's thought was greatly influenced by Aristotle. The book is mainly concerned with the development of the chick in hens' eggs, and Harvey insisted throughout that in all living things the origin of the embryo is to be found in the egg. He investigated also the embryology of deer, rejecting Aristotle's notion that menstrual blood played any part in the formation of the fetus; he also questioned whether or not semen had any influence. Having no microscope, he could not see the spermatozoa, which were not demonstrated until 1686 by Leeuwenhoek working in Holland with stronger lenses. Harvey remained uncertain of how fecundation of the ovum was accomplished and even suggested that it was by a kind of infection resembling the origin of infectious diseases. Aristotle had originated the theory of gradual formation of the embryo, part by part, as opposed to the idea of preformation, meaning that all the parts arose in miniature at the same time. Harvey agreed with Aristotle and crystallized the belief in the term epigenesis, though to him its meaning was extremely simple compared with all that is implied by it at the present time. Aristotle believed in the principle of spontaneous generation of primitive organisms; it is probable that Harvey did not support this belief, but his statements are equivocal, and his position remains uncertain.

Harvey's brothers had been successful merchants, and their advice, coupled with his skill as a doctor and his naturally austere habits, enabled him to accumulate considerable wealth. But he had become old and ill. He had met with so much opposition and disbelief that his passionate desire to establish scientific truth was partially unsatisfied. In his last years under Cromwell's Protectorate, he was regarded as a political "delinquent" owing to his long association with King Charles and was forced to spend most of his time lodging in one or another of his brothers' houses outside London. Though he corresponded with many distinguished foreign doctors, he was reluctant to engage in any further scientific research, saw few patients, and took little part in the affairs of the College of Physicians. He showed his regard for the fellows by giving them a new college building in 1652 with a library containing his own collection of books and presumably any remaining manuscripts. This was in use for less than 14 years, being destroyed in the Great Fire of London in 1666, so that very few of his books have survived to the present day. He suffered severe pain from gout and kidney stones and described himself to a correspondent as "not only ripe in years, but also a little weary and entitled to an honourable discharge" from further scientific argument. His last illness was brief. He awoke one morning partially paralyzed and unable to speak, probably owing to a cerebral thrombosis. He died in his 80th year on June 3, 1657, probably in his brother Eliab Harvey's house at Roehampton. He was buried in the family vault at Hempstead, an Essex village 50 miles (80 kilometres) from London. In 1883 he was reburied in a marble sar-

Last years

cophagus in the Harvey Chapel there, near a marble bust by Edward Marshall. This is a lifelike image of Harvey, better, probably, than any of the existing portraits of him in old age.

BIBLIOGRAPHY. SIR GEOFFREY KEYNES, *The Life of William Harvey* (1966), a full and definitive biography based on examination of all available contemporary sources, documented, illustrated, and with eight appendixes; *A Bibliography of the Writings of Dr. William Harvey, 1578–1657*, 2nd ed. (1953), a detailed account of all Harvey's books and of where they may be found; *The Portraiture of William Harvey* (1949), a catalog of all known pictures, genuine and spurious, with reproductions; J.G. CURTIS, *Harvey's Views on the Use of the Circulation of the Blood* (1915), an early study of the position of Harvey's work in the history of the knowledge of human physiology; GWENETH WHITTERIDGE, *The Anatomical Lectures of Harvey: Praelectiones* (1964), a reliable transcription of Harvey's lecture notes with a full discussion and interpretation; *William Harvey and the Circulation of the Blood* (1970), an important study of the growth of Harvey's ideas; A.W. MEYER, *An Analysis of the De Generatione Animalium of Harvey* (1936), a discussion of Harvey's second major publication, a work on animal reproduction and development; W. PAGEL, *William Harvey's Biological Ideas* (1967), a well-documented historical analysis of Harvey's physiological and embryological ideas.

(G.L.K.)

Haryana

Haryana, a state of the northern region of the Indian Union, was constituted November 1, 1966, as a result of the partition of the former state of Punjab into two separate states—Punjabi-speaking Punjab and Hindi-speaking Haryana. It is bounded on the northwest by Punjab, on the northeast by Himachal Pradesh, on the east by Uttar Pradesh and the union territory of Delhi, and on the south by Rājasthān.

The reorganization of the former Punjab followed agitation by the Sikhs for a Punjabi Suba (Punjabi-speaking province); but it also met, substantially, the aspirations of the people of the Hindi-speaking region of Punjab for a Vishal Haryana (Greater Haryana). The name means "the abode of God," from Hari (the Hindu god Viṣṇu) and *ayana* (home), although it has also been suggested that the derivation may be from the word *hari* ("green"), denoting the fertility of the countryside.

The area of the state is 17,010 square miles (44,056 square kilometres). Its population in the early 1970s was about 10,000,000. The headquarters of the state government are at Chandīgarh, the modern capital of Punjab.

History. The region now known as Haryana was the birthplace of the Hindu religion and the civilization of the Indus Valley. There, the first hymns of the Aryans were sung, and the most ancient manuscripts were written. On the battlefield at Kurukṣetra, Lord Kṛṣṇa delivered to the warrior Arjuna the teachings that are contained in the *Bhagavadgītā*, venerated in India as the highest code of ethics.

The early Aryan settlers

The earliest known settlers of Haryana were Aryans. Lying athwart the route of overland invasions into India, it underwent successive waves of migration from the time of Alexander the Great (326 BC) and was the scene of many famous battles of Indian history, including the battles of Pānīpat in 1526 (at which the Mughal leader Bābur defeated Ibrāhīm Lodī to end Muslim domination), 1556 (at which the Afghan forces were defeated by the army of Mughal emperor Akbar), and 1761 (at which the Afghan shāh decisively defeated a Marāthā army, strengthening British control in Bengal) and the Battle of Karnāl in 1739 (at which Nāder Shāh of Persia dealt a blow to the crumbling Mughal Empire).

The area included in the present state was ceded to the British East India Company in 1803. In 1832 it was transferred to the then North-Western Provinces (Uttar Pradesh). In 1858 Haryana became a part of Punjab, remaining as such after the partition of India and Pakistan in 1947. The demand for a Haryana province, however, was raised even before India's independence in 1947. Lala Deshbandhu Gupta and Asaf Ali, prominent figures in the national movement, advocated a separate state of Haryana. Sri Ram Sharma, a veteran freedom fighter, had headed a Haryana Development Committee to focus attention on the concept of an autonomous state. The demand for unilingual states by Sikhs and Hindus gained momentum in the early 1960s, and Haryana became the 17th state of the Indian Union in accordance with the recommendations of the Punjab Reorganization Commission and the provisions of the Punjab Reorganization Act of 1966.

Physical geography. Most of Haryana lies on the Indo-Gangetic Plain, but in the northeast there is an extension of the low, sub-Himalayan Siwālik Hills. The state, on the whole, is flat, the average height ranging from 700 to 900 feet (210 to 270 metres) above sea level. It is drained by one perennial river, the Yamuna (Jumna), which does not flow in the state but forms its eastern border with Uttar Pradesh. There are numerous seasonal streams, the most important being the Ghaggar, which flows out of the Siwālik Hills and forms its northern border with Punjab.

Rainfall is scanty in most areas. Although the state has a system of canal irrigation, there are chronic drought-prone areas and occasional floods from the tributaries of the Yamuna and the Ghaggar. The climate is hot in summer and markedly cold in winter. The maximum temperature in summer (May–June) goes up to 114° F (46° C). The minimum temperature of 28° F (−2° C) occurs in January.

The soils are deep and fertile except in the eroded lands of the northeast and in the southwest where, bordering the Thar Desert in Rājasthān, they are sandy. Out of a total area of 10,900,000 acres, 9,600,000 acres are classified as arable. Canal irrigation is available for only about 3,200,000 acres, or a little more than one-third of the arable land.

The land, as indicated above, is mostly arid or semi-arid, and the natural vegetation, although little remains, reflects that fact, the flora of the plains resembling that of Iran, Saudi Arabia, and North Africa. The largest indigenous trees are the shisham (*Dalbergia sissoo*) and the kikar (*Acacia arabica*).

Dominance of Hindus in the population

Population. *Composition.* Haryana was constituted as a Hindi-speaking state. According to the 1971 census, out of a total population of about 10,000,000, Hindus constituted 89 percent. Other principal religious groups were Muslims (4 percent), Sikhs (6 percent), and Jains (less than 1 percent).

Languages and distribution. The principal language spoken is Hindi. The other important languages are Punjabi and Urdu.

More than 80 percent of the population is rural, living in over 7,000 villages. Only about 18 percent of the total population is classified as urban, spread over 60 cities and towns. Ambāla, together with its army cantonment, is the largest town in the state, with a population of more than 186,000. It is followed in size by Rohtak, with a population of almost 124,760. Other towns are Karnāl, Hissār, Pānīpat, Farīdābād, Bhiwāni, and Yamunānagar. The overall population density of the state is more than 590 persons per square mile.

Administration. The governor, appointed by the president of India, is the head of the state. He is aided and advised by a Council of Ministers responsible to the state Legislative Assembly (Vidhan Sabha). The 81-member assembly had its third election in less than six years in 1972. Normally, election to the assembly is for a five-year term. Political defections, however, brought about the dissolution of the first assembly, elected in 1967. A new assembly was elected in May 1968, but the ruling party decided on dissolution for a fresh mandate from the people, coinciding with assembly elections in neighbouring Punjab and several other states. The state has a common High Court with Punjab and also a common Public Service Commission that recruits all government employees.

The state is divided into seven districts—Hissār, Rohtak, Gurgaon, Karnāl, Ambāla, Jīnd, and Mahendargarh. The system of village self-government (*pañcāyat raj*) has been extended to all of the 7,039 villages. Administration,

as is the case in the rest of the country, is oriented toward development in all sectors.

Social conditions. With education getting a high priority in the development program, there has been a large quantitative expansion in school enrollment in Haryana; however, as in the northern states in general, girls' education has not made much headway. The literacy rate is lower than in the western or southern states of India. About 55,000 students attend colleges. The state has a teaching college and a residential university (founded 1956) at Kurukshetra, which also has a regional engineering college. Panjab University, in Chandīgarh, grants degrees through affiliated colleges of which there are six in Rohtak, five in Karnāl, seven in Ambāla, and four in Hissār. Hissār is the seat of the Haryana Agricultural University.

Haryana has a network of district and subdivisional hospitals and primary health centres. The Family Planning Programme is making headway. In more than half the villages, drinking water is acutely scarce. Eighteen percent of the population belonging to scheduled castes (former "untouchables") and other backward classes are covered by schemes of assistance in education and technical guidance and for housing.

Economy. The economy is agricultural and, despite its pockets of chronic drought and scanty rainfall, Haryana is one of the few states in which substantial investments in agriculture, irrigation, fertilization, and related programs embodied in India's five-year plans have made a tremendous impact. By 1972 the state had already exceeded its Fourth Five-Year Plan (1969–74) target of 4,400,000 tons of additional food grains. The spurt in food-grain production (mainly wheat) led to a significant growth in per capita income. Except in the northeast, most of the land is used for two crops annually. In addition to wheat, the principal winter crops are gram, barley, and mustard. The principal summer crops are millet, rice, maize (corn), sugarcane, and cotton. The state is also known for the quality of its bullocks and dairy cattle.

Agricultural programs

To increase agricultural production, the state has initiated a number of irrigation schemes to tap subsoil water and has undertaken several small and medium-size irrigation projects. The state has been energizing more than 20,000 tube wells (driven wells) a year to augment irrigation facilities. There are more than 77,000 such tube wells.

The state has no heavy industry; nor has it any sizable mineral resources for exploitation. Light industries, however, have been developing, especially since the late 1960s. The increase in agricultural production, especially of sugarcane and cotton, has stimulated related industries, and there were several cotton textile and sugar mills in the state in the early 1970s. Other light industries include the manufacture of farm tools, machine tools, electrical and glass goods, cement, paper, and bicycles. Industrial areas with medium- and small-scale factories and other enterprises have developed at Farīdābād and Ballabhgarh, south of Delhi (where Haryana borders Uttar Pradesh), and at Sonepat, Jagādhri, and Yamunānagar. Industrial areas are also rising up along the Grand Trunk Road from Delhi and along railways to take advantage of transport facilities.

Transport and communication. Road development has been accorded high priority in the state's development plans. These included the construction of 2,500 miles of new roads to bring the state total to 6,200 miles. Haryana Roadways, which is a nationalized transport undertaking, operates hundreds of vehicles on local and interstate routes.

The chief lines of communication are those leading to and from Delhi, particularly the historic Grand Trunk Road and the mainline of the Northern Railway between Ambāla and Delhi, which is the centre of the nation's rail, road, and airway networks.

Culture. Haryana is a land of legends, temples, historical places, and archaeological sites. Besides the Hindu pilgrim centre of Kurukṣetra, site of a bathing fair at each solar eclipse, the state has several sacred and historic sites including Karnāl, site of many events in the epic of the *Mahābhārata;* Thānesar, an early (9th–12th century) Hindu capital; and Gurgaon, capital of the Gurjara dynasty before the 10th century. Rohtak was the centre of the Yaudheya warriors in the 5th and 6th centuries. Religious centres with sacred tanks for bathing and other purposes include Agroha, Pandara, Pehowa, and Ramrah. Surajkund, an ancient centre of sun worship, has a sacred tank built in the shape of the rising sun. Pānīpat, mentioned in the *Mahābhārata,* was the scene of important dynastic battles in the 16th and 18th centuries.

The Indus Valley civilization was crystallized at Khokhra Kot (Rohtak), Naurangabad, Meham, and Agroha in Haryana. The Gurukul (University) at Jhajjar has a rich collection of ancient coins, molds, seals, inscriptions, utensils, and statues, which throw light on the life of people inhabiting the region at different periods.

The people of Haryana are simple and hardworking and have preserved their old religious and cultural traditions and customs. They hold fairs and festivals with great enthusiasm, among the most notable being the Janam Ashtami (Lord Kṛṣṇa's birthday) Fair at Bhiwāni and the Masani Fair at Gurgaon. Folk theatre and folk dancing have also been very popular. *Saang* is the most popular variety of performances with the participants—the female roles played by men—singing and dancing continuously for five to six hours. The themes are drawn from mythology or Sanskrit literature and occasionally include love lores of Punjab.

Essentially a rural state, more than 80 percent of the people live in the villages, where their principal occupation is agriculture. At the same time, the Indian army, traditionally a favourite career for young men, finds Haryana a major recruiting ground.

BIBLIOGRAPHY. R.S. SHARMA (ed.), *Haryana Directory & Who's Who: 1967–68* (1968), gives the background of the formation of Haryana state and various details of the state's economy. The *Haryana Review* (quarterly), gives useful information about the state; *Facts about Haryana* (1970), is a small statistical handbook. Both are published by the Public Relations Department, Haryana.

(C.Ra.)

Ḥasan al-Baṣrī, al-

Al-Ḥasan al-Baṣrī (Abū Sa'īd ibn Abī al-Ḥasan Yasār al-Baṣrī), one of the most important religious figures in early Islām, was known to his own generation as an eloquent preacher, a paragon of the truly pious Muslim, and an outspoken critic of the political rulers of the Umayyad dynasty (AD 661–750). Among later generations of Muslims, he has been remembered for his piety and religious asceticism. Muslim mystics have counted him as one of their first and most notable spiritual masters. Both the Mu'tazilah (philosophical theologians) and the Ash'arīyah (followers of the theologian al-Ash'arī), the two most important theological schools in early Sunnī (traditionalist) Islām, considered Ḥasan one of their founders.

Ḥasan was born in Medina in AD 642, nine years after the death of the prophet Muḥammad. One year after the battle of Ṣiffīn (657), which proved to be one of the most important events in the history of Islām, Ḥasan moved to Basra, a military camp town situated 50 miles (80 kilometres) northwest of the Persian Gulf. From this base, military expeditions to the east disembarked, and, as a young man (670–673), Ḥasan participated in some of the expeditions that led to the conquest of eastern Iran. In a short time, Basra came to be more than a military post; its cultural importance in the history of Islām is based on the fact that it became the home of intense religious and intellectual activity that had profound effects on the development of later Islāmic thought.

Early career; leader in theological debates

After his return to Basra, Ḥasan became a central figure in the religious, social, and political upheavals brought about by internal conflicts with the Muslim community. The years 684–704 marked the period of his great preaching activity. From the few remaining fragments of his sermons, which are among the best examples of early Arabic prose, there emerges the portrait of a deeply sensi-

tive, religious Muslim. For Ḥasan, the true Muslim must not only refrain from committing sin, but he must live in a state of lasting anxiety, brought about by the certainty of death and the uncertainty of his destiny in the hereafter. He said: "I never saw a certainty of which there is no doubt bear a greater resemblance to a doubtful thing of which there is no certainty, than death does." The world is treacherous, "for it is like to a snake, smooth to the touch, but its venom is deadly." The practice of religious self-examination (*muḥāsabah*), which led to the activity of avoiding evil and doing good, coupled with a wariness of the world, marked Ḥasan's piety and influenced later ascetic and mystical attitudes in Islām.

Religious and political views

The enemy of Islām, for Ḥasan, was not the infidel but the hypocrite (*munāfiq*), who took his religion lightly and "is here with us in the rooms and streets and markets." In the important freedom–determinism debate, he took the position that man is totally responsible for his actions, and he systematically argued this position in an important letter written to the caliph ʿAbd al-Malik. His letter, which is the earliest extant theological treatise in Islām, attacks the widely held view that God is the sole creator of man's actions. The document bears political overtones and shows that in early Islām theological disputes emerged from the politico-religious controversies of the day. He wrote: "Oh Commander of the Faithful . . . tyranny and injustice are not from God's decree, but His decree is His command concerning goodness, justice, kindness, and generosity regarding one's relatives." His political opinions, which were extensions of his religious views, often placed him in precarious situations. During the years 705–714, Ḥasan was forced into hiding because of the stance he took regarding the policies of the powerful governor of Iraq, al-Ḥajjāj. After the Governor's death, Ḥasan came out of hiding and continued to live in Basra until he died in 728. It is said that the people of Basra were so involved with the observance of his funeral that no afternoon prayer was said in the mosque because no one was there to pray.

BIBLIOGRAPHY. With the exception of a few of his sayings, nothing from al-Ḥasan's works has been published in English translation. The most important Western-language studies are: LOUIS MASSIGNON, *Essai sur les origines du lexique technique de la mystique musulmane*, rev. ed. (1954), an examination of al-Ḥasan's place in Islāmic mysticism; HELMUT RITTER, "Studien zur Geschichte der islamischen Frömmgkeit," *Der Islam*, 21:1–83 (1933), the most detailed study of al-Ḥasan's thought—includes the Arabic text of the letter written to ʿAbd al-Malik; JULIAN OBERMANN, "Political Theology in Early Islam: Ḥasan al-Baṣrī's Treatise on Qadar," *Journal of the American Oriental Society*, 55:138–162 (1935), which also discusses the letter written to ʿAbd al-Malik; and "Hasan al-Basri," *Encyclopaedia of Islam*, new ed., vol. 3, pp. 247–248 (1966).

(D.E.)

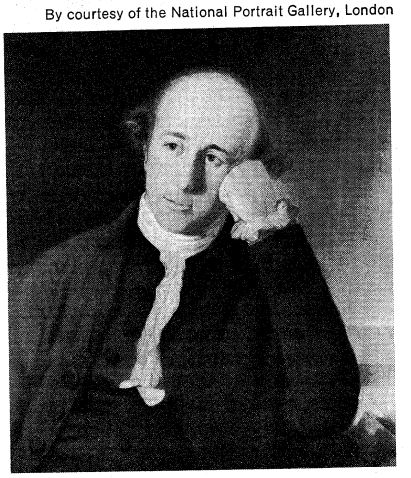
Hastings, Warren

As the first British governor general of Bengal, Warren Hastings was responsible for consolidating British control over the first major Indian province to be conquered. In the 13 years he held office (1772–85), solutions began to be devised to such problems as how vast Indian populations were to be administered by a handful of foreigners and how the British, now themselves a major Indian power, were to fit into the state system of 18th-century India. These solutions were to have a profound influence on Britain's future role in India. Hastings' career is also of importance in raising for the British public at home other problems created by their new Indian empire—problems of the degree of control to be exercised over Englishmen in India and of the standards of integrity and fair dealing to be expected from them—and the solutions to these problems were also important for the future.

Early life. The son of a clergyman of the Church of England, Hastings was born on December 6, 1732, at a village in Oxfordshire. Abandoned by his father at an early age, he was brought up by an uncle, who gave him what was probably the best education then available for a boy of his inclinations, at Westminster School in London. Hastings showed great promise as a schoolboy and seems at Westminster to have acquired the literary and scholarly tastes that were later to give him a serious interest in Indian culture and civilization. His school days were, however, cut short by his uncle's death in 1749. He was then taken away from school and granted a writership (as the junior appointments in the East India Company were called), and in 1750, at the age of 17, he sailed for Bengal.

In 1750 British contact with India was still the monopoly of the East India Company, which was engaged in

By courtesy of the National Portrait Gallery, London



Hastings, oil painting by Tilly Kettle (1735–86). In the National Portrait Gallery, London.

buying and selling goods at small settlements in Indian ports. As one of the company's servants, for the early part of his career Hastings was employed in the company's commercial business. But after 1756 the outlook for both the company and its servants was radically altered. The company became involved in hostilities in India both with the French and with Indian rulers, and under Robert Clive its army was able to depose the nawab, or Indian governor, of Bengal at the Battle of Plassey in 1757. Although the company did not at this stage intend to set itself up as the actual ruler of the province, it was now so powerful that the new nawabs became its satellites. Thus, the servants, including Hastings, began to be drawn more and more into Indian politics. Hastings served as the company's representative at the court of the nawabs of Bengal from 1758 to 1761 and then on the company's Council, the controlling body for its affairs in Bengal, from 1761 to 1764. His career was cut short, however, by bitter disputes within the Council. Finding himself in a minority, Hastings resigned from the company's service and returned to England in 1765.

Servant of the East India Company

Governorship of Bengal. Short of money, Hastings sought service in India again. In 1769 he was appointed second in Council in Madras. Two years later he received his great opportunity when he was sent back to Bengal as governor in charge of the company's affairs there. Since he had last been in Bengal, the disintegration and demoralization of the normal Indian government of the province, begun after Plassey, had gathered speed; yet the company had been reluctant to create a new system in its place. In practical terms Bengal was in the power of the British, who were also virtually its legal rulers after being granted in 1765 the powers called the dewanee by the Mughal emperor. But the business of government was still conducted by Indian officials, with very limited European participation. Hastings recognized that this situation could not go on and that the British must accept full responsibility, make their power effective, and involve themselves more closely in the work of government, even if he shared his contemporaries' objections to excessive involvement. His view of the British role in India was later to be regarded as a very conservative one. He saw no "civilizing" or modernizing mission for them. Bengal was to be governed in strictly traditional ways, and the life of its people was not to be disturbed by in-

novation. To ensure good government, however, he felt that the British must actively intervene. In what was to be the most constructive period of his administration, from 1772 to 1774, Hastings detached the machinery of the central government from the nawab's court and brought it to the British settlement in Calcutta under direct British control, remodelled the administration of justice throughout Bengal, and began a series of experiments aimed at bringing the collection of taxation under effective supervision.

End of undisputed power in Bengal

Political rivalries. Hastings' period of undisputed power in Bengal came to an end in 1774 with changes in the company's government. He acquired the new title of governor general and new responsibilities for supervising other British settlements in India, but these powers had now to be shared with a Supreme Council of four others, three of whom were new to India. The new councillors, who were led by an army officer, Sir John Clavering, and included the immensely able and ambitious Philip Francis, immediately quarrelled with Hastings. Hastings' admirers have had little patience with Clavering and Francis; but it is possible to see that Francis had a genuine point of view in his opposition to Hastings and that there was still much in Bengal, even after Hastings' reforms, to shock men fresh from Britain (bribery, extortion, and other abuses of power by Englishmen, which had been so common since Plassey, undoubtedly continued). The quarrel between the new councillors and Hastings paralyzed the government of Bengal and produced a number of squalid episodes in which the newcomers, to discredit Hastings at home, encouraged Indians to bring accusations of malpractices against him, while his friends used various methods to deter such accusations. The most notorious of these episodes concerned one Maharaja Nandakumar, who made accusations against the governor general but was in his turn accused of forgery and hanged for it. Hastings was certainly not guilty of procuring a judicial murder, but recent research does suggest that he knew in advance of the counterplot against Nandakumar.

War in India. The death of Clavering in 1777 put Hastings once again in possession of full power, although Francis' opposition dragged on for another three years. It ended in a pistol duel between Hastings and Francis; the latter was wounded, and he returned to Europe. But by 1777 the energies of the Bengal government were becoming more and more absorbed in war. War against Indian states was always a likely consequence of the company's conquest of Bengal. As full participants in the unstable world created in India by the fall of the Mughal Empire, the company now found it difficult not to be drawn into the rivalries of the powers that had set themselves up in the ruins of the empire. Hastings' policy was to avoid further conquest and war but to maintain peaceful relations with neighbouring states by a series of alliances. He had, however, already taken part in one war in 1774, when he helped the company's ally on the northwestern boundary of Bengal, the Vizier of Oudh, to take over territory occupied by a people called the Rohillas; and in 1778 he became involved in war with the Marāthās, a loose federation of Hindu peoples in western and central India. Rightly or wrongly, Hastings came to believe that it was necessary for the safety of the British in India to ensure that the Marāthā leaders were friendly to the company and that he would be justified in applying military pressure to achieve this end. After the entry of France into the U.S. War of Independence in 1778, he was also confronted with French expeditionary forces in the Indian Ocean. Finally, in 1780, Hyder (Haidar) Ali, the ruler of the south Indian state of Mysore, attacked the British at Madras. War on several fronts brought out the best in Hastings, and his achievement in organizing the company's military and financial resources to counter every threat was a remarkable one. The Marāthās were brought to peace in 1782, as was Mysore in 1784, and the French were held in check until peace was made in Europe in 1783. But war stretched the company to the limit, disrupting its trade and thus antagonizing opinion at home. War also forced Hastings (or so he believed) into dubious acts to raise extra funds, two of which—the demand for a subsidy to the company from Chait Singh, the raja of Banaras, and the requisitioning of the treasures of the begums of Oudh (the mother and grandmother of the Vizier)—were to count heavily against him later.

Wartime achievements

Retirement and impeachment. It was, however, an India at peace and with British dominions fully intact that Hastings finally left in 1785. But even before his retirement the allegations of Francis and the reports of wars, whether justly or unjustly undertaken, had damaged his reputation; and the passionate moral concern about the standards of the British in India felt by Edmund Burke, the great Whig parliamentarian, had come to be focussed on Hastings. Most historians, while recognizing Burke's absolute sincerity, now feel that Burke was attempting to pin the evils of a situation on one individual and that he had chosen the wrong one. But Hastings was vulnerable on episodes such as the execution of Nandakumar and his treatment of the begums of Oudh and Chait Singh and even on some aspects of his personal finances, where he had acquired money in excess of his official allowances. In 1786, when Burke introduced an impeachment process against him (a prosecution by the House of Commons before the House of Lords), these blemishes were enough to persuade the House of Commons and in particular William Pitt the Younger, the prime minister, that Hastings ought to be sent to trial. The trial before the House of Lords lasted from 1788 to 1795, when he was acquitted. It is difficult not to regard this long-drawn-out ordeal as a serious injustice. At the most it made some contribution to the process by which standards were being laid down for the future conduct of British rule in India.

After his acquittal, Hastings lived the life of a retired country gentleman—unassuming, mild-mannered, and of scholarly tastes, as he had been during his active career. He died on August 22, 1818, at age 85.

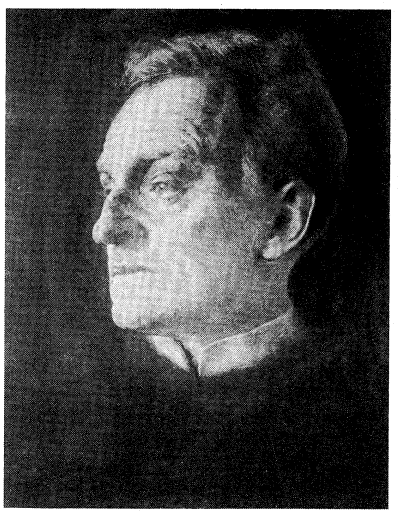
BIBLIOGRAPHY. PENDEREL MOON, *Warren Hastings and British India* (1947), is the best short account. KEITH FEILING, *Warren Hastings* (1954), is a full biography. P.J. MARSHALL, *The Impeachment of Warren Hastings* (1965), deals with the trial and the controversial episodes featured in it. For the English and Indian backgrounds respectively, see LUCY S. SUTHERLAND, *The East India Company in Eighteenth-Century Politics* (1952); and ABDUL MAJED KHAN, *The Transition in Bengal, 1756–1775* (1969). Hastings' letters may be consulted in G.R. GLEIG, *Memoirs of the Life of the Right Hon. Warren Hastings, First Governor-General of Bengal*, 3 vol. (1841). Macaulay's Essay on Warren Hastings, available in most editions of his collected essays, should still be read.

(P.J.M.)

Hauptmann, Gerhart

The most prominent German dramatist of his time, Gerhart Hauptmann established his reputation in 1889 as an exponent of Naturalism with the performance of *Before Dawn* in Berlin. This starkly Realistic tragedy, dealing with contemporary social problems, signalled the end of the rhetorical and highly stylized drama of the 19th century. Hauptmann's most celebrated works are in the Realistic style, notably, such proletarian tragedies as *The Weavers* and *Drayman Henschel*, as well as the tragicomedy *The Rats* and the satirical comedy *The Beaver Coat;* but he also produced, beginning in the 1890s, dramas in a variety of other styles: lyrical, Symbolist, historical, neo-Romantic, Neoclassical, and others. Many of these are tinged with the mystical religiosity to which he turned with advancing years. Starting in the 1880s, he also published stories, novels, verse epics, poems, essays, and autobiographical accounts, which demonstrated his extraordinary range. The unifying element of Hauptmann's vast and protean literary output is the sympathetic but by no means activist concern for human suffering: his heroes are generally passive, victims of elementary "powers." He extended this view to the concept of history in his highly controversial *Festspiel*, which commemorates the centenary of the anti-Napoleonic Wars.

Hauptmann was born on November 15, 1862, in the then fashionable Silesian resort town of Obersalzbrunn (now Szczawno Zdrój, Poland), where his father owned



Hauptmann, etching by Hermann Struck, 1904. In the Schiller-Nationalmuseum, Marbach, West Germany.
By courtesy of the Schiller-Nationalmuseum, Marbach

the main hotel. He enjoyed little formal education. Leaving school at 15, he tried to find himself in a variety of pursuits during the next ten years. After trying his hand at farming, he studied sculpture for nearly two years at the Breslau (now Wrocław, Poland) Art Institute and then joined his older brother Carl in Jena for a semester of studies in history, philosophy, and science. After spending the winter of 1883–84 in Rome as a sculptor, he continued his academic studies and took acting lessons in Berlin. It was at this time that he decided to make his career as a poet and dramatist. Having married the well-to-do Marie Thienemann in 1885, Hauptmann settled down in Erkner, a rural suburb of Berlin, dabbling in political, sociological, theological, and literary studies and associating with a group of scientists, philosophers, and avant-garde writers. The circle was influenced by the literary works of Leo Tolstoy, Émile Zola, and Henrik Ibsen; the philosophical writings of Karl Marx and Friedrich Nietzsche; the theological works of D.F. Strauss; and the writings of the German Darwinian zoologist Ernst Haeckel. Although he was never a doctrinaire Naturalist, Socialist, or Materialist, Hauptmann's first dramas grew out of the ferment of such ideas.

First success

In October 1889, *Before Dawn* made him famous overnight, though it had to be performed in the Freie Bühne, a private theatre club dedicated to the modern drama. Encouraged by the ensuing uproar, he wrote in rapid succession a number of successful plays on Naturalistic themes (heredity, the plight of the poor, the clash of natural science and religion), again artistically reproducing social reality and common speech, as well as nonverbal elements of communication. Most gripping and humane, as well as most objectionable to the political authorities, was *The Weavers*, a compassionate dramatization of the Silesian weavers' revolt of 1844. The following decade, the writer's most creative, was marked by marital difficulties, which ended in 1904 with a divorce; in the same year he married an actress and violinist, Margarete Marschalk (12 years his junior), four years after the birth of their son, Hauptmann's fourth. In 1901 they had moved to their newly built mansion, Haus Wiesenstein, at Agnetendorf in Silesia (Jagniąków, Poland). No longer suffering from his usual sickliness, Hauptmann spent the rest of his life here in his characteristically convivial style, summering on the Baltic isle of Hiddensee, wintering on the Italian Riviera or in the Ticino district of Switzerland, and travelling frequently to Berlin and Vienna and occasionally elsewhere (to England in 1905, to Greece in 1907, to the U.S. in 1932). The tranquillity of his life was interrupted only temporarily, in 1905–06, by his romance with Ida Orloff, a 17-year-old actress, who was to become the model of numerous elfish figures in plays and novels.

Hauptmann never became a partisan of any literary trend, receptive though he was to many of them. Although he was honoured like a national figurehead at his frequent public appearances, he tended to isolate himself somewhat from the literary and intellectual life of the time. His literary production became prolific—he dictated virtually every afternoon—but also more uneven over the years; yet it culminated in one of the most powerful dramatic works of the century, the *Atridentetralogie*, a tetralogy on the myth of the Atrides that he wrote during World War II. In addition to the literature and mythology of Greek and Germanic antiquity, in his later years he derived much of his inspiration from Shakespeare (whose *Tempest* he adapted in *Indipohdi*) and Goethe (to whom he felt a particularly close personal affinity), as well as from his own early life. He also delved into the theology and mythology of early Christianity, especially into its heretical movements. These studies have left their imprint most clearly on his philosophically ambitious visionary epic in terza rima, *Der grosse Traum* ("The Great Dream"). The autodidactic philosophical and scholarly pursuits and the syncretistic cosmological speculations of Hauptmann's later decades did, however, distract him from his spontaneous talent for creating characters that come alive on the stage and in the imagination of the reader. Not surprisingly, he was also an ingenious director of his own plays, as well as those of others.

Period of major literary productivity

Hauptmann's fame—the Nobel Prize (1912) being only one of many international honours—was unequalled, though not uncontested, until the ascendancy of Nazism, when he was barely tolerated by the regime and was denounced by émigrés for staying in Germany. Though privately out of tune with the Nazi ideology, he was politically naïve and tended to be indecisive. He died on June 6, 1946, in Agnetendorf and was buried on the isle of Hiddensee. His plays, the early Realistic ones especially, are still frequently performed, though not so often as during his lifetime.

Fame and honours

MAJOR WORKS

PLAYS: *Vor Sonnenaufgang* (1889; *Before Dawn*); *Das Friedensfest. Eine Familienkatastrophe* (1890); *Einsame Menschen* (1891; *Lonely Lives*); *Die Weber* (1892; *The Weavers*); *College Crampton* (1892); *Der Biberpelz* (1893; *The Beaver Coat*); *Hannele* (1894; later pub. as *Hanneles Himmelfahrt*; Eng. trans., *The Assumption of Hannele*); *Florian Geyer* (1896); *Die versunkene Glocke* (1897; *The Sunken Bell*); *Fuhrmann Henschel* (1898; *Drayman Henschel*); *Schluck und Jau* (1900); *Michael Kramer* (1900); *Der rote Hahn* (1901; *The Conflagration*); *Der arme Heinrich* (1902; *Henry of Auë*); *Rose Bernd* (1903); *Elga* (1905); *Und Pippa tanzt!* (1906; *And Pippa Dances*); *Kaiser Karls Geisel* (1908; *Charlemagne's Hostage*); *Griselda* (1909); *Die Ratten* (1911; *The Rats*); *Gabriel Schillings Flucht* (1912); *Festspiel in deutschen Reimen* (1913; *Commemoration Masque*); *Der Bogen des Odysseus* (1914; *The Bow of Odysseus*); *Winterballade* (1917); *Der weisse Heiland* (1920; *The White Saviour*); *Indipohdi* (1920); *Veland* (1925); *Dorothea Angermann* (1926); *Vor Sonnenuntergang* (1932); *Hamlet in Wittenberg* (1935); *Die Tochter der Kathedrale* (1939); *Die Atridentetralogie—Iphigenie in Delphi* (1941), *Iphigenie in Aulis* (1944), *Agamemnons Tod* (1948), *Elektra* (1948). The above translations can be found in *The Dramatic Works of Gerhart Hauptmann*, trans. by Ludwig Lewisohn *et al.*, 9 vol. (1912–29).

NOVELS: *Der Narr in Christo, Emanuel Quint* (1910; *The Fool in Christ, Emanuel Quint*, 1911); *Atlantis* (1912; Eng. trans., 1923); *Der Ketzer von Soana* (1918; *The Heretic of Soana*, 1923); *Phantom* (1922; Eng. trans., 1923); *Die Insel der grossen Mutter* (1924; *The Island of the Great Mother*, trans. by Willa and Edwin Muir, 1925); *Wanda* (1928); *Der neue Christophorus* (1943), fragment.

STORIES: *Bahnwärter Thiel* (1888); *Die Hochzeit auf Buchenhorst* (1932); *Das Meerwunder* (1934); *Der Schuss im Park* (1942).

VERSE: *Anna* (1921); *Till Eulenspiegel* (1928); *Der grosse Traum* (Part 1, 1942); *Neue Gedichte* (1946).

OTHER WORKS: *Griechischer Frühling* (1908), travel diary; *Buch der Leidenschaft* (1930); *Das Abenteuer meiner Jugend* (1937), autobiographical.

BIBLIOGRAPHY. WALTER A. REICHART, *Gerhart-Hauptmann-Bibliographie* (1969), is a selective listing of scholarly books and articles. Hauptmann's own archives (thousands of manu-

scripts, typescripts and letters, notebooks and diaries, many of them unpublished) are preserved in the Staatsbibliothek der Stiftung Preussischer Kulturbesitz, West Berlin. The Schiller-Nationalmuseum, Marbach-Neckar, West Germany, holds a sizable collection of primary and secondary materials (exhibition catalog: *Gerhart Hauptmann: Leben und Werk*, 1962) as does the Deutsche Akademie der Künste, East Berlin. 5,000 volumes of Hauptmann's library are preserved in the Märkisches Museum in East Berlin; 4,000 volumes in the Staatsbibliothek der Stiftung Preussischer Kulturbesitz; and 1,000 volumes in Hauptmann's house in Kloster on the isle of Hiddensee, East Germany.

Editions and Correspondence: Das Gesammelte Werk, series 1, 17 vol. (1942–43), series 2, drafts and unfinished works, not published. The more comprehensive 11-volume centenary edition, *Sämtliche Werke*, ed. principally by HANS EGON HASS (1962–73), contains previously unpublished completed works and substantial selections from drafts and unfinished writings. There is no edition of the correspondence; the letters to and from Ida Orloff are included in *Gerhart Hauptmann und Ida Orloff* (1969).

Biography and comprehensive appreciation: C.F.W. BEHL and FELIX A. VOIGT, *Chronik von Gerhart Hauptmanns Leben und Schaffen* (1957); HUGH F. GARTEN, *Gerhart Hauptmann* (1954), in English; KARL S. GUTHKE, *Gerhart Hauptmann: Weltbild im Werk* (1961); EBERHARD HILSCHER, *Gerhart Hauptmann* (1969); HANS DAIBER, *Gerhart Hauptmann, oder der letzte Klassiker* (1971).

Studies of phases, themes, groups of works, and problems: FELIX A. VOIGT and WALTER A. REICHART, *Hauptmann und Shakespeare*, 2nd ed. (1947); SIEGFRIED H. MULLER, *Gerhart Hauptmann and Goethe* (1949); RALPH FIEDLER, *Die späten Dramen Gerhart Hauptmanns* (1954); MARGARET SINDEN, *Gerhart Hauptmann: The Prose Plays* (1957); LEROY R. SHAW, *Witness of Deceit: Gerhart Hauptmann As Critic of Society* (1958); FREDERICK W.J. HEUSER, *Gerhart Hauptmann: Zu seinem Leben und Schaffen* (1961); *Hauptmann Centenary Lectures*, ed. by K.G. KNIGHT and F. NORMAN (1964); FELIX A. VOIGT, *Gerhart Hauptmann und die Antike* (1965).

(K.S.G.)

Havana

One of the great historic cities of the New World, Havana (Spanish La Habana) is the capital of the Republic of Cuba and its major economic, political, and cultural centre. Its core, Old Havana, still contains carefully preserved remnants of the colonial period, when Havana was the leading Spanish stronghold in the Americas and, for a time, the greatest port of the Western Hemisphere. Contemporary Havana, which includes also the modern city and its suburbs, has undergone a radical transformation since the Cuban Revolution of 1959, when it was essentially a developed, even overgrown city in an underdeveloped country. A major current goal of the government is to redress this imbalance.

The capital is located toward the western end of the northern coast of the island of Cuba. In an administrative reorganization of the country in October 1976, the boundaries of the city were extended and 15 municipalities within the new city limits were created. Thus, in the later 1970s Havana was the home of more than 1,900,000 persons, about 20 percent of the national total, concentrated in 286 square miles (740 square kilometres), less than 0.5 percent of the national territory.

The harbour

The city is located on the west side of a classic example of a "bottleneck" harbour, the narrow entrance of which, 270 yards wide and 1,500 yards long (250 by 1,400 metres), leads to a broad interior bay divided into the smaller bays of Marimelena, Guasabacoa, and Atarés. The city is crossed by the slow-moving Río Almendares, which enters the sea between the old cities of Marianao and Havana. Havana's most notable topographic feature is a limestone ridge some 200 feet (60 metres) high, which penetrates the city from the east, parallel to the coast, and terminates at the bay entrance in the heights of La Cabaña and El Morro. This same relief feature reappears in the hill occupied by the Universidad de La Habana (University of Havana) and the Castillo del Príncipe (Prince's Castle) and extends southward across Marianao. Other low hills include the Castillo de Atarés and those dotting the historic neighbourhoods of El Cerro, Jesús del Monte, and La Víbora.

Climate

The city's climate, tropical yet mild, is tempered by the location of the island in the belt of the trade winds and by the warm Gulf Stream offshore; occasionally in winter a cold front from the North American mainland brings cold northwest winds (the Norte) that temporarily displace the usual mild onshore breezes. Relative humidity averages 78 percent, and rainfall averages about 73 inches (1,850 millimetres) annually, falling largely in the May to October rainy season. Average temperature is 79° F (26° C), varying by only about 9° F (5° C) throughout the year. (For information on related topics, see LATIN AMERICA AND THE CARIBBEAN, COLONIAL; CUBA; and CUBA, HISTORY OF.)

History. *Foundation and early growth.* The city of Havana was founded in the southern part of the province of the same name by the Spanish colonizer Diego de Velázquez de Cuéllar in 1515. The first location may have been either on the site now occupied by the little town of Surgidero de Batabanó, or nearby, where the Río Mayabeque flows into the sea. Its inhabitants later moved to the northern coast in search of better climatic conditions, then began to shift toward the city's present position, finally establishing themselves at the port of Carenas, today the port of Havana, around 1519.

Because of its geographic position, Havana was a port of call for Spanish ships travelling between the colonies in America and the Iberian Peninsula. As a result, the primitive village slowly evolved from a group of rustic houses into a city of stone structures, including churches and public buildings, with carefully planned paved streets and public squares.

In 1538 the city was attacked, and in 1555 it was plundered, by French buccaneers. Philip II of Spain in 1589 ordered the erection of the two forts, La Punta and El Morro, at the entrance to the passage leading to the harbour, to reduce the danger from pirates.

In 1656 the construction of a defensive city wall began, with important consequences, both military and structural. Within the walls the area now known as Old Havana (La Habana vieja) developed, and, though it has been constantly modified, it retains many characteristically colonial buildings as well as the original crooked streets. The walls made Havana the leading Spanish stronghold in America, a status soon reflected in the designations "Key to the New World" and "Rampart of the Indies," titles assigned the city by the Spanish crown. In 1607 Havana had officially been named the island's capital by royal decree, though it had occupied this position in practice since 1553.

English occupation

The 18th and 19th centuries. In 1762, during the Seven Years' War, after bloody battles around the capital the English captured it. Their occupation lasted only 11 months, after which the city was returned to Spain in exchange for the territory of Florida. The brief occupation bestowed important benefits on the island's landholders: the English reduced some of the limitations on foreign commerce and eliminated the monopoly held by the Real Compañía de Comercio (Royal Commerce Company). On their return, the Spanish had no alternative but to continue the policies established by the English. In 1765 Havana became a free port for commerce with Spain, leading to a period of economic prosperity, which in turn consolidated the power of the nascent Cuban aristocracy.

The city's original population was remarkably heterogeneous; in addition to its customary inhabitants, it accommodated transient soldiers, sailors, clergy, merchants, and ordinary travellers who often had to spend many months in the city waiting for the last ship of the fleet to arrive so that in convoy—the so-called Fleet System, for protection against pirates—they could safely make the return voyage to Spain. There were some 50,000 persons in Havana toward the middle of the 18th century, including the population of the suburbs that were already appearing outside the wall. The third most heavily populated city in the Americas at that time, it was also the New World's most important port in terms of volume of commercial transactions.

The city continued to grow during the 19th century but



El Vedado section of Havana.
Sovfoto

was no longer merely a place of transit. It had begun to export sugar, coffee, tobacco, and rum, and its economic health was reflected in further growth of the suburbs. In 1817 the tobacco monopoly was suppressed, and a year later free commerce with other countries was authorized.

The 20th century. The city's growth continued in the 20th century, especially to the west and south, and Havana became a striking, modern city, with wide avenues and large buildings. This urban growth and concentration, disproportionate to the rest of the country, was accompanied by the creation of an incipient industrial base that has been insufficient, however, to absorb the surplus work force emigrating to the capital from rural areas. Since the revolution in 1959, attempts have been made to reverse this flow of population.

Layout and architecture. As noted above, the capital first developed within its walls but soon began to expand outward into suburban neighbourhoods. The city had already greatly exceeded its original walled boundaries by the 19th century, and its subsequent growth has been primarily toward the west and south, with increasingly more distant areas being integrated into the urban nucleus. The city's expansion to the east was facilitated by the construction of an 875-yard (800-metre) tunnel under the entrance to the bay, permitting the creation of such new suburban neighbourhoods as Habana del Este in a previously inaccessible area. Contemporary Havana is thus an urban, political, and administrative unit comprising the old cities of Havana, Marianao, Regla, Guanabacoa, Santiago de las Vegas, and Santa María del Rosario.

Composition of metropolitan Havana

Some of the city's most imposing architecture is located around the Plaza de la Revolución (Revolutionary Square, formerly the Plaza Cívica); the monument to José Martí, the "apostle of independence," in the centre, is surrounded by such modern public buildings as the Centro del Gobierno Nacional and buildings housing the Central Committee of the Communist Party of Cuba, as well as the Armed Forces, Interior, Communications, and Construction ministries, the Biblioteca Nacional (National Library), the Junta Central de Planificación (or Juceplan; Central Planning Board), and others. In the centre of the city, imposing in style although less modern in architecture, are the buildings housing the Academia de Ciencias de Cuba (Academy of Sciences of Cuba, the former National Capitol), the Museo de la Revolución (Museum of the Revolution, in the old Presidential Palace), the Centro Asturiano (Asturian Centre) and the Centro Gallego (Galician Centre), founded as clubs by Spanish immigrants, and the Palacio Aldama.

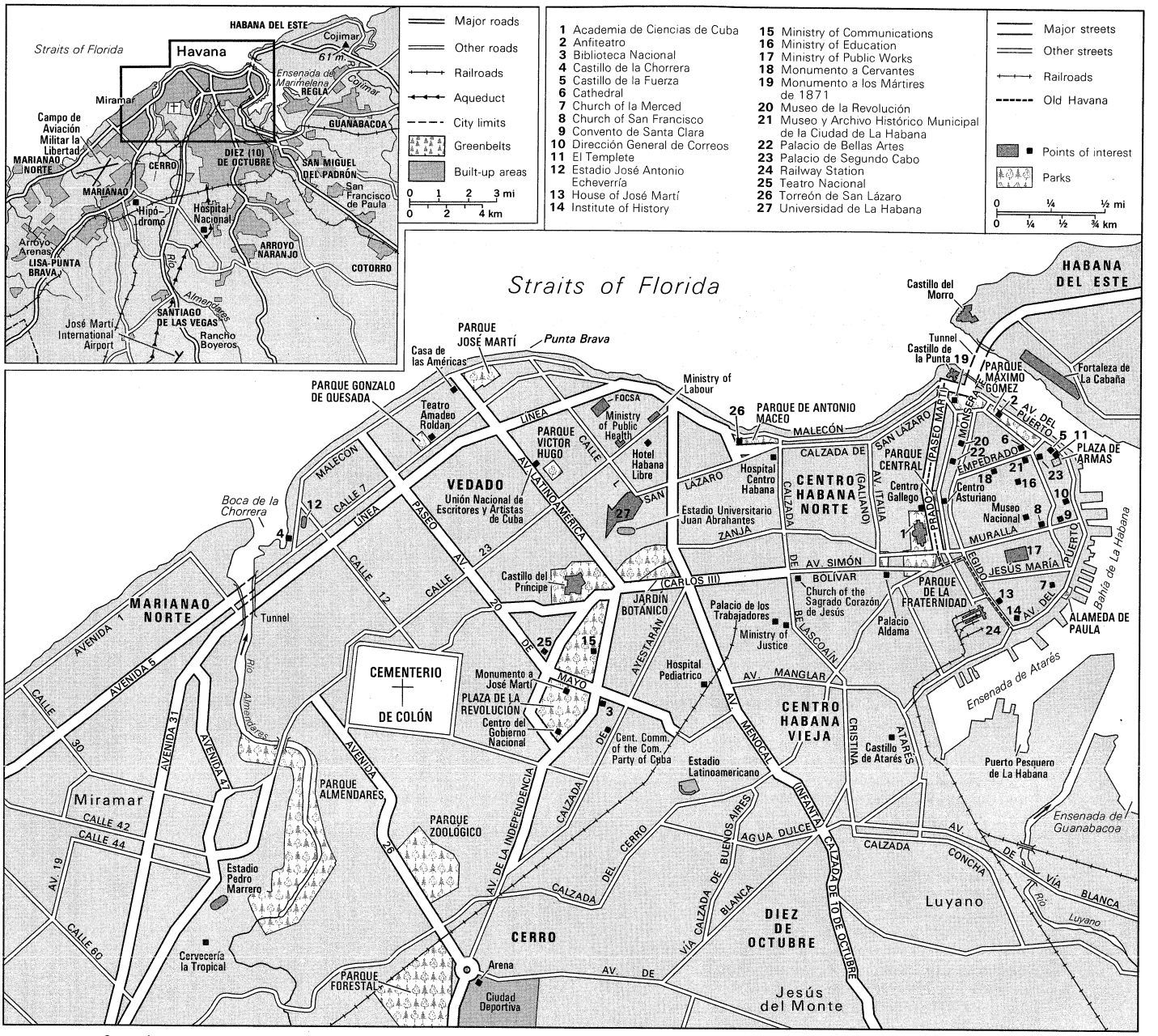
Old Havana, the original urban nucleus next to the port, is characterized by history-laden buildings in pure colonial style. Among the most outstanding of these is the Palacio de los Capitanes Generales (Palace of the Captains General), completed in 1793; from 1902 this accommodated the first presidents of the republic and later was the home of the Ayuntamiento (City Council). After the revolution it was restored to its former architectural magnificence and now houses the Museo y Archivo Histórico Municipal de la Ciudad de La Habana, the city's historical museum and archives. Old Havana also contains the Palacio del Segundo Cabo (*i.e.*, of the former second-ranking military chief of the province), also restored and now the seat of the Consejo Nacional de Cultura (National Council of Culture), and El Templete.

The architectural heritage of Old Havana

Other outstanding architectural features are those fortresses built for the city's defense, such as the Castillo del Morro, or Morro Castle (to which was subsequently added the lighthouse that guides ships to the port of Havana, a famous symbol of the city), the Fortaleza de La Cabaña, the Castillo de la Punta, the Castillo de Atarés, the Castillo del Príncipe, the Torreón de San Lázaro, the Castillo de la Chorrera, and the Castillo de la Fuerza, a monumental fortress perhaps built on the city's original site. The last named is the oldest colonial military building in the Americas, having been rebuilt between 1565 and 1583 on the site of a fortress built in 1538 on the orders of the colonizer and explorer Hernando de Soto. Havana's ancient ruined walls are still visible, as are many other buildings bearing witness to the colonial city's military power.

Old Havana also contains such fine colonial churches as the late 18th-century cathedral; the church of San Francisco, finished in 1575 and reconstructed 1731–37 (now housing the central post office); and the Convento de Santa Clara, built in 1644 and protecting within its great interior patio some of the city's earliest houses and streets.

Other fine architecture may be found in El Vedado (The Park), an area of the city characterized by large, modern buildings. Outstanding among these are the Hotel Habana Libre, the structure known for its builders as FOCSA (Ferrocarriles Occidentales de Cuba, Sociedad Anónima), the medical insurance building, the Sugar Industry and



Central Havana.

External Commerce ministries, and the Instituto Cubano de Radiodifusión (Cuban Broadcasting Institute) building. The Cementerio de Colón (Columbus Cemetery) contains many important historical monuments.

The city has numerous parks, notably the Parque Central; the Alameda de Paula next to the port; the Plaza de Armas, completed in 1792; the Parque de la Fraternidad, formerly the Campo de Marte; and the Plaza de la Revolución, as well as the Antonio Maceo and Máximo Gómez parks and the park of the Mártires de la Revolución (Martyrs of the Revolution).

The avenues

Traffic flow is facilitated by many first-class avenues. These include the Malecón, an avenue extending several miles from the port entrance and along the coast to the outlet of the Río Almendares, where it enters a tunnel beneath the river and emerges at Avenida Quinta (Fifth Avenue) of Miramar. Another very long avenue is Línea, in El Vedado, which crosses this entire section and then passes through the same tunnel under the river to Marianao Norte, where it becomes another avenue crossing the entire residential area and proceeding miles farther to a junction with an avenue completely encircling the urban perimeter. Other avenues and streets include the Paseo de José Martí (formerly the Prado), which follows the old wall, and the Avenida del Puerto, the Calzada de Belascoaín, Calzada del Cerro, Calzada de Galiano, and Avenida de la Independencia (formerly Boyeros). Multiple housing units, constructed after the revolution to house former residents of the city's slum areas, add a major new element to the urban landscape.

The people of Havana. The reorganization of Cuba in October 1976 extended Havana's city limits to what was formerly the boundary of the metropolitan area; the area within these boundaries was divided into 15 municipalities, replacing the six divisions into which the former metropolitan Havana had been divided. The city's present constituent parts are: Arroyo Naranjo, Centro Habana Norte, Centro Habana Vieja, Cerro, Cotorro, Diez (10) de Octubre, Guanabacoa, Habana del Este, Lisa-Punta Brava, Marianao, Marianao Norte, Regla, San Miguel del Padrón, Santiago de las Vegas, and Vedado. With 6,650 inhabitants per square mile, it is also Cuba's most densely populated area. Metropolitan Havana's birth rate approximates 20.2 per 1,000, its death rate 8.4, and its infant mortality rate 39.8.

The population of Havana, like that of the rest of Cuba,



Castillo del Morro (Morro Castle), Havana.
Gus Parras

is very heterogeneous. Its people include whites, who are basically descendants of the Spaniards; blacks, descendants of former slaves; and mestizos (of mixed ancestry). In the 19th century there was an immigration of Chinese to the island; in the first half of the 20th century, an influx of Spaniards, as well as Italians, French, and Germans; and, after World War II, Jews from central Europe. The city's population has increased steadily since its founding: 51,307 inhabitants were recorded in 1791; 96,304 in 1811; 184,508 in 1841; 253,418 in 1899; 363,506 in 1919; 542,522 in 1931; 676,376 in 1943; 1,880,700 in 1958; and 1,751,216 in 1970. In 1975 the population of 1,900,000 occupied an area nearly 100 square miles (260 square kilometres) larger than the area within the boundaries of 1970.

Population growth

In the 20th century, Havana's population increases in periods of economic crisis, when masses of landless farmers come to the capital seeking employment. Statistics indicate that when the sugar industry (Cuba's principal economic activity) is flourishing, the rural population grows at a more rapid rate than that of Havana, and that the reverse occurs in periods of economic crisis. This demographic rhythm has become less pronounced since the revolution, a result of government policy to develop other parts of the country and alleviate excessive metropolitan growth.

Economic life. The Cuban government is attempting to take advantage of the capital's existing infrastructure (*i.e.*, its technical and administrative resources and the competence and relative abundance of its work force), which is the nation's most developed, but only in those activities where its economy complements that being developed in the rest of the country. By the late 1960s, Havana had more than 885,000 inhabitants of working age, of whom almost 550,000 were economically active; some 290,000 were housewives; some 44,000 were nonworking students; and the rest were engaged in other forms of housework or were not working. The active labour force was divided into production (about 374,000) and services (155,000). Within the first group, industrial workers numbered 133,000; agricultural, 57,000; commercial, 64,000; transportation, 46,000; and construction, 46,000. In services, the two largest sectors (each with almost 40,000 workers) were education and health. More than 12,000 persons were self-employed.

Havana is Cuba's most important industrial, importing, and distributing centre, with a considerable part of the manufacturing and processing industries of the country concentrated in its environs. The different branches include the food processing industry, with numerous factories and a wide range of products; shipyards, devoted primarily to building various kinds of fishing boats; and the automotive industry, notably the construction of medium-sized urban and rural buses. Beverages and liquors produced include liquors and rums. Textile manufactures include knitted fabrics and ready-to-wear clothing, while the leather and footwear industry produces boots and shoes. Other activities include those associated with electric power plants; gasworks; fertilizer plants; printing (especially of books of all kinds); paper manufacturing; furniture and packaging; metallurgy; a wide range of pharmaceutical and chemical products including perfumes, soaps, detergents, plastics, and stock feed; and petroleum refineries supplying the greater part of the country's demand. The quality products of the tobacco industry, notably Havana cigars, have brought Cuba world fame. The fishing industry is particularly notable; the building of the Puerto Pesquero de La Habana (Havana Fishing Port), with all its auxiliary installations, together with the constant growth of the fleet itself, led to a quadrupling of the fish catch over the 1960s.

Expansion of the fishing industry

Most of Cuba's imports and an appreciable amount of its exports pass through the port of Havana; the total tonnage handled now exceeds the port's effective unloading capacity, making larger installations a necessity.

Since 1967, the larger part of the population of the city has been involved, directly or indirectly, in the development of the Cordón de La Habana (Havana Cordon, or Green Belt), a belt of 71,530 acres (28,948 hectares), 60 percent of it arable. In this vast sector, plantains (*plátanos*) and root crops (manioc, *malanga*, yams, and so on) are produced for home use as well as for the urban domestic market; crop failures in this labour-intensive agriculture have been largely eliminated by the use of irrigation systems. Crops from perennials, such as coffee, citrus fruits, mangoes, soursops, and guavas, are also harvested. Many people engage in dairying, poultry production, and pig raising. New roads, new urban nuclei for workers, and additional plants for processing such products as coffee and fruits all have aided in the dispersion of population.

Political, governmental, and social institutions. Havana is the political, administrative, and cultural centre of Cuba. It is the seat of the revolutionary government and of the Central Committee of the Communist Party of Cuba, as well as of all state bodies directing governmental tasks, including 31 ministries. Judicial power resides in the Tribunal Supremo Popular, divided into four chambers. Other important government organizations include the Central Planning Board (Junta Central de Planificación, or Juceplan), the agency charged with the central planning of the economy and with allocation of resources; the National Institute of the Tourist Industry (Instituto Nacional de la Industria Turística, or INIT); the National Fisheries Institute (Instituto Nacional de la Pesca, or INP); the National Institute of Agrarian Reform (Instituto Nacional de Reforma Agraria, or INRA); and several other bodies concerned with sport and recreation, technology, forestry, and agriculture. All these agencies are directly under the national Council of Ministers.

Government-sponsored mass organizations also play an important role in public life in the contemporary city and in the country. Paramount is the Partido Comunista de Cuba (PCC; Communist Party of Cuba), the only political party and the organizing and directing force of all the elements of the Cuban government. The Unión de Jóvenes Comunistas (UJC; Communist Youth Organization) is for young persons between the ages of 14 and 27; it absorbed the National Federation of University Students, important in student affairs before the revolution. An adjunct of the UJC is the Pioneros (Pioneers), for children between the ages of 7 and 14. The Federación de Mujeres Cubanas, or FMC (Federation of Cuban Women), founded in 1960, and other organizations also help advance the ideals of the revolution.

Mass organizations

The Committees for Defense of the Revolution (Comités de Defensa de la Revolución, or CDR) are a particu-

larly distinctive feature of present-day Cuban life. Their members—estimated to include almost 25 percent of the national population—are organized by districts throughout the country, and their chief function is to watch for counter-revolutionary activities on the part of their neighbours and report them to the national security apparatus; they are referred to by hostile critics as neighbourhood spy rings. They serve other functions as well, including public health and Marxist indoctrination of the public.

All these bodies are interconnected in their functions, and their objective is continued transformation of Cuban society. As the Communist authorities emphasize, increasing mass participation in what is regarded as an ever changing revolutionary process makes new forms of organization and new institutions necessary.

Services. Havana is the centre from which communications radiate to the rest of Cuba, a particularly disadvantageous situation in view of the length and narrowness of the island.

Transportation

Havana was the first city in all of Latin America to possess a railroad, a line inaugurated in 1837 between Havana and Bejucal. Railroads link Havana with the provincial capitals and other major cities. The city is also connected with the rest of the country by bus routes. The José Martí International Airport (Aeropuerto Internacional José Martí) provides flights to the principal Cuban cities and links the capital with foreign centres.

There is heavy traffic within the city itself, consisting almost exclusively of buses, which transport more than 900,000,000 passengers annually. There are also taxi services at hospitals, hotels, bus and railway terminals, and the airport. The port of Havana is visited by ships of many nations; shipping services are also available to and from Cuban ports.

The city maintains the usual modern urban services. The telephone service is well developed, and public telephones are free. The capital contains 10 radio stations, most of which broadcast to the entire country, and 14 shortwave stations broadcasting cultural, entertainment, musical, and news programs. Two television channels transmit a variety of programs; an educational television channel operates from the José Antonio Echevarría University City.

The city has 10 general hospitals, 26 clinics, and supplementary specialized clinics and dispensaries. A notable feature is the 100 or so nurseries (*círculos infantiles*), with about 20,000 children under five years of age enrolled by the end of the 1960s.

Commercial and service establishments include some 6,000 food distribution centres, almost 1,500 for industrial products, and more than 4,000 cafeterias and restaurants, as well as numerous hotels. A student and worker dining room service, with more than 3,400 establishments, supplies free meals to students and very low-priced meals to workers.

Cultural life and recreation. Spanish is spoken in Havana, as in the rest of Cuba, for Spain was the country's most important cultural influence. The capital is prominent in national cultural life. There are numerous centres

Education

of instruction in the city, outstanding among them being the University of Havana (Universidad de La Habana), which has about 54,000 students and an international reputation. In addition, there are some 2,400 educational centres of various types, with a total enrollment of almost 325,000 students; primary schools alone account for 200,000 students. Education has been given high priority since the revolution, initially with a campaign to eradicate illiteracy. By the start of the 1970s, emphasis had shifted to experiments with new methods and forms aimed at giving more students exposure to a university education. All education is directed by the national government.

Mass cultural activities are primarily under the direction of the National Council of Culture (Consejo Nacional de Cultura, or CNC). They include theatrical events, activities for children, literary functions, and many activities for hobbyists. Attendance at public exhibitions of various kinds is high.

Havana is the home of the Ballet Nacional de Cuba (National Ballet of Cuba), which, together with its first star, Alicia Alonso, has achieved continuous successes in its performances at home and abroad. The Havana Symphony Orchestra offers regular concerts, and there are numerous other musical groups performing for the public, with annual attendance approaching 500,000. Classes in ballet, modern dance, theatre, painting, sculpture, and music are available.

Literary activities in the city are encouraged by contests held by such government-sponsored institutions as the National Union of Writers and Artists of Cuba (Unión Nacional de Escritores y Artistas de Cuba, or UNEAC) and the House of the Americas (Casa de las Américas), the latter being international in character, and also by the University of Havana. The city has some 16 libraries, and the National Library (Biblioteca Nacional), with 351,000 cataloged volumes, is outstanding. Film making in the city is under the supervision of the Cuban Institute of Cinematographic Art and Industry (Instituto Cubano del Arte e Industria Cinematográfica, or ICAIC); its distinguished products, with a documentary emphasis, have won international recognition. Film projection equipment mounted on vehicles offers free programs in schools, parks, hospitals, factories, and farms.

Scientific research in the city is stimulated and coordinated by the Academy of Sciences of Cuba and other related institutions. These include the National Centre of Scientific Research (Centro Nacional de Investigaciones Científicas, or CNIC) and the Cuban Institute for Research into the Derivatives of Sugarcane (Instituto Cubano de Investigaciones de los Derivados de la Caña de Azúcar, or ICIDCA).

Newspapers

There are two newspapers in Havana, both official organs: *Granma*, the voice of the Central Committee of the Communist Party of Cuba (named for the yacht that carried Fidel Castro and his followers to Cuba in December 1956), and *Juventud Rebelde* ("Rebel Youth"), speaking for the Young Communist Organization. There are numerous popular magazines, most of which are distributed nationally and a few of which circulate throughout Latin America. Publications from other Communist countries are available on newsstands.

By the late 1960s, the city had almost 1,000 separate sports installations. Important centres are the Ciudad Deportiva (Sports City), containing the offices of the National Institute of Sports, Physical Education and Recreation (Instituto Nacional Deportes, Educación, Física y Recreo, or INDER) and noted for its architectural beauty and its large capacity (16,000). The Latin-American Stadium (Estadio Latinoamericano) has a capacity of 58,000, and the Juan Abrahantes University Stadium (Estadio Universitario Juan Abrahantes; capacity 10,000), with multiple installations, and the Pedro Marrero Stadium (Estadio Pedro Marrero; capacity 15,000) are also important. All sporting events are free to the public. Havana has also been the site of several international sporting events, including chess and fencing championships.

The city possesses a large and very popular zoo, the Parque Zoológico, and an aquarium containing the principal tropical marine species. On the right bank of the Río Almendares, in the Miramar section, the Parque Almendares is a well-equipped recreation area for children and adults. Also noteworthy are the social clubs located on the beaches bordering the city, which provide facilities for water sports and also include restaurants, cafeterias, and dance halls.

BIBLIOGRAPHY. Basic historical works are EMILIO ROIG DE LEUCHSENRING, *La Habana: apuntes históricos*, 2nd ed., 3 vol. (1963–64); JULIO LE RIVEREND, *La Habana: biografía de una provincia* (1960) and *Historia económica de Cuba*, 2nd ed. (1965); and MANUEL MORENO FRAGINALS, *El Ingenio*, vol. 1, *La Habana* (1964). ANTONIO NUNEZ JIMENEZ, *Geografía de Cuba*, 3rd ed. (1965); and the ACADEMIA DE CIENCIAS DE CUBA, *Atlas nacional de Cuba* (1970), give a general survey of the urban characteristics, boundaries, and demography of Havana. Economic life is covered in detail in the Cuban economic section of the *Economía y desarrollo* (1970–).

(I.E./R.E.Cr./Ed.)

Hawaii

A group of volcanic islands in the central Pacific Ocean, Hawaii was characterized by the American Mark Twain as "the loveliest fleet of islands that lies anchored in any ocean." It became, in 1959, the 50th state of the United States of America.

By the early 1970s, it was economically vigorous, with diversified agriculture and manufacturing; strategically important to the global defense system of the United States; a Pacific Basin transportation and cultural centre, often called "the Crossroads of the Pacific"; and a major tourist mecca. Hawaiian activities gaining in national and international importance by the early 1970s include research and development in oceanography, geophysics, astronautics, satellite communications, and biomedicine.

The capital city of Honolulu, on the island of Oahu, is 2,397 miles from San Francisco to the east, and 5,293 miles from Manila, in the Philippines, to the west. For information on related topics, see the articles PACIFIC ISLANDS; OCEANIA, HISTORY OF; UNITED STATES; and PACIFIC OCEAN.

THE HISTORY OF HAWAII

The early Hawaiians

The first Hawaiians are thought to have reached the islands by canoe about AD 400, and for nine centuries such contacts continued with other parts of Polynesia, notably Tahiti. Powerful classes of chiefs and priests emerged and established themselves, as did internecine conflicts not dissimilar to the feudal struggles in Europe, with complicated land rights contributing to the disputes. The early Hawaiians lacked a written language, and their culture was entirely oral, rich in myth, legend, and practical knowledge especially of animals and plant life. The material life of the islands was hampered by the lack of metal, pottery, or beasts of burden, but there was great skill in the use of wood, shell, stone, and bone, and the huge outrigger canoes were technical marvels. Navigational methods were well developed, and there was an elaborate calendar. Athletic contests encouraged warrior skills.

European discovery. Capt. James Cook, the English explorer and navigator, is regarded as having made the first European discovery of Hawaii, first landing at Waimea, Kauai Island, on January 20, 1778, returning the following year to meet his death in an affray with a number of the local inhabitants at Kealakekua Bay, Hawaii.

The initial discovery was followed by a period of intermittent contact with the West during which the remarkable Kamehameha I, utilizing European military technology and weapons, emerged as an outstanding Hawaiian leader, seizing and consolidating control over most of the island group. For 85 years thereafter, monarchs ruled over the Hawaiian Kingdom. In the early years of the 19th century, the American whaling fleet took up the practice of wintering in Hawaii, and the islands were visited with mounting frequency by explorers, traders, and adventurers. Capt. George Vancouver had introduced livestock to the island in 1794. In 1820 the first of 15 companies of New England missionaries arrived. By midcentury there were frame houses, horse-drawn vehicles, schools, churches, taverns, and mercantile establishments. A written language had been introduced, and European and American skills and religious beliefs—Protestant and Catholic—imported; Hawaiian culture was irrevocably changed.

Establishment of United States dominance. Political manoeuvring between United States, British, and French consuls and naval forces caused uncertainty in the governmental situation. The foundations of constitutional government were nevertheless laid down with the promulgation, by Kamehameha III, of a Declaration of Rights (June 7, 1839), an Edict of Toleration (June 17, 1839), and a written constitution (October 8, 1840). This step forward—made under missionary influence—was followed by formal avowals of Hawaiian independence by the United States, Britain, and France. The ambitions of these powers continued unabated, however, with a succession of overt and covert diplomatic moves, culminating in the signing of a reciprocity treaty with the United States in 1875. Hawaiian kings continued to attempt to preserve their peoples' culture and society, but the turbulent second half of the 19th century was to witness, by the joint annexation resolution of Congress in 1898, the final establishment of United States domination. This status was confirmed by the establishment of a territory on June 14, 1900.

The period until 1940 was to see a great growth in population, the development of a modern economy based on the production of sugar and pineapples for consumption on the mainland, and the growth of transport and military links. Movements for statehood, based in part on the fact that Hawaii had to pay United States taxes without corresponding legislative representation, began to make themselves heard. The Japanese attack on Pearl Harbor, on December 7, 1941, precipitated not only Hawaii but the United States as a whole into the conflict of World War II, and the islands witnessed an upsurge of military activity and a sometimes controversial curtailment of civil liberties. The post-1945 period was marked by further economic consolidation and a long constitutional path to statehood, a status finally achieved in 1959.

THE NATURAL AND HUMAN LANDSCAPE

The natural environment. *Surface features.* The state actually consists of the tops of a chain of submerged volcanic mountains that form eight major islands and 124 small islets, stretching in a 1,500-mile crescent from Kure Island in the west to the island of Hawaii in the east, with a combined land area of 6,425 square miles (16,641 square kilometres). With the exception of Midway, a United States naval reservation near the western end of the archipelago, the leeward coral atolls and central lava islets, forming a total of only 3.2 square miles, are in the Hawaiian Islands National Wildlife Refuge.

Islands comprising Hawaii

The eight major islands at the eastern end of the chain are, from west to east, Niihau, Kauai, Oahu, Molokai, Lanai, Kahoolawe, Maui, and Hawaii. The volcanic activity that molded the island landscapes has long since become dormant, with the exception of the volcanoes of Mauna Loa and Kilauea on the easternmost and largest island, Hawaii, where spectacular eruptions and lava flows take place from time to time. The highest Hawaiian mountains are Mauna Kea and Mauna Loa, reaching 13,796 feet (4,205 metres) and 13,677 feet (4,169 metres) above sea level, respectively.

There is little sign of erosion in the geologically young areas, where the terrain is comparatively domelike and the volcanic craters are clearly defined. In the contrasting older areas, during the glacial age, the mountains were shaped and eroded by ice and by the action of sea, rain, and wind. Their aspects thus include sharp and craggy silhouettes; abrupt, literally grooved cliffs pocked with caves; deep valleys; smoothed saddle areas; and coastal plains. The powerful Pacific surf, churning and crashing against the fringing coral shelves and the lava shorelines, has carried minute shells onto the shore and reduced coral and large shells to sand, creating the state's famous expanses of beach.

Soils and drainage patterns

Volcanic ash, gravel, rotted vegetation, crumbling lava, and windblown sand and dust all help to make up the alluvial, residual, and organic soils found in various depths and densities in valley floors, the regions between mountain ranges, and along the shores. Oxidation of iron causes a ubiquitous, bright-red soil and rock strata. The iron content is, however, insufficient for smelting, and there are no coal or oil deposits.

As the topography is generally abruptly descending or sloping, there are few surface collecting basins or lakes. Excess rainfall seeps through porous mountain areas to collect in subterranean chambers and layers retained by less permeable lava and ash beds or is prevented by underlying salt water from seeping to the sea. The resultant artesian water supply is tapped by man in many places, for use in irrigation and also for human consumption.

Where rainwater has converged in high mountain

pockets and then plunged great distances, it has dug deep pools at the foot of the falls, overflowed, and cut streams downward in its pursuit to the sea. Some streams and small rivers meander and wind through wide valleys to lowlands and debouch concisely into the ocean or disperse themselves into swampland.

Climate. Though Hawaii lies in the earth's tropical zone, its climate is temperate. The Pacific anticyclone, a large atmospheric eddy located northeast of Hawaii, is the generator of prevailing trade winds from the northeast, which are cooled by their journey over the ocean. The ocean is, in effect, a natural air-conditioner.

The average temperature in Honolulu is 71.9° F (22.2° C) in the coolest month and 78.4° F (25.8° C) in the warmest, with extremes from 57° F (14° C) to 88° F (31° C) having been recorded. Average water temperatures off Waikiki Beach, near Honolulu, range from 75° F (24° C) in late February to 79° F (26° C) in late September. Mountainous regions are considerably cooler, especially during the winter months, when there can be frost; temperatures as low as 14° F (−10° C) have been recorded on the summit of Haleakala, on Maui, and winter snows frequently blanket the crests of Mauna Kea and Mauna Loa, on Hawaii.

Rainfall variations throughout the state are dramatic. Mt. Waialeale, on the island of Kauai, is probably the wettest spot in the world, with an annual rainfall of 486.1 inches. The driest area in the islands is at Puako, on the island of Hawaii, where the average annual rainfall is only 9.46 inches. The average yearly rainfall in Honolulu is 23.9 inches, and, in Hilo, it is 136.62 inches.

As moisture-laden air is carried over the islands, most frequently by the trade winds, it is apt to condense, form cloudcaps, and dissipate against the shores and mountains of the windward coasts, which are therefore more lush in foliage than the leeward coasts.

Plant and animal life. The seeds of endemic plant species were carried to Hawaii by birds, riding winds, or moving in currents and tides, bringing about extensive forestation, shrubbery, and grasslands where soil and precipitation were favourable. From the time of the first Polynesian settlement to the present day, a tremendous variety of food and ornamental plant life from many parts of the world has been introduced. Food plants grown commercially or in backyards for home consumption include sugarcane, pineapples, papayas, bananas, mangoes, guavas, lichee, coconuts, avocados, breadfruit, macadamia nuts, limes, passion fruit, taros, and tamarinds. Nearly all varieties of common garden vegetables are raised in the islands, and flowers abound all year.

The effects of isolation on natural life

Endemic birds, long isolated from others of their kind, have taken on certain characteristics of their own. These include the Hawaiian goose (nene), the Hawaiian stilt, and a variety of small forest birds. Some species have become extremely rare, but as the result of an increased environmental awareness, great strides have been taken in recent years to preclude their extinction. Seabirds nest in profusion on the western islands of the archipelago and to a far lesser extent among the major eastern islands. There has been considerable importation of birdlife during the last 100 years. Quantities of mynas, sparrows, cardinals, and doves inhabit the trees in both urban and country areas. Every fall the small golden plover make an awe-inspiring, nonstop 3,000-mile flight from Alaska to Hawaii, where they spend the winter, together with ducks from Alaska, Canada, and the northwestern United States.

Most forms of common domestic animals and poultry are raised on farm and beef ranches, and a large percentage of Hawaiian households keep dogs and cats as pets. What wild-animal life there is consists of mongooses, rats, frogs, toads, and, in the more remote regions of some of the ranches, deer, pigs, sheep, and goats. The Hawaiian insect population is multitudinous, and marine life abounds in Hawaiian waters.

Patterns of settlement. Agricultural and fishing activities bring about extensive and scattered rural settlement, ranging from tiny fishing villages far off the main roads, scant clusters of small houses in isolated valleys, solitary farm and ranch houses, to large coastal and upland villages and plantation and ranch towns.

The older houses in the smaller villages are largely single-family, raised, frame structures, with corrugated-iron roofs. Plants of native origin skirt the foundations of houses, and the yards are informally planted with fruit and flower trees. In all but the very small villages, there are a school, markets, post office, firehouse, and at least one church. The day's activities traditionally begin early and end early, following the sun. The life-style of the rural people appears to be simpler and less sophisticated than that of the urban populations, who are exposed to constant and varied activity, and the country dwellers tend to retain more of the speech patterns and customs of their distinctive ethnic backgrounds.

Village and urban life

During the 1950s and 1960s, Hawaii experienced a building boom of such magnitude that the configuration of entire towns was altered. The most graphic example of this was in the city of Honolulu, where construction of 20- and 30-story buildings gave the city, once sprawling and low, a thrusting, many-levelled skyline. On Oahu, erstwhile vacation or agricultural towns have become expansive residential areas for commuters to Honolulu and Pearl Harbor.

Urban settlement once consisted almost entirely of single-family dwellings, individual business houses and shops, small markets, and three- or four-story hotels; but, with the increase of resident and touring population during the preceding two decades, Hawaiian towns and cities were, by the early 1970s, building more and more high-rise apartment houses, hotels, and business establishments, with the traditional individual shopkeepers becoming absorbed into the complexes of shopping centres and supermarkets.

The planned city, a new housing concept in Hawaii, is gathering momentum in the 1970s and is being carried out in areas that were previously open spaces or given over to agriculture.

Throughout the state there are large hotels and resort complexes, creating communities that are not without colour in themselves.

New construction throughout the state is thriving, and thus the physical aspect of Hawaii, in terms of the landscape under human settlement, continues to undergo considerable change.

THE PEOPLE OF HAWAII

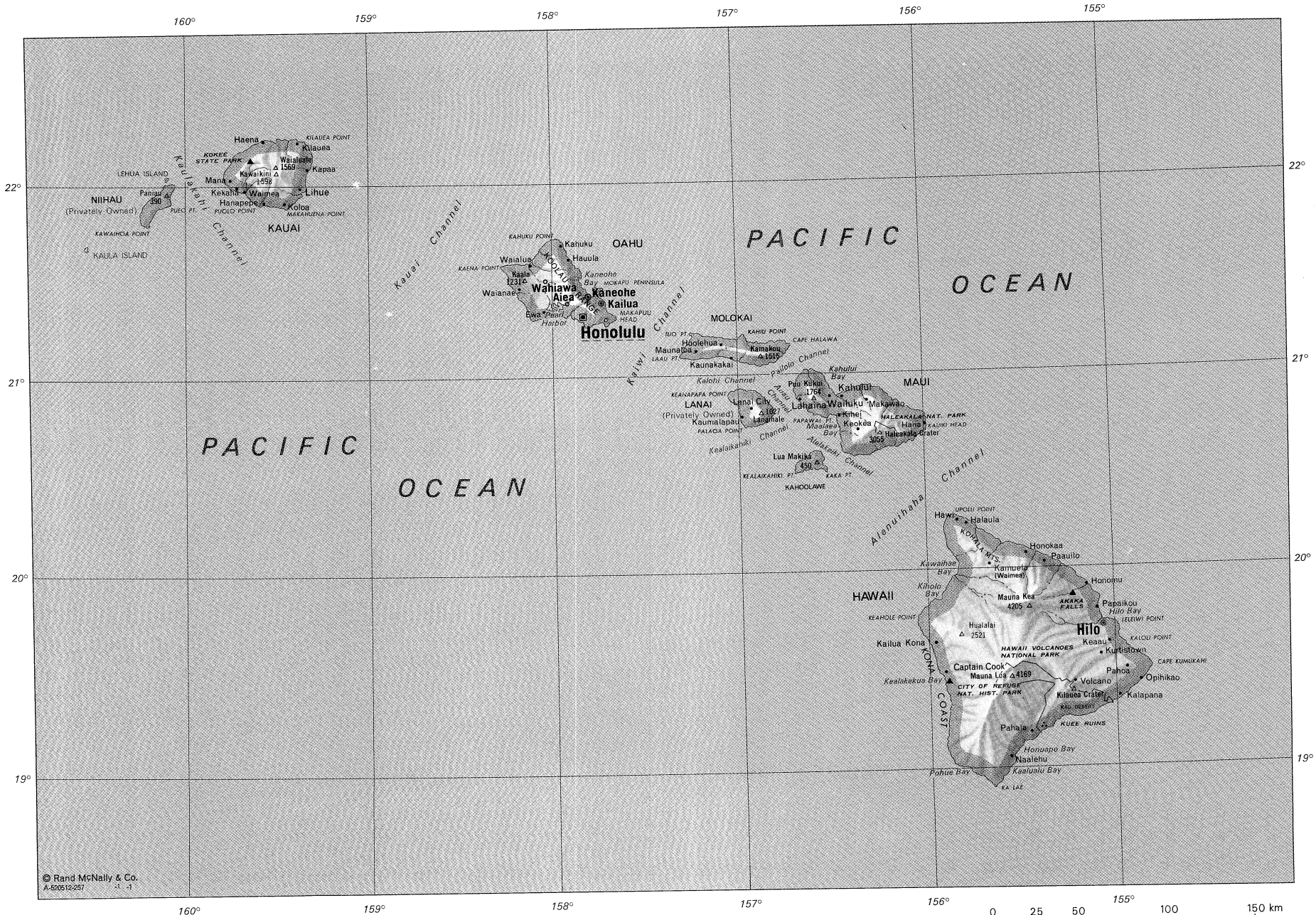
Ethnic origins. Most anthropologists feel that the original settlement of Hawaii was by Polynesians who migrated northeast from the Marquesas Islands perhaps as early as AD 400, to be followed by a second wave of immigration from Tahiti in the 9th or 10th century. Once having established themselves in Hawaii, the Hawaiians had no further need for supplies from their old homeland and underwent centuries of isolation. Although there remained close resemblance in linguistics, physical characteristics, and general customs and lifestyles between the Hawaiians and their Polynesian relatives, a degree of racial individuality evolved.

The original Hawaiians were a brown-skinned people of large stature, highly skilled in fishing and farming, who adhered to an extremely rigid and strict system of laws set down by their chiefs and the priests. They worshipped and feared a group of gods not unlike, in character and power, the ancient Greek deities of Olympia.

The first recorded contact between the Hawaiians and the people of the European world took place in 1778, when the English explorer James Cook came upon the islands. During the ensuing four decades, the influence of further European and American explorers, adventurers, trappers, and whalers stopping for fresh supplies at Hawaiian islands was to have a profound effect.

Impact of the cultural mix

Contact with men of totally different cultures, with a belief in only one god substituted for fear of swift punishment from vengeful gods, eventually caused a spiritual revolution among the Hawaiians. In a series of defiant acts led by members of the royal family, the basic beliefs of the Hawaiian religion were undermined, the priests were overthrown, and by 1820—the year



HAWAII

Size of symbol indicates relative size of town • ○ ◉ ■

Elevations in metres

when the first company of Christian missionaries arrived from America—Hawaiians were experiencing something of a religious void. Loss of faith in the old gods, intense interest and curiosity about the ways of the people of America and Europe, avid interest in learning to read and write, and a desire for spiritual identity caused a swift adoption of Christianity on the part of the Hawaiians. By the mid-19th century, the Hawaiian kingdom was largely a Christian nation. It has been estimated that the population of the Hawaiian Islands at the time of Captain Cook's discovery was approximately 300,000. Virtually disease free, this population had no natural immunity to the diseases introduced from both west and east and fell easy prey to venereal disease, cholera, measles, bubonic plague, and leprosy, all of which contributed to the decimation of the native peoples.

It is recorded that in 1853 the population of the Hawaiian kingdom consisted of 70,036 native Hawaiians, 983 part-Hawaiians, 1,687 Caucasians from Europe and the United States, 364 Chinese, five Filipinos, and 62 people of other racial extractions, making a total of 73,137 people.

The racial and religious makeup of Hawaii has undergone dramatic change since then. Thousands of settlers from the Pacific Basin—primarily from China, Japan, Korea, and the Philippine Islands—as well as immigrants from Europe and from America, carried their own customs, languages, and religions into the Hawaiian way of life. Today, they and their descendants now far outnumber the descendants of the original Hawaiians.

The contemporary population. In 1970 there were 16,591 live births in Hawaii, a 21.5 rate per 1,000 civilian resident population, and 4,216 deaths, a 5.5 rate per 1,000 civilian resident population, indicating the large natural rate of increase in the state. In the ten years between 1960 and 1970, 140,000 American civilians and 30,000 persons from foreign countries, largely the Philippines, moved to the state, while only 123,000 residents moved away. This, too, contributed to population growth. There is also a continuous influx and outflow of military personnel and their dependents, linked closely to the American presence in Southeast Asia.

The majority of the state's 769,000 people reside on the island of Oahu, with 325,000 in the city of Honolulu proper and 305,000 in outlying districts, making a total of 630,000 residents. As there are vast areas of Oahu devoted to agriculture and forest reserve, the majority of the population actually resides in high-density clusters. The high birth rate and continuing influx from the mainland United States and from foreign countries lead statisticians to expect the population to pass the million mark in the not too distant future.

Linguistic diversity

Linguistically, Hawaii is English speaking. Although Hawaiian, formerly a major means of communication, is all but extinct, it remains alive in place-names and street names and in songs, and the local residents liberally sprinkle their speech with words and phrases from the traditional language. A pidgin English, differing from standard English in both word usage and inflection, is spoken throughout the state in varying degrees of richness, while some of the older immigrants from Japan and China continue to speak their native tongues. As Filipinos continue to move to Hawaii, their language, too, is frequently heard in the state.

THE STATE'S ECONOMY

On the national scene, Hawaii ranks between 38th and 45th among the United States in terms of personal income, farm products sold, value of manufacturing shipments, retail sales, and bank deposits. Its cost of living and personal income per capita are among the highest.

Resources and power. Since Hawaii has no significant mineral deposits, its only natural resources are its temperate climate, water supply, soil, vegetation, and surrounding ocean and rock, gravel, sand, and earth quarried for use in construction and landscaping. Electric power is supplied throughout the state by a small number of power companies operating oil-powered steam and diesel generators. Certain military installations and some private institutions generate their own power, and a small amount of hydroelectric power is generated on the islands of Hawaii and Kauai.

Agriculture. Although the largest source of income to the state of Hawaii is the federal government, primarily through defense expenditures amounting to over $650,000,000 per year at the start of the 1970s, and while tourism annually attracts over 1,000,000 visitors to the state, with gross expenditures of more than $500,000,000, agriculture remains the backbone of the local economy.

Sugar and pineapples are the principal crops, the plantations maintaining high yields through the use of sophisticated irrigation systems and planting and harvesting machinery. Hawaii produces 40 percent of all of the canned pineapple consumed in the world. There is extensive diversified agriculture, livestock, poultry, and dairy production, together with some lumbering and commercial fishing. Approximately half of the commercial fish catch is aku (skipjack tuna).

Manufacturing. Hawaii has more than 600 companies engaged in diversified manufacturing. Heavy-manufacturing plants, using raw materials for the most part imported from the United States mainland, include an oil refinery that produces a variety of petroleum products and chemical compounds, a steel mill manufacturing reinforcing bars, two cement plants, a concrete-pipe plant, and an aluminum-extrusion plant. Most building lumber is imported from the mainland. Eighty garment manufacturers, largely situated in Honolulu, produce printed fabrics and apparel bright with colours and designs inspired by the local flora and marketed locally, nationally, and abroad.

A wide variety of Hawaii-grown foodstuffs are bottled, canned, jarred, or packaged in the state; sold in local grocery stores; and exported to the mainland. These include Oriental and Hawaiian food specialties, tropical fruit juices, jams and jellies, candies, coffee, macadamia nuts, and various alcoholic beverages.

Foreign trade. Exports to foreign countries are largely in the form of sugar, canned pineapple, garments, flowers, and canned fish. Major foreign imports are fuel, vehicles, food, and clothing.

Management of the economy. By mid-1969, some 13,000 corporations and partnerships were registered to do business in Hawaii, and 51 percent of the land in the state was owned by private individuals or corporations, although the state itself, holding 39 percent of the land, was the largest single landowner. State and county governments hire a larger number of employees (nearly 40,000) at a higher wage scale than do most of the other states. Hawaii is the regional headquarters of the federal government, which owns 10 percent of the land in the state and employs 34,700 persons.

State and private interests

Taxation. By the 1970s, annual fiscal tax collections by federal, state, and county governments totalled almost $1,000,000,000, with nearly two-thirds in federal taxes, almost a third in state taxes, and the remainder in licenses and fees from the four counties making up Hawaii's local government system. There are no municipal governments or incorporated villages. State taxes are collected under a centralized tax system. The chief sources of revenue are a general excise tax, corporate income taxes, and real property tax.

Trade unions. Major Hawaiian industries are unionized, as are many service and construction industries. The largest union in the state, and one with a turbulent history, is the International Longshoremen's and Warehousemen's Union.

Economic prospects. Economic policies, influenced by and relative to the economic state of the country and the world, tended toward the conservative at the start of the 1970s. Investments, by Hawaii-based financiers in the mainland United States and abroad, increase yearly, as further outlets for economic energy are located, and investment in Hawaii by mainland and foreign capital is on the increase. Within the confines of the state, a certain mood of cautiousness regarding overextension prevails, in sharp contrast to the headlong development, particularly in construction, characteristic of the decade of the

Movements of capital

1960s. Governmental budgets indicate increasing allotments for public housing and recreation and for the preservation of the environment.

A major economic problem in Hawaii is the high cost of living due in large part to Hawaii's insularity and dependency on imports. Transportation costs are included in the purchase prices of nearly all consumer goods. As population increases, it finds housing difficult to acquire and when acquired, by purchase or lease, disproportionately expensive when compared with housing costs in many of the mainland states. Prospects for the economic future in the light of views expressed by both private and governmental study groups are optimistic. A major Hawaiian goal is said to be to bring the economy out of the growing pains of the 1950s and 1960s and, hopefully, into a more secure maturity.

It has been suggested that this may be accomplished by limiting permitted encroachment by urban development on agricultural lands. Carefully planned housing, located in communities in which, in addition to high-rise, high-density dwellings, the single-family home gives way to cluster housing around recreational areas, is indicated as one solution to the shortages and expense associated with urban housing. Finally, an increase in the quantity and variety of manufacturing is envisaged for the entire decade of the 1970s.

TRANSPORTATION

Shipping. Ocean surface transportation is Hawaii's lifeline. Consumer goods and raw materials are brought to Hawaii, and Hawaiian goods are exported in freighters, tankers, container ships, and barge "sea trains." Gigantic, high-speed cranes load and unload the containers.

The importance of Honolulu Harbor

Honolulu Harbor, with its extensive docks, warehouses, and storage sheds, is the focal point of Hawaiian shipping, handling over 8,000,000 tons of incoming cargo and nearly 5,000,000 tons of outgoing cargo annually. A large percentage of the cargo ships ply between Hawaii and California ports, a few between Hawaii and the East Coast of the United States via the Panama Canal, and others from western Pacific ports. Around-the-world passenger ships carry approximately 28,000 passengers through Honolulu every year, and passenger liners from the West Coast of the United States alone bring 38,000 visitors to the islands annually. Tug-pulled barges and small freighters transport goods from Honolulu to the outer islands, returning with agricultural crops and livestock.

Air transport. The majority of voyagers to and from Hawaii travel by air, as do nearly all interisland passengers. The Honolulu International Airport, on Oahu, and the General Lyman Field at Hilo, on Hawaii, are the state's two civilian airports capable of serving large-jet traffic. There are 12 smaller airports among the islands and a number of small private airfields. Military authorities maintain a number of airports throughout the state. About 80 cargo planes come in from the east each month, with over 40 from the west, bringing an annual 78,000,000 tons of cargo into the state and departing with in excess of 42,000,000 tons. There are two interisland air carriers.

Roads. Throughout the state, there are 3,529 miles of roads, most of them following the lowland contours, circling the islands along or near the shorelines, and only crossing islands between mountain ranges. In addition, there are many spectacular mountain roads zigzagging down cliff faces. On Oahu two tunnels bring traffic from the heads of two valleys behind Honolulu through the Koolau Mountain Range and out into the windward, or northeastern, side of the island. Hawaiian roads range from narrow country tracks to an eight-laned divided freeway, which crosses the city of Honolulu.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. Hawaii is governed by a state constitution that was originally adopted in 1950, was amended in 1959, at the time of admission to statehood, and further amended at the constitutional convention of 1968. The governor and lieutenant governor, elected for concurrent terms of four years, must be members of the same political party. The only other elected members in the 17 departments of the executive branch are the members of the Board of Education. Hawaii's bicameral legislature consists of a Senate, with 25 elected representatives from eight senatorial districts, serving four-year terms, and a House of Representatives, consisting of 51 members elected from 25 districts for two-year terms. The state judicial system consists of a Supreme Court, four circuit courts, and 27 district courts, a family court, and a land court.

Executive, legislative, and judiciary

Hawaii's governmental structure is unique in the United States in that it is limited to two levels of government: the state and the four counties, each with a mayor and a council. There are no separate municipal governments.

The political process. Primary elections are held in October and general elections in November. Party competition is intense in Hawaiian politics. During the first half of the century, the Republican Party remained dominant, but party success at the polls began to seesaw somewhat after this, and the Democratic Party has captured a majority of House, Senate, and council seats in recent years.

Military influence. Hawaii holds a strategic position in the defense system of the United States. Pearl Harbor, a vast shipyard for repair and overhaul, is home port to many naval ships. It serves as a training base for Polaris submarine and antisubmarine warfare forces. The headquarters of the commander in chief, Pacific, and of the Fleet Marine Force, Pacific, are at Camp H.M. Smith. The major army, Marine Corps, and air force bases are Schofield Barracks, Ft. Shafter, Ft. De Russy (a rest and recreation centre for personnel fighting in Asia), Corps Hickam and Wheeler air force bases, and the Kaneohe Marine Corps Air Station. In addition, there are military installations, camps, and airfields of varying sizes throughout the state. The total of approximately 116,000 United States military personnel and their dependents are stationed or home ported in Hawaii, a fact not without significance in the local economy and social life.

Social conditions. The minimum wage under Hawaii's Wage and Hour Law, at the start of the 1970s, was $1.60 per hour. Agricultural workers in Hawaii were more highly paid than those in most of the mainland states, receiving an average of $2.37 an hour, with other wage earners obtaining an average of between $1.97 and $3.43 an hour. Federal and county employees are among the most highly paid in the country.

The per capita annual income of $4,530 at the start of the 1970s was higher than the national average, although living costs, as has been noted, were among the highest in the country. Hawaii's high health standards are generally reflected in the vital statistics. Expectation of life at birth is approximately 72 years, among the highest in the nation. Hawaii is free of most tropical diseases, smallpox, and rabies.

Hawaii enjoys a unique reputation as a place in which the population, stemming from many different roots, has created a harmonious society; a degree of social division takes place, following the patterns of ethnical and cultural backgrounds, but the groups tend to appreciate and enjoy the variations of other such groups. As in other parts of the United States, poverty and wealth coexist, sometimes in startling contrast.

Education. Hawaii's school system provides educational facilities from nursery school through the Ph.D. level. By the 1970s, there were 217 public schools below the college level, with an enrollment of almost 180,000 and over 7,000 classroom teachers. Well over 100 private schools added over 30,000 students to the total enrollment. At the college level Hawaii has the University of Hawaii, five smaller private colleges, and a state-established system of five two-year community colleges. By the 1970s, nearly 60,000 full- and part-time students were attending colleges and universities. Private business, technical, and specialized schools provide further educational facilities and opportunities.

Higher education

A unique educational institution is the Center for Cultural and Technical Interchange Between East and West,

commonly referred to as the East–West Center. A project of the federal government, it is housed at the Manoa campus of the University of Hawaii and annually provides specialized and advanced academic programs and technological training to 600 students, from countries in Asia, the Pacific, and the United States.

Public Services. There are 31 hospitals in the state and nearly 1,000 licensed doctors. A state department of health maintains hospitals, health centres, clinics, care centres, and nursing services. Annual public welfare costs at the start of the 1970s exceeded $50,000,000, which came from state and federal government funds. The Hawaiian Homes Commission controls the transfer of land use to qualified persons of Hawaiian racial origin for homesteading.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu. Hawaii's cultural milieu, the result of overlay after overlay of varied cultural groups, is rich in its admixture. The force of the original culture remains highly evident, although the Hawaiian race has become diminished and diluted through death and intermarriage.

Uniqueness of Hawaii's cultural heritage

Vestiges of American culture in the New England tradition remain, as do the traditional cultures of the early Asian and Filipino immigrants. With the advent of fifth-, sixth-, and seventh-generation descendants of Asian and Caucasian immigrants and the massive influx of Americans from all parts of the country, the cultural overlays have melded to form a uniquely Hawaiian culture.

Interest in the arts is high in this state, which has produced distinguished artists, photographers, and theatrical and musical performers. Appreciation of classical, modern, and experimental art forms is manifest in attendance figures at galleries, concerts, legitimate-theatre performances, and museums. Many ethnic groups preserve the traditions of their ancestors or combine or modify music and dance forms. In addition, islanders very literally keep in step with whatever dance forms are currently popular in the nation as a whole and make a distinctive contribution to the music of the times.

Cultural institutions. An assortment of cultural and scientific institutions in Hawaii provides a wide variety of opportunity for the appreciation and understanding of the fine arts, history, traditions, and sciences. The Bernice P. Bishop Museum, founded in 1889 in Honolulu, is a research centre and museum dedicated to the study, preservation, and display of the history, sciences, and cultures of the Pacific and its people. The Honolulu Academy of Arts, often called the most beautiful museum in the world, houses a splendid collection of Western art, including works by the late 19th- and early 20th-century masters of modern art: Monet, van Gogh, Matisse, Gauguin, and Picasso. Its collection of Oriental art is also one of the finest in the Western world. The University of Hawaii art, music, and drama departments contribute to the expanding cultural life of Hawaii, while the state has several legitimate-theatre organizations, professional and amateur. The 80-member Honolulu Symphony Orchestra performs 100 concerts a year in Honolulu and on the other major islands and 300 educational concert demonstrations. Its home is the Honolulu International Center, a municipal theatre–concert-hall–arena complex, where touring opera companies and ballet troupes and musical artists of international renown also perform. Honolulu's Chamber Music Society gives a concert series each year. The active art, music, and drama departments in Hawaiian schools and colleges and the innumerable students receiving private instruction in the arts also contribute toward the maintenance and growth of Hawaii's cultural life.

The performing arts

Press and broadcasting. In the city and county of Honolulu, two daily newspapers and several specialized weekly and semiweekly papers are published in English, and there are also two daily papers in Japanese and English and two daily papers in Chinese. Honolulu papers are delivered daily by air to the other islands. A daily paper, a biweekly paper, and two weekly papers are published on the island of Hawaii, and a biweekly paper is published on Maui. There are 28 radio stations and 12 television stations in the state.

Problems and prospects. Long famed for its beauty, Hawaii had, by the 1970s, also acquired distinct national and international stature as a Pacific cultural, transportation, economic, and scientific centre. At that time, the state's future appeared to be one of continued growth in production, manufacturing, investment, foreign trade, agricultural, scientific exploration and research, population, and visitor services, with the government and the private sector utilizing imaginative modern methods and concepts to bring about successful development and continued change while preserving the environment and the unique pattern of Hawaiian life.

BIBLIOGRAPHY. Since Captain Cook penned his journal at the time of his discovery of Hawaii for the Western world, the islands have been the source material and inspiration for hundreds of writers from many nations; the complete bibliography is vast.

An understanding of the insular nature of the state, its geographical reference to the Pacific, and its topography and still-active volcanoes are basic to an understanding of present-day Hawaii as well as its history. HAROLD THORNTON STEARNS, *Road Guide to Points of Geologic Interest in the Hawaiian Islands* (1966), is especially valuable for its descriptions of points of interest along the roadside, maps, statistical tables, and a glossary of volcanic terms. See also GORDON A. MACDONALD and DOUGLASS HOPWOOD HUBBARD, *Volcanoes of the National Parks in Hawaii*, 3rd ed. rev. (1965).

PETER H. BUCK, *Vikings of the Sunrise* (1938, reprinted 1959), covers Polynesian migration in the Pacific by a noted anthropologist. Fact, legend, and myth brought to the written word are a mine of information, passed by memorization through generations, about the early Hawaiians. MARTHA BECKWITH, *Hawaiian Mythology* (1940), is scholarly and valuable.

Personal journals from the late 18th and early 19th centuries form a major source of information regarding life in Hawaii. Outstanding are ARCHIBALD CAMPBELL, *A Voyage Around the World, From 1806 to 1812*, 3rd ed. (1822, reprinted 1967), one of the best accounts of life in Hawaii before any significant settlement by non-Hawaiians; and LAURA FISH JUDD, *Honolulu: Sketches of Life in the Hawaiian Islands from 1828–1861* (1880, reprinted 1966), an excellent documentary by the wife of an American medical missionary.

Good histories include ALAN GAVAN DAWS, *Shoal of Time: A History of the Hawaiian Islands* (1968), one of the best single-volume histories of Hawaii; ANDREW W. LIND, *An Island Community* (1938, reprinted 1968), an excellent in-depth study of the racial migrations of Hawaii and subsequent race relations; and RALPH S. KUYKENDALL and A. GROVE DAY, *Hawaii: A History, from Polynesian Kingdom to American State*, rev. ed. (1961), another good history of the period covered.

Descriptions of flora and fauna include LORRAINE E. KUCK and RICHARD C. TONGG, *Hawaiian Flowers and Flowering Trees* (1958); MARIE C. NEAL, *In Gardens of Hawaii*, new and rev. ed. (1965), the latter containing a plant guide as well as legends; and GEORGE CAMPBELL MUNRO, *Birds of Hawaii*, rev. ed. (1970), the best illustrated guide to Hawaiian birds.

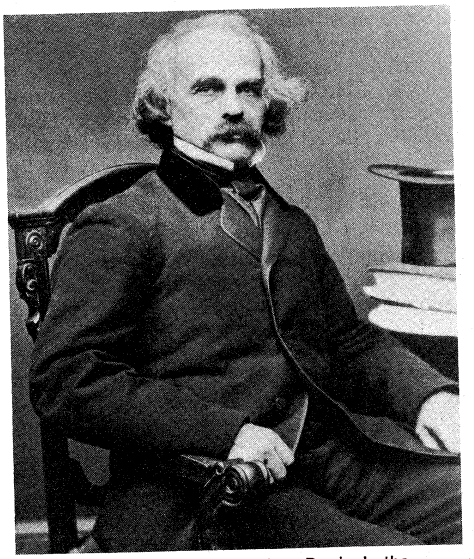
THE AMERICAN INSTITUTE OF ARCHITECTS, HAWAII CHAPTER, *A Guide to Architecture in Honolulu, 1957* (1957), illustrated; and RUTH L. HAUSMAN, *Hawaii: Music in Its History* (1968), are two good books on the arts.

JEAN SCOTT MacKELLAR, *Hawaii Goes Fishing* (1956, reprinted 1968), is a very readable book about different methods of fishing; THOMAS EDWARD BLAKE, *Hawaiian Surfboard* (1935, reprinted 1961), includes stories of surfing and instructions; PETER L. DIXON (ed.), *Men and Waves: A Treasury of Surfing* (1966); BEN R. FINNEY and JAMES D. HOUSTON, *Surfing, the Sport of Hawaiian Kings* (1966), a history and guide; RICKY GRIGG and R. CHURCH, *Surfer in Hawaii* (1963), places to surf and instructions; JOHN MELVILLE KELLY, *Surf and Sea* (1965), a good complete guide, including an explanation of the tides and currents that create waves; and H. ARTHUR and M.C. KLEIN, *Surf's Up! An Anthology on Surfing* (1966), are all recommended books on sports in Hawaii.

(J.P.M.S)

Hawthorne, Nathaniel

Nathaniel Hawthorne was the first American writer to produce a distinguished body of fiction so clearly reflecting American experience, particularly that of 17th-century New England, that one cannot imagine its having been produced in another culture with a different history.



Hawthorne, photograph by Mathew Brady. In the National Archives, Washington, D.C.

By courtesy of the U.S. Signal Corps

Hawthorne was born on July 4, 1804, in Salem, Massachusetts, where his ancestors had lived since the 17th century. The first Hathorne (Nathaniel added the *w* to the name when he began to write) was a magistrate who had sentenced a Quaker woman to public whipping. He had acted as a staunch defender of Puritan orthodoxy, with its zealous advocacy of a "pure," unaffected form of religious worship, its rigid adherence to a simple, almost severe, mode of life, and its conviction of the "natural depravity" of "fallen" man. According to an apocryphal family legend, a later Judge Hathorne, presiding during the witchcraft trials of 1692, was supposed to have brought down upon himself and his descendants the curse, uttered by a woman he sentenced to death, "God will give you blood to drink."

Cursed or not, the Hathornes declined in prosperity and prominence during the 18th century, while other Salem families were growing wealthy from the lucrative shipping trade with China. When Nathaniel's father—a ship's captain—died during one of his voyages, he left his young widow, Elizabeth Manning Hathorne, without means to care for her two girls and young Nathaniel, aged four. She moved in with her affluent brothers, the Mannings. Hawthorne grew up in their house in Salem and, for extensive periods during his teens, in Raymond, Maine, on the shores of Sebago Lake. He returned to Salem in 1825 after four years at Bowdoin College, in Brunswick, Maine. Unlike his classmate Henry Wadsworth Longfellow, who was soon to become a popular American poet, Hawthorne did not distinguish himself early. Instead, he spent nearly a dozen years in what he would later refer to as his "dismal third-floor chamber," reading and trying to master the art of writing fiction.

Early work. In college Hawthorne had excelled only in composition and had determined to become a writer. Upon graduation, he had written a novel, *Fanshawe*, which he published at his own expense—only to decide that it was unworthy of him and to try to destroy all copies. The novel was an amateurish blend of the popular mode of supernatural melodrama and Hawthorne's college moods and experiences. Hawthorne, however, soon found his own voice, manner, and subjects; within five years of his graduation he had published such impressive and distinctively "Hawthornesque" stories as "The Hollow of the Three Hills" and "An Old Woman's Tale." By 1832, "My Kinsman, Major Molineux" and "Roger Malvin's Burial," two of his greatest tales—and among the finest in the language—had appeared. "Young Goodman Brown," perhaps the greatest tale of witchcraft ever written, appeared in 1835.

First short stories

Increasing success in placing his stories—first in newspapers, then in Christmas "gift-book" annuals, and finally in magazines—brought him slight fame (most of his early work was unsigned and pseudonymous) but hardly any money. Writing, he became convinced, was not a "career" in America in the 1830s. Unwilling to depend any longer on his uncles' generosity, he turned first to editorial hackwork in Boston and then to a job in the Boston Custom House, where by frugal living he managed to save some money. Finally, he became a shareholding member of the agricultural cooperative Brook Farm, in West Roxbury, Massachusetts, where he dreamed of earning a living by manual labour while still having enough time to write. Even when his college friend Horatio Bridge had financially guaranteed the publication of his first signed book, *Twice-Told Tales*, in 1837, the work had brought gratifying recognition but no dependable income. By 1842, however, Hawthorne's writing had brought him a sufficient income to allow him to marry Sophia Peabody, a Salem neighbour to whom he had been engaged for several years, as well as to realize his dream of being just a writer.

Life in Concord. The young couple rented the Old Manse in Concord and began a three-year period that Hawthorne would later look back on as the happiest in his life.

The presence of some of the leading social thinkers and philosophers of his day, such as Ralph Waldo Emerson, Henry Thoreau, Bronson Alcott, and Ellery Channing in Concord, made the village the centre of the new philosophy of Transcendentalism—a school of thought that encouraged man to transcend the materialistic world of experience and facts and become conscious of the pervading spirit of the universe and the potentialities for human freedom. But despite his new-found contentment, Hawthorne identified with the Puritans enough to write such anti-Transcendentalist sketches as "The Celestial Railroad" and "Earth's Holocaust" during the Concord years. When his Transcendentalist neighbours paid calls, he and Sophia were gratified but he had little to say to them. Artists and intellectuals never inspired his full confidence, but he thoroughly enjoyed the visit of his old college friend and classmate Franklin Pierce, later to become the 14th president of the United States. A happy man, Hawthorne thought, would have no questions to ask of Emerson, who seemed to him to dismiss natural evil and human guilt too easily. Though his writing of the Concord period was generally lighter in tone than his earlier sombre tales of Puritan history, Transcendentalism touched him only lightly.

Reaction to Transcendentalism

Mature novels. A growing family and mounting debts brought an end to his idyllic life in 1845, and the Hawthornes returned to Salem where Nathaniel became surveyor of the Custom House. Salem was declining as a port, and the job was essentially a political sinecure with minimal duties. Nevertheless, Hawthorne did no writing beyond book reviews during the more than three years of his tenure. When the next presidential election brought a new administration to power, he lost the job; but in a few months of concentrated effort he produced his masterpiece, *The Scarlet Letter*. The bitterness he felt over his dismissal is apparent in "The Custom House" essay prefixed to the novel. The work also reflects the grief he felt at his mother's death—in the unrelieved sombreness of the story of two lovers kept apart by the ironies of fate, their own mingled strengths and weaknesses, and the Puritan community's interpretation of moral law, until at last death unites them under a single headstone. The cemetery, the prison, and the rose compete for symbolic dominance, but the cemetery has the last word. Hawthorne decided the work was not "representative" of him and resolved to make his next book a happier one.

Sombre or not, the work sold far better than any of his books yet had, well enough to renew his hope of being able to live by writing alone. His dismissal from the Custom House job had become a partisan cause célèbre across the country, a fact that no doubt greatly helped the book's sale. Sales were further stimulated by the condemnations of the novel's sympathetic treatment of an adultress by certain reviewers in religious publications.

Bitter about his dismissal and determined to leave Salem

forever, Hawthorne decided that the mountain scenery of the Berkshires in western Massachusetts would be conducive both to concentrated writing and to that lighter mood he wished to achieve in his new romance. He settled in Lenox and began work on *The House of the Seven Gables*, the story of the Pyncheon family, who for generations had lived under a curse until it had been removed at last by love.

The 18 months he spent in Lenox were, at first, a happy and productive time for Hawthorne. Some five months of steady writing in the fall and winter of 1850–51 produced the new romance, and its favourable reception cheered and encouraged Hawthorne. His publisher now saw him as a "valuable property" and continually pressed him for new works, to which Hawthorne responded with two new editions of *Twice-Told Tales* and two books for children designed to make the past seem as vividly present to their readers as it did to Hawthorne.

Friendship with Melville

The friendship he formed in the Berkshires with the author Herman Melville, although important for the younger writer and his work, was much less so for Hawthorne. Melville praised Hawthorne extravagantly in a review of his *Mosses from an Old Manse*, comparing him with Shakespeare in depth of meaning. He also dedicated *Moby Dick* to Hawthorne and wrote some of his *Piazza Tales* in a manner that owed much to the symbolic writing he had first discovered in Hawthorne's tales. But eventually Melville came to feel that the friendship he so ardently pursued was one-sided. Later he was to picture the relationship with disillusion in his introductory sketch to *The Piazza Tales* and depicted Hawthorne himself unflatteringly as "Vine" in his long poem *Clarel*.

Hawthorne's final months in "the little red house" at Lenox found him increasingly discontented. Deciding that he preferred to be near the coast rather than in the mountains after all, he moved his family to another temporary residence, this time in West Newton, near Boston. There he quickly wrote *The Blithedale Romance*, based on his disenchantment with Brook Farm, as well as another book for children. Then he purchased and—urged on by his wife—extravagantly redecorated and remodelled Bronson Alcott's house in Concord, the Wayside, hoping apparently to recapture something of the idyllic contentment of the earlier Concord period. But *Blithedale* was disappointingly received and did not produce the income Hawthorne had expected. He was hoping for a lucrative political appointment that would bolster his finances; in the meantime, he began writing the kinds of books he could produce quickly and that would make money. Two more works for children and a campaign biography of his old friend Franklin Pierce resulted.

Consulship in Liverpool

When Pierce won the presidency Hawthorne was, in 1853, rewarded with the consulship in Liverpool, Lancashire, a position he hoped would enable him in a few years to leave his family financially secure. (He was now 48 and already anticipating an early death, as his first college-day protagonist Fanshawe had.)

Last years. The remaining 11 years of Hawthorne's life were, from a creative point of view, largely anticlimactic. Though the official duties as consul, which terminated in 1857, left him a great deal of free time, he found himself unable to write any new work of fiction, filling his notebooks instead with descriptions of places visited on sight-seeing trips. A year and a half of sightseeing in Italy similarly swelled the notebooks, but they were full of his anxiety over serious family illnesses and the tedium of museum visiting. Determined to produce yet another romance, he finally retreated to a seaside town in England and quickly produced *The Marble Faun*. In writing it, he drew heavily upon the experiences and impressions recorded in the Italian notebooks to give substance to an allegory of the Fall of man, a theme that had usually been assumed in his earlier works but now received direct and philosophic treatment.

Back in the Wayside once more in 1860, Hawthorne found American summers unbearably hot and the winters unpleasantly cold. Although he devoted himself entirely to his writing, he was unable to make any progress with his plans for a new novel. The drafts of unfinished works he left are mostly incoherent and show many signs of a psychic regression, already foreshadowed by his increasing restlessness and discontent of the preceding half dozen years. He managed to draw upon his English notebooks for a sensitive and moving book of English sketches, *Our Old Home*, to complete a magazine piece called "Chiefly About War Matters," and finally to begin *The Dolliver Romance*, which starts off in something very like the manner of the early tales. Some two years before his death he began to age very suddenly. His hair turned white, his handwriting changed, he suffered frequent nosebleeds, and he took to writing the figure "64" compulsively on scraps of paper. He died in his sleep on May 19, 1864, in Plymouth, New Hampshire, on a trip in search of health with his friend Pierce.

Assessment

Hawthorne's work initiated the most durable tradition in American fiction, that of the symbolic romance that assumes the universality of guilt and explores the complexities and ambiguities of man's choices. His work directly influenced that of Melville, Henry James, Flannery O'Connor, and Robert Penn Warren, among others, and analogues of his manner and his vision may be found in William Faulkner's works and that of many other writers. His greatest tales and *The Scarlet Letter* are marked by a depth of psychological and moral insight seldom equalled and never surpassed by any American writer.

MAJOR WORKS

NOVELS: *Fanshawe, a Tale* (1828); *The Scarlet Letter* (1850); *The House of the Seven Gables* (1851); *The Blithedale Romance* (1852); *The Marble Faun: Or, the Romance of Monte Beni* (British title, *Transformation*, 1860). (UNFINISHED NOVELS): *Septimius Felton* (1872); *The Dolliver Romance, and Other Pieces* (1876); *Doctor Grimshawe's Secret* (1883); *The Ancestral Footstep* (1883).

STORIES: *Twice-Told Tales*, including "The Gray Champion," "The Gentle Boy," "A Rill from the Town Pump," "The Great Carbuncle," "Sights from a Steeple," and "Dr. Heidegger's Experiment" (1837); 2nd enl. ed., including also "The Celestial Railroad" (1842); *Mosses from an Old Manse* (1846); *The Snow-Image, and Other Tales* (1851; also published as *The Snow-Image, and Other Twice-Told Tales*, 1852). (STORIES FOR CHILDREN): *Grandfather's Chair* (1841); *Famous Old People* (1841); *Liberty Tree* (1841); *Biographical Stories for Children* (1842); *A Wonder Book for Girls and Boys* (1851); *Tanglewood Tales for Girls and Boys* (1853).

BIOGRAPHY: *Life of Franklin Pierce* (1852).

AUTOBIOGRAPHICAL: *Our Old Home: A Series of English Sketches* (1863); *Passages from the American Note-Books of Nathaniel Hawthorne* (1868); *Passages from the English Note-Books of Nathaniel Hawthorne* (1870); *Passages from the French and Italian Note-Books of Nathaniel Hawthorne* (1871).

BIBLIOGRAPHY. NINA E. BROWNE, *A Bibliography of Nathaniel Hawthorne* (1905, reprinted 1967); BUFORD JONES, *A Checklist of Hawthorne Criticism, 1951–1966* (1967); *Hawthorne's Works*, "Riverside Edition," 12 vol. (1904); *The Centenary Edition of the Works of Nathaniel Hawthorne* (in progress); N.H. PEARSON (ed.), *The Complete Novels and Selected Tales of Nathaniel Hawthorne* (1937); HENRY JAMES, *Hawthorne* (1879), the earliest critical study, still valuable; RANDALL STEWART, *Nathaniel Hawthorne: A Biography* (1948), still definitive, though it lacks insight into Hawthorne's inner life; HYATT H. WAGGONER, *Hawthorne: A Critical Study*, rev. ed. (1963); R.H. PEARCE (ed.), *Hawthorne Centenary Essays* (1964); FREDERICK C. CREWS, *The Sins of the Fathers: Hawthorne's Psychological Themes* (1966); B. BERNARD COHEN (ed.), *The Recognition of Nathaniel Hawthorne* (1969), a collection of representative critical responses from the earliest to the present.

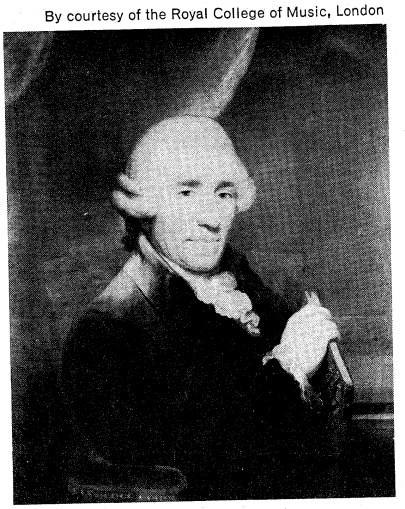
(H.H.W.)

Haydn, Joseph

Closely linked to the flowering of classical music in his 18th-century Austria, Franz Joseph Haydn, after experimenting with various stylistic trends prevailing in his youth (the pompous and complex idiom taken over from the preceding Baroque period; the light and gay *style galant*, a style distinguished by formal elegance and clarity imported from Italy and France; and the strongly emotional and expressive *Empfindsamkeit*, or "sensitive style," preferred by North German composers), eventually achieved his own distinctive musical identity by us-

ing some elements of all three styles simultaneously. From the age of nearly 50 onward, he wrote his greatest works, which display a classical synthesis of seemingly incompatible characteristics. Inspiration and heartfelt expression were as important to him as the work of an alert intellect. He established a well-balanced mixture of cheerful folk song elements and serious, at times even tragic, moods. Haydn was significantly helped by the younger composer Wolfgang Amadeus Mozart, who at first accepted the older man's guidance and later enthusiastically collaborated with him.

By courtesy of the Royal College of Music, London



Haydn, portrait by Thomas Hardy, 1791. In the collection of the Royal College of Music, London.

Haydn is often called the father of the symphony and of the string quartet. Although the statement is in the nature of an oversimplification, works of this kind having existed before him, he did endow the two forms with an artistry and significant content that secured for them a high place within musical literature. His last two oratorios, his great symphonic masses, as well as some of his piano works, songs, and concerti stand on an equally high level. Beethoven was strongly indebted to Haydn, and in the 19th century such composers as Franz Schubert, Johannes Brahms, and Anton Bruckner all were to build on foundations erected by him.

Early years. Haydn was born on March 31, 1732, in Rohrau, a village in eastern Austria near the Hungarian border. War was anything but a rarity in the district, and Haydn's ancestors had suffered from invasions by the Turks and by Hungarian raiders. Stubborn tenacity and resourcefulness were needed to survive under such conditions, and these were qualities inherited by Haydn. His parents, like their forefathers, were humble people; his father a wheelwright, his mother, before her marriage, a cook for the lords of the village. Joseph, their second child, revealed, to everybody's surprise, unusual musical gifts, and the parents were at a loss how to provide the proper type of education for him. The problem was solved when a cousin, serving as a school principal and choirmaster in the nearby city of Hainburg, offered to take the boy into his home and train him. The parents, seeing no other way, agreed; and thus Joseph, not yet six years old, left home, never to return to the parental cottage except for rare, brief visits.

Education at Hainburg and Vienna

From a professional point of view, the move to Hainburg was an important one. The boy sang in the church choir, learned to play various instruments, and obtained a good basic knowledge of music. In other respects, life at Hainburg was less satisfactory. The cousin was poor and his modest salary hardly sufficed to support a growing family. Joseph was not given the love and care a child needs; as he later reported, he received "more floggings than food." But he was able to take such disadvantages in his stride; his nature was endowed with wiry resilience and a contented disposition.

A decisive change in his life occurred when Haydn was eight years old. The musical director of Saint Stephen's Cathedral in Vienna had observed the boy on a visit to Hainburg and invited him to serve as chorister at the Austrian capital's most important church. Haydn's parents accepted the offer with delight, for it secured for their son a thorough musical training and relieved them of all financial cares while he boarded at the choir school. Thus, in 1740 the eight-year-old moved to Vienna full of expectation of the glorious experiences in store for him. He stayed at the school for nine years, acquiring an enormous practical knowledge of music by constant performances but, to his disappointment, receiving far too little instruction in music theory. The pattern of life in Hainburg seemed to repeat itself. Again he had to work hard to fulfill all his obligations as a chorister; again he was badly fed and sometimes suffered, as he reported, from "ravenous hunger." Gradually his voice deteriorated; when it broke, the cathedral choir had no use for him. Seizing on the pretext of punishing a practical joke committed by the high-spirited youth, the director had him expelled from the choir school.

With no money in his pocket and three ragged shirts and an old coat as his only possessions, Haydn at 17 was left to his own devices. He found a refuge for a while in the humble garret of a fellow musician and supported himself "miserably" with odd musical jobs, performing at dances and serenades, playing the organ at Sunday services, and teaching at very modest fees. Hand in hand with these activities went an arduous course of self-instruction through the study of musical works—notably those of Carl Philipp Emanuel Bach (1714–88)—and of leading manuals of musical theory. A fortunate chance brought him to the attention of the successful Italian composer and singing teacher Niccolò Porpora (1686–1768), who accepted him as accompanist for voice lessons and, in return for this as well as for his services as a valet, corrected Haydn's compositions.

Thanks to tremendous persistence and unflagging energy, Haydn made progress. He was eventually engaged to teach some aristocratic pupils and introduced by them to the music-loving Austrian nobleman Karl Joseph von Fürnberg, in whose home he played chamber music. For the instrumentalists there he wrote his first string quartets, a form that he cultivated throughout his professional career, composing some 80 works in this genre.

In 1758 Haydn received his first steady appointment. Through the recommendation of Fürnberg he was engaged as musical director and chamber composer by the Bohemian count Ferdinand Maximilian von Morzin, who resided most of the time in his country estate of Lukaveč in western Bohemia. Haydn had an orchestra of about 16 musicians at his disposal, and, for this ensemble, he wrote his first symphony, the other musical form in which he pioneered. These first attempts in the field of instrumental composition were still conventional in character, yet a certain freshness of melodic invention and sparkle marked them as the work of a future master.

Bohemian post with Count von Morzin

The first period of Esterházy patronage. Haydn did not stay long with Count von Morzin, as financial difficulties forced his patron to dismiss the orchestra. Before long he was invited to enter the service of Prince Pál Antal Esterházy, who had heard his works at Count von Morzin's castle. The Esterházys were one of the wealthiest and most influential families of the Austrian empire and boasted a distinguished record of supporting music and art. Prince Pál Antal had a well-appointed orchestra performing regularly in his sumptuous castle at Eisenstadt, a small town some 30 miles from Vienna. As his aged music director was ailing, the prince had vision enough to appoint the relatively unknown Haydn as assistant conductor. The contract concluded on May 1, 1761, informs us about the new appointee's duties. While church music was still entrusted to the director, three other spheres of activity were assigned to Haydn: to conduct the orchestra and coach the singers (which meant almost daily rehearsals); to compose the greatest part of the music required; to do administrative work by serving as music librarian, supervisor of instruments,

and chief of the musical personnel. Haydn carried out these duties extremely well and even managed to find time to supervise the work of the copyists of music and to tune his own keyboard instruments. His service as chief of personnel revealed tact, good nature, and skill in dealing with people. He exerted himself to support his subordinates against other officials on the Prince's staff, and, thanks to his friendliness and his sense of humour, he managed to maintain good relations with staff and employers. The musicians loved and respected their "Papa," as they nicknamed him.

"Papa" Haydn

In 1766 Haydn became musical director at the Esterházy court. He raised the quality and increased the size of the Prince's musical ensembles by appointing many choice instrumentalists and singers. His ambitious plans were supported by Prince Miklós, who, on the death of his brother in 1762, had become head of the family and maintained his leadership for 28 years. A passionate and discriminating music lover himself, he was able to appreciate Haydn's contributions and created an atmosphere beneficial for the development and maturing of his music director's art. Miklós "The Magnificent," as he was usually referred to, loved splendour and display. Having admired Versailles in France, he decided to match it by a creation of his own. At enormous expense he built in 1766 the splendid castle of Esterháza, situated in the western part of Hungary. Every night a performance of a German play or of an Italian opera would be offered in the palace theatre. Haydn, in addition to his other staggering duties, was in charge of all operatic activities. He not only selected, rehearsed, and conducted operas by other composers but contributed more than a dozen musical works of his own. No lesser person than Empress Maria Theresa of Austria testified to the high level of operatic performance achieved in the Esterházy theatre. She was once overheard saying, "When I want to hear a good opera, I have to go to Esterháza," a verdict that was not gratifying to the musicians of Vienna.

Operas

Some of Haydn's operas, such as the hilarious *La canterina* (*The Songstress*) of 1767 and *Lo speziale* (*The Apothecary*) of 1768, are musical comedies; other stage works, such as *L'isola disabitata* (*The Deserted Island*) of 1779 and *Armida* of 1784, are of a serious character. Most significant are those operas in which Haydn succeeded in combining comic and serious elements by conjuring up gay characters contrasted with more earnest-minded ones. One of his last and most successful works for the stage, *Orlando paladino* (*Knight Roland*) of 1782, bore the fitting subtitle of "drama eroicomico" (heroic–comic drama). Even when performing operas by other composers, Haydn made adaptations to render them suitable for the personnel of the Prince's theatre. Following the common practice of the time, he often added arias of his own to works by other composers, which were display pieces for the singers and added additional sparkle to the work performed.

In addition to his operatic duties, he composed symphonies, string quartets, and other chamber music. The Prince was a passionate performer on the baryton, and Haydn provided for his patron more than 150 compositions featuring this now obsolete cello-like instrument.

Haydn seems to have enjoyed his busy life and served Prince Miklós for nearly 30 years. Living in the country was no hardship for him. He loved hunting, fishing, and other outdoor activities. Nor did he feel cut off from the cultural world, for travelling dramatic troupes often brought works by eminent writers to Esterháza, and he frequently visited Vienna in the Prince's retinue.

Friendship with Mozart

On these occasional visits a close friendship developed with Mozart. The fact that Mozart was 24 years younger than Haydn did not matter. They felt inspired by each other's work. Mozart declared that he had learned from Haydn how to write quartets and dedicated a superb set of six such works to his "beloved friend." Haydn's music, too, shows in various details the impact of his young friend's idiom. The mature composer was by no means set in his ways; he was flexible and receptive to new ideas, an admirable trait in an artist who had already won wide recognition.

Unlike Mozart, Haydn became internationally famous in his own lifetime. His works were performed throughout Europe and were published in Austria, Germany, Holland, France, and England. He received official commissions from many European music lovers. The city of Cádiz in Spain, for example, commissioned *The Seven Last Words* for a Good Friday service; the king of Naples requested compositions for the lira organizzata, a hurdy-gurdy (wheel fiddle) with a built-in tiny pipe organ; and the *Paris Symphonies* (numbers 82–87) were composed between 1785 and 1786 for the French capital.

The success of Haydn's professional career was not matched in his personal life. The girl he loved entered a convent, and her parents induced Haydn, who had just obtained his position with Count von Morzin, to espouse the elder sister. The decision proved disastrous for his realization of a pleasant, peaceful home. No children were born to the couple. Haydn's wife, two years his senior, was quarrelsome and bigoted. She did not understand music and showed no interest in her husband's work. It was reported that her expression of disdain went to the extremes of using his manuscripts for pastry linings or curl papers. As a result the composer was not insensitive to the attractions of other women, and for years carried on a love affair with Luigia Polzelli, a young Italian mezzo-soprano in the Prince's service.

Unhappy marriage

English period. In 1790 Prince Miklós died. His son, Prince Antal, did not care for music and dismissed most of the court musicians. Haydn was retained, however, and continued to receive his salary. No duties were required of him, enabling Haydn to do whatever he pleased. But after such a long time at the Esterházy court, the composer was eager to try a different way of life. He, therefore, considered seriously two tempting offers that were made as soon as his availability was generally known. First, the king of Naples urged him to accept an invitation extended several years earlier to visit his court; second, a violinist and concert manager, Johann Peter Salomon, arrived from England and commissioned, at excellent financial terms, six new symphonies and 20 smaller compositions to be conducted by the composer himself in a series of orchestral concerts in London sponsored by Salomon. Haydn, moreover, was to compose a new Italian opera for the King's Theatre in London, for a fee of £300. Haydn felt far more attracted by the English than by the Italian invitation. He felt it would be a great experience to work with a large, excellently trained orchestra and to live in one of the great musical centres of the 18th century. Furthermore, he would no longer have to submit to the rigid etiquette of the Esterházy court and continually be reminded of his subservient position.

First visit to London

Undeterred by the dire warnings of his friends, including Mozart, he left for London. What the prospect of the visit to England meant to him may be realized when one considers that though 58 he had yet to travel beyond the borders of Austria. On New Year's Day, 1791, he arrived on British soil, and the experience of the following 18 months proved as rewarding and stimulating as he had anticipated. The many novel impressions, the meeting with eminent musicians, the admiration bestowed on him had a powerful impact on his creative work. The symphonies written for the first and second visit to London represent the climax in his orchestral output. One admires their virtuosity of instrumentation, the masterly treatment of musical forms, the freely flowing melodic inspiration, and the sense of humour that endeared the works so much to the British audiences. Their popularity is reflected in the various nicknames bestowed on them (*Surprise*, *The Clock*, *Military*, *Drum Roll*). As Haydn conducted the works himself, he exercised, in the words of the great contemporaneous British musicologist Charles Burney, "an electrical effect on all present and such a degree of enthusiasm as almost amounted to frenzy." The opera he composed fared less well because the King's Theatre, for which it was commissioned, was refused the license to perform such works. Haydn, however, suffered no financial loss, and the event seems not to have greatly disturbed him.

In England Haydn was feted by famous men as well as by royalty. In July 1791 Oxford University awarded him the honorary degree of Doctor of Music. He performed at the residences of the Prince of Wales and the Duke of York. His diary describes visits at which he was "gloriously entertained," and references to beautiful women he met at such occasions are anything but rare. Sixty years old, his face pitted by smallpox, with a large aquiline nose, and too-short legs, he remained attractive to women admirers.

In spite of his great success, Haydn was forced to submit to one of the artistic contests fashionable at the time. A rival music organization, which had failed to induce Haydn to join them, engaged a young pupil of his, one Ignaz Pleyel, to conduct its concert series. A vicious campaign was started emphasizing Haydn's advanced age and the pupil's expertise, but it was to no avail; Haydn could report that he had "kept the upper hand." The concerts he conducted, consisting largely of his own works, continued to be a series of triumphs.

In June 1792 Haydn left London for Germany, as Prince Esterházy wanted the famous composer in his retinue at the coronation of the emperor Francis II in Frankfurt am Main. On his journey he stopped at Bonn where the 22-year-old Beethoven was introduced to him, and it was arranged that the young composer should move to Vienna to receive the aged master's instruction. On July 29 Haydn arrived in Vienna where he was able, thanks to his considerable earnings in England, to buy a pleasant house in the suburb of Gumpendorf. This dwelling still stands today and is the site of a Haydn museum.

Second visit to London

The composer did not stay in Vienna for long, however. His English admirers urged him to visit London once more, and in January 1794 he left the Austrian capital to return to England until August 15, 1795. Again he won great acclaim. Various members of the royal family made efforts to keep him in England, but Haydn declined, because a change had occurred in the leadership of the Esterházy family. Miklós II, the new head of the princely house, was eager to restore the former orchestra under Haydn's direction, and the composer felt compelled to accept the assignment. The duties he now undertook would, of course, not be as burdensome as previously. Moreover, the Prince's intention to spend the winter months in Vienna provided a strong incentive. Although Haydn had to separate from the many friends he had made in London, he could look back on the period spent in England as a profitable one. His trunks bulged with no less than 768 pages of music he had written in that hospitable country, and the musical stimulus and the financial rewards for his labour had been gratifying.

The late Esterházy and Viennese period. In 1791 Haydn had attended the Handel Commemoration at Westminster Abbey in London. He was deeply moved by the superb rendition of Handel's masterly oratorios and by the veneration with which they were received by the English audience. Deciding to compose in this musical genre, he was able to obtain a suitable libretto, allegedly prepared for Handel himself. After settling in Vienna and resuming his duties for Prince Esterházy, Haydn started work on the oratorio *The Creation*, the text of which had been translated into German by Baron Gottfried van Swieten. The libretto was based on the epic poem *Paradise Lost* by England's foremost 17th-century poet, John Milton (1608–74) and on the Genesis chapter of the Bible. Composing the oratorio provided Haydn with a truly congenial task. He was a deeply religious man who throughout his long career had contributed Latin masses of great beauty for the Roman Catholic Church. In the oratorio he could express his gratitude to God in a work with German text and at the same time depict in music the beauties of nature that so greatly delighted him. The years devoted to this task were among the happiest in Haydn's life. He felt uplifted and in communion with the divine spirit. In April 1798 the oratorio had its first performance in a princely palace and produced a profound effect. As a critic wrote: "Three days have gone since that enrapturing evening . . . still the mere memory of all the flood of emotions then experienced constricts my heart." Before long a public performance took place with equal results, and, henceforth, *The Creation* was again and again performed with great success, the proceeds going, at the composer's request, to charitable institutions.

Oratorios

This encouraged Haydn to produce another oratorio, which absorbed him until 1801. A British text, the poem *The Seasons*, by the Scottish poet James Thomson (1700–48), was again chosen for the libretto and translated by van Swieten. Though of limited poetical value, the adapted work allowed him to compose delightful musical genre pictures of events in nature. This second oratorio was also triumphantly successful, both at court, where the Austrian empress sang the soprano solos, and in public performances.

Haydn's late creative activity, however, was not confined to the oratorio. The six masses written for his patron Esterházy are among the most significant works of this kind from the 18th century. He continued to compose magnificent string quartets and, in 1797, he gave to the Austrian nation the stirring song "Gott erhalte Franz den Kaiser" ("God Save Emperor Francis"). It was used for more than a century as the national anthem of the Austrian monarchy and as the patriotic song "Deutschland, Deutschland über alles" ("Germany, Germany above all else") in Germany. It is heard today under many titles as a Protestant hymn in the English-speaking countries. The song was so beloved that Haydn decided to use it as a theme for variations in one of his finest string quartets, the *Emperor Quartet*, Opus 76, No. 3.

Honours bestowed

Not surprisingly, honours from various parts of Europe gladdened the composer in his last years. Stockholm, Amsterdam, St. Petersburg, and Paris made him an honorary member of their music associations and the French capital had, after the premiere of *The Creation*, a gold medal engraved in Haydn's honour. Nor did Austria lag behind. In the village of Rohrau, Haydn's birthplace, a monument was erected to its famous son, and Haydn had the gratification of seeing it. The city of Vienna conferred on him the great golden Salvator Medal and named him an honorary citizen. Particularly notable was the Vienna concert of 1808 in celebration of Haydn's 76th birthday. *The Creation* was performed by eminent musicians in the presence of the ailing composer, carried in on an armchair and seated amid jubilant exclamations among members of the high nobility. Poems were read in his honour, applause shook the hall, and on Haydn's departure Beethoven knelt down and kissed the hands of his former teacher. This was Haydn's last public appearance. Henceforth, conditions in Vienna prevented such outings and undermined the little strength still left to the aged artist. Austria faced the Napoleonic armies in 1809—devastating battles were fought and Vienna was bombarded, a cannon ball falling near Haydn's residence. Napoleon, however, eventually placed a guard of honour outside Haydn's house.

In these turbulent days a moment of joyful respite was granted to the invalid. A French officer called and, professing admiration of Haydn's music, gave a rendering of an aria from *The Creation*, which allegedly made the composer shed happy tears. Soon afterward the singer fell in battle. Haydn survived him by only a few days, dying on May 31, 1809. At the official obsequies, members of the French army joined forces with the municipal militia to form a line before the catafalque. Vienna's cultural elite and high-ranking French officers paid solemn tribute to the genius who had given so much to the musical world.

Reasons for contemporary success

The success achieved by Haydn in his lifetime is often attributed to his slow development, enabling music lovers to progress step by step in their understanding of his aims; his long life endowed with creative power up to old age; and, most of all, his attunement to the spirit of the times. Haydn was a true representative of the age of Enlightenment. His positive, optimistic approach to life, his striving for balance between the work of the intellect and the flow of emotions; his sense of moderation, lead-

ing to the avoidance of strongly discordant moods, found superb expression in his music and thus gave his contemporaries the kind of music they were able to appreciate. Music lovers also found irresistible the nobility and deceptive simplicity of his idiom, sparked by delightful outbreaks of humour. In the 19th century the ideals of the Enlightenment no longer were prevailing. The age of Romanticism liked to probe dark and complex moods; morbidity seemed attractive, and emphasis was laid on conjuring up ambivalent emotions in music. To this epoch the gaiety and naturalness of Haydn's idiom seemed to be rather philistine and people patronized "good old papa Haydn," whose works were, as a matter of fact, hardly known. In the 20th century there was a re-evaluation of his work; and the outstanding intellectual nature of his thematic elaborations, the originality of his modulations, and the artistry and superb craftsmanship of his orchestration were again appreciated.

MAJOR WORKS

Instrumental music

SYMPHONIES: 106 symphonies, including *No. 6 in D Major* (*Le Matin;* completed, 1761); *No. 7 in C Major* (*Le Midi;* 1761); *No. 8 in G Major* (*Le Soir;* 1761); *No. 45 in F Sharp Minor* (*Farewell;* 1772); *No. 49 in F Minor* (*La Passione;* 1768); *No. 83 in G Minor* (*La Poule;* 1785); *No. 84 in E Flat Major* (1786); *No. 88 in G Major* (1787); *No. 92 in G Major* (*Oxford;* 1789); *No. 94 in G Major* (*Surprise;* 1791); *No. 96 in D Major* (*Miracle;* 1791); *No. 100 in G Major* (*Military;* 1794); *No. 101 in D Major* (*The Clock;* 1794); *No. 102 in B Flat Major* (1795); *No. 103 in E Flat Major* (*Drum Roll;* 1795); and *No. 104 in D Major* (*London;* 1795).

CONCERTI: 3 concerti for violin; 2 concerti for cello; 11 keyboard concerti; *Symphonie concertante* for violin, oboe, cello, bassoon, op. 84 (1792); *Horn Concerto* (1762); 5 concerti for two lire organizzate (hurdy-gurdies with small built-in organ pipes; 1786); *Trumpet Concerto* (1796).

OTHER ORCHESTRAL WORKS: Various overtures, divertimenti, marches, and dances.

STRINGS: About 79 quartets, including some 20 early works; 6 in op. 17 (1771); 6 in op. 20 (1772); 6 in op. 33 (1777–81); op. 42 in D minor (1785); 6 in op. 50, dedicated to Frederick William II of Prussia (1787); 7 in op. 51 *Die Sieben Worte des Erlösers am Kreuze* (*Seven Words of Our Saviour on the Cross,* originally an orchestral work, string quartet version 1787, vocal version 1796); 6 in op. 54 and 55 (1788?); 6 in op. 64 (1790); 6 in op. 71 and 74 (1793); 6 in op. 76, dedicated to Count Joseph Erdödy (1797); 2 in op. 77, dedicated to Prince Lobkowitz (1799); op. 103 (1803); various string trios, duos, and divertimenti for 3 instruments.

OTHER COMPOSITIONS FOR STRINGS: 126 trios for baryton (cello-like instrument with additional sympathetically-vibrating metal strings under the fingerboard), viola, and cello; 41 piano trios.

PIANO: 54 sonatas; *Variations in F Minor* (1793); various separate movements.

Theatre music

OPERAS: About 16, including *La canterina* (1767); *Lo speziale* (1768); *Le pescatrici* (1770); *L'infedeltà delusa* (1773); *L'incontro improvviso* (1775); *Il mondo della luna* (1777); *L'isola disabitata* (1779); *La fedeltà premiata* (1780); *Orlando paladino* (1782); *Armida* (1784); and *L'anima del filosofo* (composed 1791).

Other vocal music

ORATORIOS: *Die Sieben Worte* (*The Seven Last Words;* 1796); *Die Schöpfung* (*The Creation;* 1798); *Die Jahreszeiten* (*The Seasons,* 1801).

MASSES: 13 masses, including *Missa Celensis* (1782); *Missa in tempore belli* in C major (*Paukenmesse,* 1796); *Missa St. Bernardi de Offida* in B flat major (*Heiligmesse,* 1796); *Missa in angustiis* in D minor (*Nelson Mass,* 1798); *Mass in B Flat Major* (*Theriesienmesse,* 1799); *Mass in B Flat Major* (*Schöpfungmesse,* 1801); *Mass in B Flat Major* (*Harmoniemesse,* 1802).

OTHER CHURCH WORKS: 2 Te Deums (both in C major; one from the 1760s, one *c.* 1800); *Stabat Mater* (1767); *Ave Regina* in A major (*c.* 1756); 3 Salve Reginas (E major, 1756; E flat major, *c.* 1770; G minor, 1771); various offertories, motets, and arias.

SONGS: 12 canzonettas (in English, 1794–95); about 34 German songs; 13 partsongs; various arias for other composers' operas.

CANONS: 10 sacred (*The Holy Ten Commandments*) and 46 secular.

BIBLIOGRAPHY. ANTHONY VAN HOBOKEN, *Joseph Haydn: Thematisch-Bibliographisches Werkverzeichnis,* vol. 1 (1957), lists the instrumental compositions; the second volume dealing with the vocal works is in preparation. The largest number of Haydn's autographed scores is owned by the National Library Széchényi in Budapest, Hungary. H.C. ROBBINS LANDON, *The Collected Correspondence and London Notebooks of Joseph Haydn* (1959), is a sensitive translation of all Haydn letters known. Biographical and critical studies include: VERNON GOTWALS (trans.), *Joseph Haydn: Eighteenth-Century Gentleman and Genius* (1963), a translation of the two earliest biographical sketches of the life of Haydn, both based on information received from the composer himself; KARL and IRENE GEIRINGER, *Haydn: A Creative Life in Music,* 2nd rev. ed. (1968), a biography first published in a German version in 1932 now expanded and updated; ROSEMARY HUGHES, *Haydn* (1950), a compact biography of the composer; and H.C. ROBBINS LANDON, *The Symphonies of Joseph Haydn* (1955), a basic investigation of the composer's symphonic output.

(K.G.)

Headache

Headache in one form or another probably is man's most frequently troublesome symptom, affecting nearly everyone at some time in his life and inflicting recurrent discomfort on perhaps one person out of ten. Headaches vary widely in their intensity and in the seriousness of the underlying conditions that cause them, and there is no necessary correlation between the severity of the pain and its cause. Pain is an individual matter—a biological danger signal initiated by tissue injury, carried along fairly specific nervous system pathways, but ultimately experienced according to the psychological endowment and past experience of the sufferer. The understanding of headache, therefore, requires an awareness of the mechanisms that cause head pain and of the fact that different genetic and emotional backgrounds produce very different individual responses to tissue injury of similar intensity.

Mechanism of headache. Most headaches occur because specific pain-sensitive structures in or around the head are overstimulated or damaged. Some of these are inside the skull, or intracranial; the remainder are in the tissues surrounding or covering the skull, or extracranial. Extracranial structures are by far the more common sources of recurrent headache.

Intracranial headache

Intracranial headache results when the arterial blood vessels at the base of the brain become excessively distended (as during fever, "hangover," or a severe and sudden attack of high blood pressure), or when an inflammation or hemorrhage surrounds these same arteries and their adjacent meningeal tissues (as during an attack of meningitis or a cerebral hemorrhage). Intracranial headache also can occur if a tumour or some other space-taking mass displaces certain of the large veins or other tissues inside the skull (the brain itself is not sensitive to pain). Intracranial headaches are usually of recent origin, particularly if they are severe, and often begin abruptly; only rarely do they last throughout the waking hours, and they usually change in character with a change in posture. Most headaches from brain tumour are felt in a particular part of the head, especially when they first begin. Intracranial causes of headache almost always produce associated abnormalities that the physician can detect by physical examination or laboratory tests.

Extracranial headache

Extracranial headaches have six major sources:

1. Painful dilatation and distention of the extracranial arteries that supply the surface tissues of the head. This category includes migraine and many "tension" headaches.

2. Sustained contraction of skeletal muscles about the face, scalp, and neck. This includes most psychologically associated "tension" headaches as well as those associated with excess fatigue, orthopedic difficulties of the neck, and simple eyestrain.

3. Distention, inflammation, or destruction of pain-sensitive tissues in the nose, paranasal sinuses, eyes, ears, and teeth. Headache from these sources is accompanied by objective signs of active disease.

4. Specific neuritis or neuralgias such as trigeminal neuralgia, all of which have very specific and predictable pain patterns.

5. Direct injuries, infections, or growths involving the extracranial tissues. Head pain following trauma is often a complex matter, since it can herald troubles as serious as a blood clot on the brain or as medically trivial as a desire for greater financial compensation after injury.

6. Rare inflammatory diseases of the cranial arteries.

More than 90 percent of all headaches result from distention of extracranial arteries or from sustained contraction of face and neck muscles, and accompany feeling states of frustration, resentment, anxiety, fatigue, or depression. Thus the individual's life situation and emotions are particularly important in the genesis of headache, both because disturbed feeling states initiate changes in arteries and muscles that cause pain, and because the presence of pain is often particularly threatening to emotionally troubled persons. Usually psychological factors important in the genesis of headache actually induce hurtful tissue changes. Only rarely does headache develop as a delusion for which no responsible tissue abnormality exists.

Migraine. Migraine is a paroxysmal disorder in which vascular headache is combined with several other bodily disturbances, including periodic changes in water balance and in autonomic nervous system function. Attacks may occur daily or as infrequently as once a year. Characteristically, patients with migraine have initial symptoms that precede headache by as much as an hour and are believed to be due to constriction of selective intracranial arteries: there may be illusions of flashing lights, loss of vision, or transient defects in speech or body movement. As a rule, the early symptoms subside in 10 to 30 minutes and are followed by severe, throbbing headache located over one or the other temple or occipital region. Less commonly the headache is bilateral or generalized. During the attack, bright lights hurt the eyes and many patients have nausea and vomiting. Dilation of one or more extracranial arteries causes this early painful stage. If the headache persists for more than a few hours, secondary painful contraction sets in in scalp and neck muscles, compounding the early symptoms. In typical cases, the attack lasts 12 to 24 hours or more unless relieved by vasoconstrictor and analgesic drugs early in its course.

Initial symptoms of migraine

The underlying cause of migraine is unknown. The illness occurs in families with an incidence suggesting an autosomal dominant inheritance (*i.e.*, the defect is transmitted by a nonsex chromosome; only one of the two genes involved in the trait need be transmitted). At one time or another, various investigators have attempted to link migraine to allergy, epilepsy, or other conditions, but the evidence has been unconvincing. Recently, several studies have suggested an abnormality in the regulation of serotonin, a chemical involved in transmitting impulses in the brain. The observed chemical changes in patients with migraine, however, have not always been consistent, and the evidence has not yet led to any change in treatment, which consists of using vasoconstrictor and analgesic drugs during the attacks and psychological guidance for the patients between attacks. Persons with migraine are characteristically perfectionistic, self-demanding, hard-working, and inflexible. They commonly set impossibly high standards for themselves and, less often, for others. When affected persons come to understand these feelings and realize the fatigue and resentment they engender in themselves, they are often able to establish more realistic ways of approaching life. Headaches tend to decline in frequency at the same time.

Cause of migraine

Medical significance of headache. The medical significance of headache varies widely as the causes themselves range from relatively minor emotional difficulties to illnesses as serious as brain tumour. Any headache that is recent in onset, different from previous patterns, or altered in quality requires close and careful medical evaluation.

BIBLIOGRAPHY. H.G. WOLFF, *Headache and Other Head Pain*, 2nd ed. (1963), an outstanding reference work covering all aspects of diagnosis, treatment, and experimental work; J.W. LANCE, M. ANTHONY, and H. HINTERBERGER, "The Control of Cranial Arteries by Humoral Mechanisms and its Relation to the Migraine Syndrome," *Headache*, 7:93–102 (1967), an experimental paper providing evidence that links the chemical serotonin to the genesis of migraine headache; F. SICUTERI, "Vasoneuroactive Substances in Migraine," *Headache*, 6:109–26 (1966), a review of pharmacological factors that affect blood vessels in migraine, and a presentation of the hypothesis that vascular headache is related to changes in the body's indole-peptide metabolism.

(F.P.)

Healing Cults

William James held that the basic element in religion is the cry "Help! Help!" This cry might be addressed to a god or to a religious specialist; it might arise out of a physiological disorder or a sense of sin and transgression. Many would hold that the most important function of religion has been that of healing—the diagnosis of the cause of evil and mental and physical sickness, and the development of techniques for its cure. From the earliest Neolithic evidence for trepanning (operating on the skull) to the techniques of the contemporary "new" religions of Africa, Melanesia, and Japan, from the shaman's hut to the big city revival hall, healing and religion have been inextricably interrelated. Rarely can a religious leader succeed unless he can heal; no religion has survived that does not heal.

Every religious tradition engages in healing and restorative activities, especially in the use of blessings, exorcisms, and purificatory rites. In almost every religious tradition there is some form of blessing; *i.e.*, a ritually repeated formula, often accompanied by gestures, that invokes the supernatural powers to protect the devotee and to grant him health, safety, and prosperity. As the verbal expression of a wish that good may befall someone and that evil may be averted, the blessing appears in a variety of religious contexts: therapeutic (designed to heal), apotropaic (designed to avert evil), and, especially, contexts in which hope and protection are expressed. Objects (*e.g.*, charms and amulets) may also be blessed and perform similar protective functions for their possessor.

Blessings, exorcisms, and purificatory rites

Many religious traditions associate illness with evil spirits that take possession of either a person or a place. They must be expelled by a series of complex incantations and actions most usually performed by a specialist.

Illness is frequently viewed as the result of ritual or moral transgression. To set things right, rituals of purification are required to remove the taint of the offense, procedures for confession alleviate the guilt of the offense, and acts of expiation or repentance restore the proper relationship with the offended power. Most religious traditions also preserve stories of gods, heroes, or holy men who were particularly adept at healing; and many religious traditions maintain hospitals and utilize various procedures for the care or visitation of the sick.

Although every religious tradition and functionary is concerned with healing (either in the physiological sense of curing disease or in the extended sense of overcoming transgression) and with restoring the world to its presumptive proper state, some movements have been so dominated by these concerns that they have come to be called healing cults. These are usually local and limited in membership; and they either become a schismatic group, a "new" religion, or are absorbed within a larger tradition.

TYPES OF HEALING CULTS

Shrine cults. Pilgrimage to a sacred place and devotion before a sacred object is a major means of religious healing. From earliest times, healing and healing cults have been associated with springs and other sources of water. An ancient Indian Vedic source says: "The waters are indeed healers; the waters drive away and cure all illnesses" (*Atharvaveda*, 6.91.3). Water—as the source of life in many myths, as that which is an absolute necessity for existence, and as that which cleanses—is the most

Water and other shrines

all-encompassing means of restoring health. As in the spa-therapy (bathing in mineral waters) of contemporary health resorts, so thermal and mineral springs were conceived to be curative in ancient times. There is evidence of Neolithic and Bronze Age devotion at the sites of a variety of such springs in western Europe (*e.g.*, Grisy and Saint-Sauveur in France; Forlì, Italy; Saint Moritz, Switzerland). Every country in which they occur has healing traditions associated with such springs. In ancient Greece the most famous shrines were at Thermopylae and near Aedepses; they were sacred to Heracles, the legend being that they were created for his refreshment after his labours. In ancient Rome, the springs at Tibus and the hot sulfur wells of Aquae Abulae were well known. In the Middle East, Callirrhoe, where Herod attempted to find relief from his fatal illness (Josephus, *Antiq.* XVII.6.5), was perhaps the best known; in ancient Egypt many of the temples dedicated to Asclepius (the Greek god of medicine) are adjacent to mineral springs; Muslim traditions identify several hundred springs as having healing powers. There is relatively little cultic activity at such springs; one simply goes to them and bathes when ill.

More elaborate cultic practices surround those sources of water that have been the scenes of epiphanies (manifestations of deities or sacred beings), or in which divinities are believed to dwell. The power of the divine presence or blessing, it is believed, charges the waters with healing abilities. The most famous Western example of this type of shrine is that at Lourdes in France where the Virgin Mary is believed to have appeared to one Bernadette Soubirous in a series of visions in 1858 and to have indicated a miraculously flowing stream that would heal the ill. A number of other European water shrines are associated with epiphanies of Mary (*e.g.*, the Shrine of the Madonna of the Baths at Scafati, Italy). Because of his association with water, many streams and wells are believed to have healing powers on the feast of the Conception of St. John the Baptist. More frequently, however, it is minor local water spirits (nymphs, water serpents, etc.) or wells and streams blessed by saints or other holy men to which devotion is made and from which healing is expected after immersion. A special, but universal, belief is the faith in the efficacy of bathing in certain streams for the restoration of fertility to barren women.

Certain great landmark rivers, the scene of both civic cults and private devotions, are believed to have general therapeutic and apotropaic powers. By immersion in the Euphrates (Iraq), the Abana, the Pharpar (in Damascus, Syria), the Jordan (Israel), the Tiber (Italy), the Nile (Egypt), or the Ganges, Jumna or Saravatī (all in India), one might be cured of disease, purified from transgression, or protected against future disorders.

These same basic features—unusual natural characteristics, scenes of epiphanies, locations associated with the life or the burial place of holy men, or great national landmarks—are present in each variety of healing shrines (*e.g.*, those associated with sacred trees, stones, or mountain peaks).

Healing cults centred around sacred organizations. As in the case of various monastic orders throughout Europe that have as their primary function the care of the sick (*e.g.*, the Knights Hospitallers, the Augustinian Nuns, the Order of the Holy Ghost, and the Sororites Order), healing has frequently been delegated to certain groups. Among these are special classes of priests (*e.g.*, the Akkadian Āshipu or Kalū priests, the Greek Asclepiads); religious castes (*e.g.*, various Brahmin groups in India, the Vaidya caste in Bengal); secret societies (*e.g.*, the Midēʻwiwim type groups among the American Indians—such groups can be highly specialized, for among the Sia Indians there are eight societies: one specializes in treating burns, one in ant bites, etc.); or dynasties of healers who trace their knowledge back to the gods (*e.g.*, the Physicians of Myddvai in Wales, who have been active herbalists for more than five centuries). The formation of such groups has been and is in part due to the priests' services at shrines and their possession and manipulation of certain sacred objects and relics that are the sources of the priestly charisma (supernatural power) of office. Most prominent are those priests who serve in the cults of healing deities (*e.g.*, Asclepius, Hygieia in Greek religion) or at shrines devoted to healing saints (*e.g.*, St. Cosmas and St. Damian in Christianity). The tendency to concentrate healing activities in specialized sacred organizations is also due to the length of training required to master the arts of healing, the need for special equipment and libraries, and the expense of maintaining such facilities—all of which may be readily borne by settled religious communities. Thus, many important religious leaders have also been physicians (*e.g.*, Mani, Moses Maimonides) and the origin of hospitals in both East and West is linked to religious orders.

Specialized healing societies and healers

Healing cults centred around individual healers. Healing may be accomplished by those who possess powers due to their office, such as priests and kings. More frequently, however, individuals are believed to cure by means of a special gift or sacred commission. They are holy men, and one means by which their sacrality is manifested is their power to heal. This power may be revealed in a vision; it may be sought after; or it may be accidentally discovered that an individual possesses such abilities.

Almost every religious founder, saint, and prophet has been credited with the ability to heal—either as a demonstration of or as a consequence of his holiness. In every culture there are also specialists who frequently have gone through extraordinary initiations that confer curative powers upon them. These individuals (*e.g.*, shamans, "medicine men," "folk doctors") may fill a cultural niche alongside certain religious groups, fulfilling their functions when circumstances so require. Some work within an established religious tradition but concentrate their energies primarily upon healing (*e.g.*, well-known Christian faith healers of the 19th and 20th centuries, such as John of Kronstadt; Chlodwig Karl Viktor, Fürst zu Hohenlohe-Schillingsfürst; Leslie Weatherhead; Edgar Cayce; and Oral Roberts). Others have founded their own religious communities that maintain a focus on healing (*e.g.*, Phineas P. Quimby and the New Thought movement; Mary Baker Eddy and Christian Science; the various independent churches of Africa; the Spiritist and Hallelujah movements among the Carib peoples in Brazil).

THE NATURE OF ILLNESS AND TECHNIQUES OF HEALING

The various practices of religious healing may be divided into three groups according to the varying views of the origin and nature of illness. The method of cure is determined by what is believed to be the cause of illness in a particular case.

Diseases related to spiritual powers. The first group assigns the cause of disease to the actions of various spiritual powers. Deities and demons, according to this view, in the beginning of creation acted in a universal way so that the forces of disease and death were unleashed, and the narrative of their actions is given in myths of origin. Usually, however, the action is specific and is directed against an individual.

Demon possession, object intrusion, and soul loss

In demon possession, the demon or spirit enters into the body of the afflicted individual, either obviously—manifesting itself in erratic behaviour—or secretly. In the latter case the healer must first determine, usually through some form of divination, that the sick person has in fact been possessed and also the identity of the spirit that possesses the individual. The spirit will then be driven out by various means: expelled through exorcistic techniques; purged by the use of herbs or other concoctions; removed surgically by opening the means for its egress; or transferred by magical means to another person, animal, or object.

Closely related is the notion that illness is caused by a spirit manifested in the form of a foreign body (usually a small stone) in the patient. Hence, it is believed, re-

moval of the object by exorcism, purgation, or surgery will restore health.

Illness is also believed to be caused by the loss of soul—the seat of vitality—most usually during sleep when the soul is believed to leave the body and wander. Many misadventures are possible: the soul may be captured by demons; it may come under the influence of a hostile magician; or it may simply lose its way. A specialist (usually a shaman) is required to retrieve the soul by means of magical flight or do combat with hostile powers in order to win its release.

Diseases related to human maleficence. The second group locates the cause of disease in the maleficence of other men. While witchcraft, in this view, may adopt the techniques of possession, object intrusion, or soul stealing to cause disease, the initiative comes from a human being who has special power because of his relation to the supernatural world. In curing, divinatory techniques are most important in order to determine the identity of the witch. Once this is accomplished countermagic, exorcism, or forms of placation will negate the baneful influence and restore health.

Diseases related to the afflicted individual. The third group places responsibility for illness on the afflicted individual. The sufferer, either consciously or unconsciously, has neglected a religious obligation or violated a sacred prohibition, and is thus punished through illness. If unconsciously, divination is required to learn the nature of the offense. Confession, expiation, and repentance will lighten or eliminate the effects of the transgression. In some religious traditions illness is seen as either an illusion or the result of lack of knowledge about the nature of the human condition. Curing is accomplished through enlightenment bestowed by a prophet or saviour.

EVALUATION

The assessment of healing cults remains a difficult problem. Some of the illnesses are psychosomatic in origin, some of the techniques are physiologically effective (*e.g.*, close to half of the herbs used in curing ceremonies are medically active); but such interpretations do not do justice to the religious power of these cults. They stem from and are a response to a profound and universal diagnosis of the human condition—man is weak and there are deep anomalies in his situation. By describing, mapping, interpreting, and organizing those elements in human experience that resist man's control and understanding, these cults aid man to understand himself and his world, and to live in it.

BIBLIOGRAPHY. There is no recent, reliable survey of this topic. The articles on "Disease and Medicine" and "Health and Gods of Healing," in J. HASTINGS (ed.), *Encyclopaedia of Religion and Ethics*, vol. 4, pp. 723–772, vol. 6, pp. 540–556 (1908–27), remain the best overall treatments. G.H.T. BUSCHAN, *Über Medizinzauber und Heilkunst im Leben der Völker* (1941), is a rich collection of material primarily drawn from primitive practices and contains a full bibliography. F.E. CLEMENTS, "Primitive Concepts of Disease," *University of California Publications in American Archaeology and Ethnology*, 32:185–252 (1932), is the classic study of primitive theories of the origin of disease. CLAUDE LEVI-STRAUSS, *Structural Anthropology*, ch. 9–10 (1963), provides a model of the serious anthropological interpretation of the "truth" of primitive healing practices. H.E. SIGERIST, *A History of Medicine*, vol. 1 (1951), surveys primitive and ancient Near Eastern medical practices; W.A. JAYNE, *Healing Gods of Ancient Civilizations* (1925), is most useful for Greco-Roman cults. D. MCKENZIE, *The Infancy of Medicine* (1927); and G.G. DAWSON, *Healing: Pagan and Christian* (1935), are informative, anecdotal descriptions of early Western and Christian folk medicine. WILLIAM JAMES, *The Varieties of Religious Experience: A Study in Human Nature* (1902), contains much documentary material on modern Christian healing cults and is the most profound treatment of the religious role and function of healing.

(J.Z.S.)

Health, Human

Health, like the weather or fortune, may be good or bad and is variously defined. It may be thought of as the extent of an individual's continuing physical, emotional, mental, and social ability to cope with his environment. René Dubos frames the problem succinctly in "Determinants of Health and Disease" (Britannica Perspectives, vol. I, p. 281; 1968):

> The all-inclusive and consequently vague meaning of the word health can be traced all the way back to its Anglo-Saxon root, which means "hale," "sound," "whole." Irrespective of precise medical criteria, the experience of feeling healthy has always consisted in being able to function well physically and mentally and to express the full range of one's potentialities. The preamble of the charter of the World Health Organization attempts to convey this utopian ideal in the following words: "Health is a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity." Health so defined is a utopian state indeed.

Definitions of health

Within this framework, good health for the librarian, the artist, or the department store clerk, for example, might be quite different from good health for the logger or the farmer or the steelworker. Such a definition has its drawbacks. The librarian, for instance, may be a rather fragile individual who stays "well" within the ordinary environment of his or her existence but succumbs to a heart attack from heavy shovelling after a snowstorm; or a sea-level dweller may move to a new home in the mountains, where the atmosphere has a lower content of oxygen, and suffer from shortness of breath and anemia until his red blood cell count adjusts itself to the altitude. Thus, even by this definition, the conception of good health must involve some allowance for change in the environment.

Bad health can be defined as the presence of disease, good health as its absence—particularly the absence of continuing disease, because the person afflicted with a sudden attack of seasickness, for example, may not be thought of as having lost his good health as a result of such a mishap. The same might apply to a pregnant woman, perfectly healthy in the afternoons and evenings but suffering from morning sickness a few hours each day.

Actually, there is a wide variable area between health and disease. Only a few examples are necessary to illustrate the point: (1) It is physiologically normal for an individual, 15 to 20 minutes after eating a meal, to have a high blood sugar content. If, however, the sugar content remains elevated two hours later, this condition is abnormal and may be indicative of disease. (2) A "healthy" individual may have developed an allergy, perhaps during early childhood, to a single specific substance. If he never again comes in contact with the antigen that causes the allergy, all other factors remaining normal, he will remain in that state of health. Should he, however, come in contact with that allergen, even 20 or 30 years later, he may suffer anything from a mild allergic reaction—a simple rash—to severe anaphylactic shock, coma, or even death, depending upon the circumstances. (3) An apparently healthy individual may imbibe a relatively large amount of alcoholic beverage. After the alcohol reaches certain levels in his bloodstream he exhibits certain behavioral changes, and his condition is referred to as intoxication. During this period, and usually the morning after, he cannot be considered as being in good health. He may regain his normal health in a few days or less; but frequent repetition of this performance, or daily ingestion of large amounts of alcohol, even when no apparent behavioral or hangover effects are noted, insidiously convert a condition of good health to bad health—sometimes irreversibly. Thus it can be seen that, unlike disease, which is frequently recognizable, tangible, and rather easily defined, health is a somewhat nebulous condition, difficult to define and never in a state of perfection—*i.e.*, one can be "really sick" but never "perfectly healthy."

Moreover, physical condition and health are not synonymous terms. A seven-foot-tall basketball player may be in excellent physical condition (although outside the range of normality for height) but may or may not be in good health—depending, for example, on whether or not he has fallen victim to an attack of influenza. The one-

Table 1: Normal Values for Human Blood

characteristic or property	range of normality	chemical component	range of normality*
Volume	7–9% body weight (about 5 quarts)	Glucose (blood)	80–120
pH (acid–base measurement)	7.35–7.45	Serum protein (total)	5.9–7.5 g
Red blood cells	4,500,000–5,500,000 per cu mm	Albumin–globulin (Alg) ratio (plasma)	1.3:1–2.9:1
White blood cells	5,000–10,000 per cu mm	Fibrinogen (plasma)	290–500
Platelets	200,000–400,000 per cu mm	Nonprotein nitrogen (blood NPN)	25–40
Hemoglobin content	14–16 g per 100 ml	Blood urea nitrogen (BUN)	8.0–20
Hematocrit	47 ± 5%	Uric acid (blood)	3.2–5
Colour index	0.9–1.1	Cholesterol, total (serum)	130–250
Volume index	0.9–1.1	Iodine, protein-bound (blood)	4.0–8.5μg
Bleeding time	1–3 min	Sodium (serum)	312–342
Coagulation time	5.5–12.5 min	Potassium (serum)	14–21
Specific gravity (25° C)	1.05–1.06	Calcium (serum)	8.5–11.5
Relative viscosity (38° C)	4.7	Phosphorus, inorganic (serum)	2.5–4.0
Sedimentation rate	0–20 mm (first hour)		
Prothrombin time	10–15 sec		

*Values except where indicated are milligrams per 100 millilitres.

armed gymnast, the colour-blind skater, or the pianist born blind—all may be in good health; but are they in good physical condition? Again, this depends upon definitions, and definitions vary.

Moreover, factors relating to specific physical fitness may complicate the estimate of a particular person's state of health. Specific physical fitness is the possession of the physical attributes and ability to function that are required for a job or situation that is particularly taxing. Physical strength, endurance, ability to work without oxygen in brief bursts of activity, ability to consume adequate oxygen for taxing efforts that take a longer time, ability to make sudden changes in posture, ability to move quickly, accurately, and with agility—these all are qualities that might be required in particular situations. A person who had been "in perfect health" in one environment might suddenly seem unhealthy if he were confronted with a new job or a new situation that made demands upon him not previously encountered.

There are further problems in settling upon a definition of human health. A person may be physically strong, resistant to infection, able to cope with physical hardship and other features of his physical environment, and may still be considered unhealthy if his mental state, as measured by his behaviour, is deemed unsound. What is mental health? Some say that a person is mentally healthy if he is able to function reasonably well. Others hold that a person is healthy mentally if his behaviour is like that of a majority of his fellows. A third group makes comparisons with an ideal. According to these physicians mental healthfulness may be approached but not attained. The originating of this attitude is attributed to Sigmund Freud, who is quoted as saying, "A normal ego is like normality in general, an ideal fiction." Still another concept stresses the changes in a person's behaviour that take place with the passage of time as criteria of his mental health.

In the face of this confusion, it is most useful, perhaps, to define health, good or bad, in terms that can be measured, can be interpreted with respect to the ability of the individual at the time of measurement to function in a normal manner and with respect to the likelihood of imminent disease. These measurements can be found in tables of "normal values" printed in textbooks of clinical medicine, diagnosis, and other references of this type. When an individual is given a health examination, the examination is likely to include a series of tests. Some of these tests are more descriptive than quantitative and can indicate the presence of disease in a seemingly healthy person. Such tests include the electrocardiogram to detect some kinds of heart disease; electromyogram for primary muscle disorders, liver and gall bladder function tests; and many types of X-ray techniques for determining disease or malfunction of internal organs.

The health examination

Regarding the changing pattern of health from primitive times, Dubos wrote:

> In the course of evolution, as his characteristic structural, physiological, and mental equipment gradually emerged, ancient man developed the fitness to resist environmental threats, those dangers emanating from cosmic forces, food shortages, microbial parasites, wild animals, or human competitors.
>
> Most of the skeletal remains found in Paleolithic and Neolithic sites are of vigorous adults essentially free of organic diseases at the time of death. Human remains of more recent origin provide further evidence of primitive man's ability to resist harsh natural conditions, at least until he is exposed to influences of Western civilization. A large burial ground from a period preceding Captain Cook's discovery of the Hawaiian Islands was recently excavated in Honolulu; the skeletons recovered from the site had healthy teeth and strong bones with powerful muscle attachments. Apparently life on the Hawaiian Islands was compatible with health and vigour even under the primitive conditions of pre-European settlements. Similar discoveries made in other parts of the world validate the legend of the healthy, happy savage appearing in accounts of primitive life written by 17th- and 18th-century explorers.
>
> Recent medical surveys of contemporary African, American Indian, and Australian tribes give us even more convincing evidence that health and vigour can be achieved under primitive conditions in extremely harsh climates. During the 1960s, Western physicians, biologists, and anthropologists studied the Meban Negroes of East Africa and the Chavanté Indians of the Brazilian Mato Grosso in their own undisturbed environments. The Meban and Chavanté tribes live in isolated primitive villages, with limited food resources, under difficult climatic conditions and out of contact with Western technology or medicine. Men in both tribes were found to be extremely vigorous and of magnificent physique. They were essentially free from dental caries, high blood pressure, cancer, and other degenerative diseases so common in civilized, prosperous countries.

Other tests give numerical results (or results that can be assigned numerical values—such as photometric colour determinations) that can be interpreted by the examiner. These are physical and chemical tests, including blood, urine, and cerebral spinal fluid analyses. The results of the tests are compared with the normal values; and the physician receives clues as to the health of his patient and, if the values are abnormal, for the methods of improving the health.

A major difficulty in the interpretation of test results is that of biological variability. Almost without exception these normal values for variables are means or adjusted means of large group measurements. To give these values significance—hence the frequent use of the word

Table 2: Normal Values for Human Urine

property or constituent	range of normality (grams)*
Acidity	slightly acid
Colour	pale yellow to amber
Specific gravity	1.0025–1.025
Amount	1.25–1.5 litres
Urea	18–32
Uric acid	0.5–0.75
Creatinine	0.35–0.45
Ammonia	0.5–15.0
Nitrogen, total	9.5–16.5
Sulfur, total	1.0–3.5
Phosphate	1.95–3.75
Chloride	10–15
Sodium	2.5–7.0
17-Ketosteroids	
Male	.005–.027
Female	.005–.015
Solids, total	45–75

*Quantitative values, unless otherwise indicated, are grams in a 24-hour sample of urine from an average adult; variation may be wide depending on dietary intake.

standard rather than normal in describing them—they must be considered as lying somewhere near the centre point of a 95 percent range; *i.e.*, the so-called ordinary range or, with reservations, the range from normal to the upper and lower borderline limits. Thus, the 2.5 percent below the lower limit and the 2.5 percent above the upper limit of the 95 percent range are considered areas of abnormality or, perhaps, illness. Some areas have wide 95 percent ranges—blood pressure, for example, may vary considerably throughout the day (*e.g.*, during exercise, fright, or anger) and remain within its range of normality.

Other values have ranges so narrow that they are termed physiological constants. An individual's body temperature, for example, rarely varies (when taken at the same anatomical site) by more than a degree (from time of rising until bedtime) without being indicative of infection or other illness.

Table 3: Normal Values for Human Cerebrospinal Fluid

property or constituent	range of normality (mg per 100 ml CSF)*
Volume	90–150 ml
Specific gravity	1.006–1.008
Colour	clear, colourless
pH (acid–base measurement)	7.35–7.70
Freezing point	−0.6–0.5° C
Pressure	70–80 ml mercury
Solids, total	850–1,700
Protein, total	12–43
Chloride	420–450
Sodium	500–545
Nonprotein nitrogen	12–30
Cells	0–10 per cu mm

*Quantitative values, unless otherwise indicated, are milligrams per 100 millilitres of CSF.

In the accompanying tables may be found a selection of ranges of normality that are frequently used in physical and clinical diagnosis and in health examinations. In general, values are calculated averages for adult males and nonpregnant, nonlactating females.

For conditions in which values are higher or lower than the range of normality, see BLOOD DISEASES; CEREBROSPINAL FLUID; DIAGNOSIS; DISEASE, HUMAN; ENDOCRINE SYSTEM DISEASES AND DISORDERS; EXCRETORY SYSTEM DISEASES. (W.Sp.)

Health and Disease, Economics of

Health economics is the study of the use of resources for health services and the effects of health and ill health on the economy. Health economists attempt to measure the quantity and economic effects of health services for a given period and how these services are financed. They also study the economic effects of different types of health-service organization and payment systems; the efficiency with which resources are deployed; and the effects of disease, the cure of disease, and standards of health on individual and national production.

Health resources. The resources used in health services consist of (1) buildings, such as hospitals, nursing homes, health centres, clinics, surgeries, offices, dispensaries, shops where health goods are sold, and factories where health goods are made; (2) personnel, such as physicians, dentists, opticians, pharmacists, auxiliaries—such as medical assistants and medical aides, who may diagnose and treat and auxiliaries who provide treatments normally under medical direction (for example, physical therapists, speech therapists, radiographers)—nursing staff, research staff, teachers of the health professions, and the administrative, secretarial, and domestic staff who service health institutions; and (3) materials and equipment, such as pharmaceuticals, nonprescription drugs, dressings, X-ray equipment, surgical and scientific instruments, uniforms, and hospital food.

Many of the goods and services used in health services are used for other purposes as well—for example, food and domestic labour. The precise border of the health "industry" can only be defined by considering the purpose for which particular resources are used. Sunglasses, tonics, and mineral water, for example, may or may not be used or consumed for health purposes. There is no internationally accepted definition of pharmaceuticals; the scope of pharmaceutical preparations that can be obtained only on the prescription of a physician varies among the different countries of the world. Even the precise scope covered by the term hospital is not the same in different parts of the world. Many countries have residential institutions—which may or may not be called nursing homes, infirmaries, or hospitals—that provide care both for those requiring continuous nursing and medical supervision and those who need only welfare care; *e.g.*, the aged and mentally retarded.

EXPENDITURE ON DIFFERENT TYPES OF SERVICE

Categories of health expenditure

Health-service expenditure can be subdivided into capital expenditure—expenditure on buildings (new construction and conversion) and the purchase of durable goods for health purposes—and current expenditure. Current expenditure includes expenditure for (1) personal health services, including all curative, preventive, and promotive services given to identifiable individuals (such services can be given inside or outside a hospital, and can be received by inpatients, ambulatory patients, or patients in their own homes); (2) nonpersonal public health services, to promote a healthy environment, the benefits of which accrue to individuals not readily identifiable; and (3) teaching and research.

A study organized by the World Health Organization of the United Nations, the latest study of this scope that is available, attempted to ascertain expenditure on health services in different countries in as comparable a form as possible. The study was for the year 1961 or the nearest year for which information could be made available. Data from this study are used in the tables and diagrams in this article. Table 1 shows current expenditure (whether financed collectively or privately) on health services in 13 countries. About one-half of the expenditure on health services was channelled through hospitals, though the exact proportion varied considerably between countries.

About one-half of the expenditure (ranging from 37 percent in Czechoslovakia to about 60 percent in the United Kingdom) on personal health care was for staff. The earnings of medical and dental staff represented about 16 percent of health expenditure in the United Kingdom and nearly 28 percent in Canada. Expenditure on nursing personnel was about one-seventh of expenditure in both countries. Expenditure on medicaments was about 18 percent of expenditure in Canada and 9 percent in the United Kingdom.

Table 1: Percentage Distribution of Expenditure on Health Services
(1961 or nearest year)

	personal health services		nonpersonal public health	teaching and research
	through hospitals	outside hospitals		
Australia	43.7	52.9	1.9	1.5
Canada	42.7	54.5	1.0	1.8
Ceylon (Sri Lanka)	50.0	43.9	4.5	1.7
Czechoslovakia	47.7*	43.6	2.4	6.3
Finland	59.9	35.9	2.5	1.7
France	40.7	55.8	1.7	1.6
Israel	45.3	50.0	1.6	2.7
Poland	39.9*	53.2	2.8	4.1
Sweden	52.9	42.4	1.2	3.5
Tanganyika	44.8	50.2	4.3	0.8
U.K.	51.4	44.3	1.9	2.1
U.S.	36.2	56.9	0.3	4.6
Yugoslavia	40.7	49.9	3.9	3.0

*Including balneological institutions (health baths).
Source: Brian Abel-Smith, *An International Study of Health Expenditure*, 1967. Further qualifications to the above data are fully set out in that volume.

Capital expenditure on health services (excluding that on factories making medical goods and the construction of offices by private practitioners) normally amounts to between 3 percent and 12 percent of health expenditure in different countries. In most but by no means all countries, the bulk of the capital expenditure goes for hospitals.

FINANCING HEALTH SERVICES

The right to receive medical service. Public health measures to protect the environment (for example, the control of food, water, and the disposal of sewage) have to be collectively financed; this function is normally performed by government (central, regional, or local). Similarly, medical research has to be collectively financed. In both cases this is justified by economists on the ground that the benefits are expected to accrue to society as a whole and that these social benefits are greater than the benefits that go to individuals and for which individuals might be willing to pay. Similar arguments are used to support the case for collective provisions to contain and treat communicable diseases. The benefit from the treatment of a person suffering from communicable disease is not confined to the individual being treated, but extends to all for whom the risk of infection is reduced.

The special problems of financing health services stem from the growing belief that people ought to receive medical care for humanitarian reasons, regardless of their ability to pay. The preamble to the Constitution of the World Health Organization states, "The enjoyment of the highest attainable standard of health is one of the fundamental rights of every human being." Many countries have attempted to provide for this right to medical care by legislation. In some countries, the question of whether medical care should be provided as a legal right or as a privilege is still not resolved. In others, it would involve a high proportion of national resources to attempt to provide comprehensive health services. Indeed, resources may be inadequate to secure adequate levels of nutrition.

The question of medical care as a legal right or a privilege

Inability to pay for medical care is not of the same character as inability to pay for such other necessities as food and shelter. Though the total incidence of sickness may be predictable for society as a whole, sickness, like unemployment, is for the individual an unpredictable risk that can stop the flow of earnings. While provision for unemployment can be made by a cash allowance that is related to previous earnings or that provides or secures a subsistence or minimum level specified for the particular country or area, a standard weekly or monthly cash benefit cannot ensure that any bills for medical care can be paid. This is because the expenditure required per day or week to meet medically determined need can vary enormously according to the nature of the illness or injury.

The problem of ensuring that the whole population is in a position to purchase the medical care that physicians recommend has become more and more acute, because, although levels of living have been rising in many countries, the possible useful expenditure on medical care per unit of time has been rising much faster, largely because of the growth in medical knowledge and the development of more and more expensive diagnostic and treatment procedures.

Thus, families with modest savings and moderate incomes—families that would not be regarded as poor or indigent—may, nevertheless, be unable to pay for the medical care of a family member. Hence the use of the term medically indigent to describe people who normally can pay their way but cannot afford to bear the extraordinary costs of illness. Some countries have a higher income level of eligibility for health and medical care under public assistance or relief programs than is used to provide for ordinary living expenses. Criteria used to define medical indigence vary widely. Much depends on the extent to which medical costs are related to the capacity to pay and on the level of living of the particular country or community, quite apart from the policies and sources of income of the public authority providing for the indigent.

The history of financing health services. Many societies once responded to the problem of medical indigency by the provision of services on a charitable basis. In Europe and America, hospitals and dispensaries were endowed by public subscription or by religious bodies to provide medical services for the poor. The physicians working part time in such hospitals gave their services or received a nominal payment. Soon after they were founded, some such hospitals in the United States began to take paying patients, whom the physicians billed for their services. It was not until the 1890s that the extensively developed voluntary hospitals for acute disease in England opened beds for paying patients.

In other countries (for example, Sweden and the Soviet Union), government played a major role in developing hospitals. In general, governments have tended to make larger and earlier provision of hospital facilities for mental illness and infectious disease than for acute sickness. There was a clear advantage for the rest of the public if the infectious and the insane were isolated in hospitals, quite apart from the treatment provided for the individuals with these afflictions.

In the case of physician care, the physician has long been considered to have an obligation to provide for the poor. This obligation has led to the acceptance by many societies of the physician's right to discriminate in his fees—to charge the rich more than those with middle and low incomes. As other solutions to the problem of providing for the medically indigent have been developed, it has been questioned whether the continuation of price discrimination of this kind can still be justified.

In Spain, an early solution to the problem of paying medical bills was for the physician to receive regular payments from local inhabitants in return for his provision of services as required. Informal physician-run insurance also developed in The Netherlands and in some other countries, but, in most of Europe, informal and, later, formal health insurance was developed most extensively under the control of local consumer or occupational groups. The origin of this mutual insurance can be traced back to the early guilds, medical clubs, and benefit and friendly societies. Local or occupational sick funds and friendly societies became extensively developed in Germany, Denmark, The Netherlands, the United Kingdom, and other countries by the end of the 19th century. Gradually the arrangements were formalized and developed under national-health-insurance legislation, starting in Germany in 1883 and extending to other European countries during the 20th century.

Informal health insurance

A fourth solution to the problem of making available the services of physicians to poorer people was by the employment of physicians by the government. There is a long tradition of employment of government medical officers in Sweden to work in designated districts, and the system was also introduced in Russia under the tsars. The developing countries of Africa and Asia have also gener-

ally provided for the needs of the medically indigent—the majority of the population—by providing health care in and out of hospitals by a national or regional comprehensive service employing salaried physicians. This service often is available to private patients also, but in most such countries the more affluent section of the population obtains the bulk of its health services through private arrangements. National- or state-planned services, available to the whole population, are the pattern in eastern Europe. In the Union of Soviet Socialist Republics, the long tradition of government hospitals and government clinics was extended by the Communist regime in the interwar period, and a health service run by the government has been adopted by eastern European governments influenced by the Soviet Union.

Since 1948 the United Kingdom has provided a national health service, offering comprehensive services to the whole population either free or subject to nominal charges, and most hospitals were taken into public ownership at that time. Physicians' services outside the hospitals are provided by generalist physicians who are not employees of the government but have contracted to provide services to designated patients registered with them. A small private sector of hospital care continues in Britain—largely financed by voluntary insurance.

In nearly all countries of the world, the majority of hospital beds are owned by government. The main exceptions are Canada, The Netherlands, and the United States. In the latter, most hospitals for the mentally ill are run by public authorities, but most hospitals for other patients are controlled by nonprofit corporations.

ASSESSING PAYMENTS

Physicians' services. *Fee for service.* Under private practice, the physician normally bills the patient for each service he has performed (such as an office call, home visit, injection, or operation). This system of remuneration is currently used under most voluntary and national health-insurance systems. In most cases elaborate lists of services, with a designated fee for each service, have been negotiated with the profession by the insurance system. Usually the negotiated fee represents the total that the physician can claim for the service, whether (as is usually the case) he makes his claim direct to the insurer or whether he claims from the patient, who in turn is reimbursed in whole or part by the insurer. In some countries (for example, in Sweden), however, physicians (or certain classes of physicians) are free to bill the patient with an addition to the fee he receives from the scheme. In the United States, there is no negotiated fee under Medicare (compulsory health insurance for the aged) or Medicaid (the program for the medically indigent). The physician is free to charge his customary fee.

Types of fee for service

Problems in fee-for-service payments

One difficulty with fee-for-service payments is that the physician may encourage unnecessary visits and perform services that are not strictly necessary. Regulations and provisions within fee structures are used to limit the effect of this type of incentive. In France, Switzerland, and The Netherlands, control physicians are employed by the sick funds to detect and disallow or recommend disallowance of claims by physicians. In the Federal Republic of Germany (West Germany), this function is performed by the profession itself. While at first sight, fee for service would suggest greater freedom for the physician, in practice there tend to be more controls under health-insurance schemes than in other systems of remuneration.

When all or part of the fee falls on the patient, some persons postpone seeking treatment. Persons with low incomes may also postpone treatment under a system of total reimbursement, because of the lack of ready cash to finance the bill that is awaiting settlement. It has been suggested that a fee-for-service system leads to inadequate attention to preventive medicine. When fees are standardized under compulsory-insurance systems, there is no clear incentive for work of superior quality. While the better physicians can be expected to attract more patients (if patients are aware of their superior skill), in so far as more time is required to provide good care, the physician tends to receive lower rewards.

Salary. Public-health physicians throughout the world are generally paid on a salaried basis. This system of remuneration is also the pattern for nearly all physicians in eastern Europe and most physicians in Israel and in the health services of most of Asia and central and southern Africa for the general population.

In most countries in western Europe, separate specialist physicians work in the larger hospitals to whom generalist physicians working in the community transfer the care of patients who are admitted. These hospital physicians are generally paid on a full-time or part-time salaried basis. Part-time salaries are extensively used for hospital physicians in Spain and Latin America.

In theory, salaried payment could lead to lack of diligence in caring for patients, brusque behaviour, and poor attendance, but much depends on the extent of commitment and supervision. The possibility of promotion based on merit within a hierarchical salaried structure can be an effective incentive to painstaking work. Difficulties are greatest when a part-time salary is paid and the physician spends the rest of his time in private practice, which is more highly remunerated.

Capitation. Capitation is a fee paid normally to an out-of-hospital physician for offering services to a patient for a period of time (*e.g.*, a quarter year or a year). The payment is made whether the patient does or does not need to use the services of the physician during the period. The rate of capitation may be reduced when a physician's list exceeds a particular figure.

This system of payment is used under compulsory health insurance in Italy, The Netherlands, Spain, and the United Kingdom. In the latter, a higher capitation fee is paid for a person over age 65, in view of the greater need for medical care. The generalist physician also receives certain flat-rate allowances, which do not vary with the size of his list of patients, and a few fees for specific services (preventive services, maternity cases, and out-of-hours calls).

In theory, capitation payment could lead to less than diligent care. In addition, there is a clear incentive for unnecessary referral to specialists. In both Italy and The Netherlands, this aspect is closely examined by officials of sick funds and may lead to complaints to the local medical association. On the other hand, capitation payment creates an incentive for preventive care and encourages continuity of generalist care.

Incentive for preventive care

Capitation is used in some programs in the United States as a payment by the insurer to a medical group for comprehensive health care. The medical group may make other types of arrangement for the remuneration of physicians working in the plan.

Case payment. Under private practice a physician may make a quotation per case; for example, for the total care of a maternity case. The fee might cover care while the patient is in the hospital and checkups afterward.

Case payment has also been used as a system of paying generalist physicians under health insurance. It differs from capitation in so far as payment is made only for those patients actually treated. It developed in the Germanic countries in the 19th century and still survives in parts of Austria. In general, this system has been replaced by fee-for-service remuneration.

The case method is also used for specialists in The Netherlands. While they are paid on a fee-for-service basis for technical, surgical, and similar defined procedures, they collect case payments for the routine care of ambulatory patients and hospital inpatients.

Payments for hospital care. The most widely used system of charging for hospital care is by the length of stay in the hospital. Charges are generally calculated per day of residence. When the physicians are in a salaried relationship to the hospital, it is common for the charge to cover all services provided, including those of physicians. When a physician is not in a salaried relationship with the hospital, he generally submits a separate bill for his services, per case or per item of service. Thus a hospital patient may receive bills from several physicians if he has needed the services of more than one specialty. In a profit-oriented hospital system, the daily charge may

cover only the charge for the room and basic nursing and food. All other items may be billed separately (*e.g.*, such items as special nursing, use of the operating theatre, pathological tests, X-rays, pharmaceuticals, blood transfused, and special diet).

SOURCES OF PAYMENT OF COSTS

Funds for payment of costs may come from indirect or direct payments.

Indirect payments. *Government.* Government sources of funds include expenditure from all government agencies, whether central, state, or local, or those made by the health or other departments or from special funds belonging to government or statutory bodies. All charges and fees paid by the recipient of services provided by government agencies are classified as payments by recipients (see below).

Compulsory social insurance. Compulsory social insurance covers all payments for goods and services, irrespective of the original source of the money (but excluding fees or charges), under an official compulsory scheme; the right to benefit is limited in principle to persons for whom obligatory contributions have been paid (either by themselves or by others) and whose possession of private means does not offset the right to benefit. Such schemes may be designated as sickness, invalidity, or maternity insurance or as workman's compensation.

Other agencies. Other indirect sources of payment are voluntary insurance funds, including nonprofit and profit-making insurance; voluntary subscriptions and charitable funds; grants from outside the country; and other corporate funds.

Types of voluntary insurance

Insurance is regarded as voluntary even when membership is a condition of employment. Contributions from the government as an employer for its employees are counted as private insurance. Voluntary insurance includes (1) nonprofit insurance, which covers voluntary social insurance and insurance that may be operated by friendly societies, mutual-benefit societies, trade unions, and any other bodies that are non-profit-making, and (2) profit-making insurance, which covers all private profit-making carriers and includes cash payments made for the reimbursement of medical costs.

Voluntary subscriptions and charitable funds are moneys contributed in such a way that the contributor himself has no right to benefit from the fund contributed. Also included in this category is any income from endowments.

Grants from outside the country include grants received from foreign governments under aid programs, from international agencies, and from trusts and individuals abroad. Many developing countries receive extensive help from international agencies and missions.

Other corporate funds include services provided in kind by employers for their employees.

Direct payment by recipient. All nonreimbursed payments, charges, and fees falling upon individuals in return for which health services were received are included in this category.

The percentage of funds estimated to have been contributed from these four main sources of funds for years approximating to 1961 in 17 countries are shown in Table 2.

INTERNATIONAL COMPARISONS OF HEALTH EXPENDITURE

Because prices of health goods, levels of remuneration for those working in health services, and the types of services provided under different health systems may differ substantially, it is not possible at present to make meaningful comparisons of the quantity of health services in different countries. Moreover, such comparisons lack meaning without some consideration of quality, which is particularly hard to define in this field.

It is possible, however, to make comparisons of the proportion of each country's total resources that is used in the health sector. These comparisons relate health-service expenditure in each country to the gross national product or national income of that country. Comparisons of this kind involve distortions if the relationship between health prices and earnings and other prices and earnings differ between countries. In Figures 1 and 2 the percentage of national income devoted to total and to indirect current expenditure on health services (excluding investment in plant) are plotted against per capita national income in United States dollars. On the whole, the countries with higher levels of income tend to spend a higher proportion of their national income on health services, but there are interesting variations. Chile and Israel stand out as high spenders for their level of income. The United Kingdom is a low spender for its level of income.

By courtesy of the World Health Organization

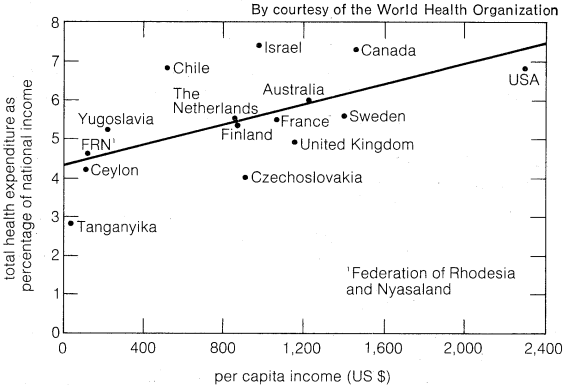


Figure 1: Percentage of national income devoted to health expenditures in 15 selected countries for the year 1961, or the year closest to 1961 for which information is available.

Table 2: Sources of Current Expenditure on Health Services

	unit of currency	total (millions)	general government (millions)	percent	compulsory social insurance (millions)	percent	other agencies (millions)	percent	recipients (millions)	percent
Australia (1960–61)	Australian pound	346	176	50.9	8	2.3	29	8.2	133	38.5
Canada (1961)	Canadian dollar	2,045	376	18.4	604	29.6	319	15.6	745	36.4
Ceylon (1957–58)	rupee	214	125	58.2	...	...	9	4.1	81	37.7
Czechoslovakia (1961)	koruna	7,264	6,259	86.2	377.5	5.2	124	1.7	503.5	7.0
Federation of Rhodesia and Nyasaland (1960–61)	pound	21.3	3.6	40.4	...	...	4.5	21.1	8.2	38.5
Finland (1961)	markka	691	402	58.2	...	...	10	1.4	279	40.4
France* (1963)	French franc	16,374	2,456	15.0	10,981	67.1	1,218	7.4	1,719	10.5
Israel (1961–62)	Israeli pound	320	91	28.5	10	3.25	133	41.5	86	26.8
Kenya (1961–62)	pound	8.55	3.18	37.2	0.47	5.5	0.51	6.0	4.39	51.4
The Netherlands* (1963)	guilder	2,344	486	20.7	994	42.4	98	4.2	767	32.7
Poland (1961)	zloty	17,157	13,628	79.4	...	...	...	...	3,529	20.6
Sweden (1962)	krona	3,683	2,440	66.3	450	12.2	...	...	793	21.5
Tanganyika (1961–62)	pound	5.36	2.82	52.6	...	...	0.84	15.7	1.7	31.7
U.K. (1961–62)	pound	1,088	922	84.7	...	...	9.6	0.9	156.4	14.4
U.S. (1961–62)	U.S. dollar	29,859	6,683	22.4	491	1.7	8,456	29.3	14,228	47.6
Yugoslavia* (1961)	dinar	176,873	17,363	9.8	135.529	76.6	12,981	7.3	11,000	6.2
Chile (1961)	escudo	304.3	70.3	23.1	20.5	6.7	41.7	13.7	171.8	56.5

*Including depreciation.

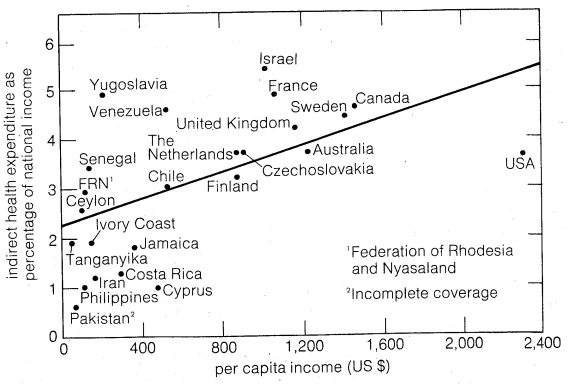


Figure 2: Percentage of national income devoted to indirect health expenditures in 24 selected countries for the year 1961, or the year closest to 1961 for which information is available.

By courtesy of the World Health Organization

Figures for indirect expenditure are more meaningful for developing countries, where the private sector involves only a small minority of the population. Figure 2 shows a slightly closer relationship between indirect expenditure and level of income.

Financing by local government

Among the high-income countries, those with multiple systems of financing health services seem to spend more than those with centralized health-service systems. In Scandinavian countries, local government plays a major role in financing health services. In continental western Europe and now Canada, compulsory insurance is the major system for financing health services. In the United States, voluntary insurance plays a major role, but the role of the public sector has increased in recent years.

Value for money. It does not follow that those countries that spend the most on health services necessarily provide either the largest amount of care or provide care of the highest quality or have the highest standard of health. High spending may be caused by relatively high remuneration for the professions involved, by lavish services enjoyed by a few while the needs of others (for example, the rural population or the minority groups) are relatively neglected, by high levels of profit (for example, for insurers or suppliers of medical goods—notably pharmaceuticals), by heavy administrative costs, or by duplicated and underutilized facilities.

A special feature of the medical-care market is that the consumer is in a poor position to assess the quality of care that he is given, quite apart from the comfort and courtesy with which it is delivered. As the consumer is not well equipped to assess the product he is purchasing, he can demand efficiency among suppliers of medical services only to a limited extent. The services that he needs may be largely selected by his physician, but the latter may have a regular association with a particular local health facility and may be neither well placed nor motivated to obtain the most favourable price, as distinct from desirable care, for his patient. If the bulk of the cost is paid by a third party, much depends on the extent to which the latter uses or is in a position to use bargaining power to monitor the quality of care and thus to promote the efficient use of health resources, to prevent unnecessary services, and to steer patients toward facilities that provide the greatest value for money.

The most expensive part of the medical-care system is the hospital. If care out of hospital has to be paid for, and care in hospital is free or nearly free, there will be incentives for patients to overuse the hospital, even though less expensive care in other types of institutions or in the community may be no less efficacious. The knowledge, incentives, and powers of patients, physicians, and financing agencies can all have important implications on the efficiency with which health resources are used.

Trends in cost. It might be expected that providing extensive health services and making them available to the whole population would in time reduce the amount that needs to be spent because of the general improvement in the health of the population; for example, the great reduction in the incidence of infectious diseases in the more developed countries (particularly with regard to tuberculosis since World War II) has undoubtedly led to substantial savings in expenditure on persons with these diseases.

Nevertheless, most developed countries are steadily spending more on health services—both absolutely and as a proportion of their total resources. In the 1950s it appeared that such countries devoted roughly an additional 1 percent of their total resources (gross national product) to health services over the decade. To this rule the United Kindgom was an exception, but it achieved about this rate of increase in the later 1960s.

Rising expenditure on health care

Expenditure on personal health care in Canada has risen steadily, from 3.1 percent of gross national product in 1957 to 4.9 percent in 1969. In the United States, total health expenditure rose from 3.5 percent of gross national product in 1929 to 6.6 percent in 1968. Projections suggest that the proportion of health expenditure may rise to between 8 and 9.8 percent of gross national product by 1980. If present trends continue, there may be several countries that devote a tenth of their total resources to health services by the end of this century.

Three main reasons can be identified for the growth in health expenditure. The first is demographic. Older people consume more health services than younger people. Thus, part of the increase in health expenditure is caused by the changing age structure of the population. People who would have died relatively young from an infectious disease are living to become victims of the degenerative diseases of old age, many of which are costly to treat. The second reason arises from the fact that health services consist to a large extent of personal care given by physicians, nurses, and others. While mechanization can reduce the labour requirement in certain areas of the health sector, the scope for labour saving is much more limited than in manufacturing industry. Third, and most important of all, the expansion of medical knowledge, new pharmaceuticals, and new treatment procedures, from transplants to kidney machines, have greatly extended the range of care that physicians can provide for the individual patient. These new techniques, both of diagnosis and of treatment, have greatly increased the staff and equipment needed to care for the average hospital patient, and it is in this sector of health services that costs are rising most rapidly. Expenditure on all hospitals in Canada, for example, rose from 1.8 percent of gross national product in 1957 to 3.1 percent in 1969.

BIBLIOGRAPHY. H.E. KLARMAN, *The Economics of Health* (1965), provides a useful introduction to the application of basic economic principles to the health service. J. HOGARTH, *The Payment of the General Practitioner* (1963), describes how general practitioners are paid under European health insurance systems and the development of and debates concerning these systems. W.A. GLASER, *Paying the Doctor* (1970), covers hospital doctors as well as general practitioners and discusses the effects of systems of remuneration. B. ABELSMITH, *An International Study of Health Expenditure* (1967), makes comparisons of health expenditure, both public and private, in 29 countries. H.M. and A.R. SOMERS, *Doctors, Patients and Health Insurance* (1961), is still a basic text that describes comprehensively the dilemmas of medical care in the United States; while the Canadian ROYAL COMMISSION ON HEALTH SERVICES, *Report*, vol. 1 (1964), performs a similar function for Canada.

(B.A.-S.)

Health and Safety Laws

The protection of public health and safety has been one of the first and most important objectives of government and law. In primitive cultures and in Western civilization until the Middle Ages, public health usually was protected in a negative way by either isolating or destroying the diseased in order to protect the healthy. Modern public-health laws, social legislation, and preventive medicine have been largely the products of scientific advancements of the 19th and 20th centuries.

The interests of the individual and of society in health and safety find expression in the orders of executives, the ordinances and statutes of legislatures and councils,

and the decisions of courts and administrative bodies. That such interests are among the most important values is suggested by the deference that courts have shown for legislative and executive formulations of health and safety policy. In the United States—*e.g.*, even during periods when courts wielded their power of judicial review to unprecedented degrees and readily substituted judicial preference for legislative judgment—health and safety laws were subject to milder scrutiny than measures involving, for example, economic experimentation. The common law at times gave priority to health and safety matters over property rights and economic interests.

Hazardous undertakings and the creation or maintenance of dangerous conditions or objects have long been a basis for special classification. English doctrine has imposed strict liability for the escape of dangerous substances, and the concept of nuisance has been construed to cover such cases in other jurisdictions.

Health and safety implications of the common law

The common law created an implied warranty of fitness for human consumption in the sale of foodstuffs, and liability for deleterious products today may be based on the fact of sale rather than on negligence in processing or failure to inspect. Summary procedure and something less than the due process usually required may be invoked when health authorities or fire marshals act to cope with emergencies, and both the police power of the state and the power of eminent domain may be utilized to protect or promote public health and safety.

During the 1960s measures to promote public health and safety proliferated. In the United States, Congress for the first time undertook to establish safety standards for automobile manufacturers; and dangers occasioned by atomic reactors, radioactive isotopes (unstable, radiation-emitting forms of substances), and related devices or substances led to the enactment of amendments to the Atomic Energy Act. Pollution-control legislation became widespread, as did efforts to prevent child abuse. Laws governing the voluntary and involuntary commitment of addicts were enacted in several jurisdictions.

During the same decade, there was increased international activity in the promulgation and implementation of health and safety laws as a result of programs sponsored by the World Health Organization and meetings and conventions on international harm occasioned by environmental pollution. With regard to the latter, the emphasis was on private liability under international law for harm done by pollution.

HISTORY

In addition to isolation of the sick, principles of hygiene and sanitation were developed by many preliterate peoples. Presumably, whenever these matters were deemed to be of sufficient social concern, they received the coercive sanction of law, as well as the informal sanctions of custom or religion. The preservation of peace and order, the conservation of manpower, and a desire to prevent offenses that may provoke the wrath of the gods against the community have been among the immediate objectives of law that may be included under the general title of public-health and safety laws. They have been the forerunners of what is termed the police power of the modern state.

The history of health and safety laws has been determined in large measure by the influences of religion and science and by political and economic forces. A public concern with the health problems of community life, the control of communicable diseases, and improvement of environmental and sanitary conditions, such as the provision of food and water, medical care, and sewerage and drainage facilities, may be found in varying degrees in primitive, ancient, and modern societies. Moreover, religious and health objectives may be inextricably intermingled, as, for example, in dietary laws, the rules of hygiene, and precautions against contagion specified in the Old Testament, prohibitions against consanguineous marriages, and the merging of functions in priests and shamans. In the Western world there was a close relationship, if not identity, between church and state until the Reformation, and for centuries virtually all literate men were clerics, which meant that most scientific knowledge was in their hands and subject to church control.

Intermingling of health and religious objectives

Early influences. The emphasis of the early church upon spiritual matters and the belief that disease and death are the wages of sin long stood as major obstacles to a positive program for public health and social reform. On the other hand, the church and its religious orders did establish and maintain institutions for relieving travellers, the sick, and widows. Such institutions included that founded by Basil the Great in AD 370 at Caesarea. Pope Leo XIII's encyclical *Rerum Novarum* (1891) and Pope Pius XI's *Quadragesimo Anno* (1931) were most important for the social-reform movement of the 19th and 20th centuries. (Concerned with social and economic reform, the *Rerum Novarum* expressed approval of legislation to protect labour and ensure its receiving a just reward. This position was reaffirmed by the *Quadragesimo Anno*.) It is only since the Industrial Revolution that cleanliness has been officially promoted.

The Middle Ages, about 500 to 1500, saw a decline from the standards of hygiene and sanitation of classical Rome. Although the world is greatly indebted to the Greeks for principles of personal hygiene and to the Romans for the development of public sanitation, there are records of elaborate programs in even more ancient times. Ruins in the Indus Valley and at Harrapa, in the Punjab, reveal that building codes were in effect and that sanitation engineering was far advanced as early as 4000 BC. Egyptian ruins dating from the Middle Kingdom (2100–1700 BC) include bathroom facilities and sewerage systems, as do those of the Incas in the New World. In addition to the formulation of principles of hygiene, the Greeks were responsible for the first attempt to show a causal relation between environmental factors and disease. The book in the Hippocratic collection known as *Airs, Waters and Places* served for over 2,000 years as the basic text on epidemiology and helped to sustain the miasmatic (contaminating-atmosphere) theory of disease until use of the microscope led to the discovery of microbes and the beginning of bacteriology. The Romans saw a relation between swamps and malaria, devised crude respirators to protect workers from dust, in the 2nd century instituted a public medical service, built sewerage systems and public baths, and engineered 14 great aqueducts. They also built the first hospitals and codified and administered health laws. Officials were appointed to maintain the banks of the Tiber, to guard the aqueducts against pollution, to inspect and maintain sewers, to destroy impure foodstuffs, to check weights and measures, and to regulate public baths, brothels, and burial grounds. Justinian I, when one of the worst plagues in history came to Byzantium in 532, set up quarantine posts and required certificates of health for admission to that city.

Contributions of Greeks and Romans

Middle Ages. Upon the disintegration of the Roman Empire, there was a general decline of urban culture and with it an abandonment of public-health measures. Byzantium became the cultural and medical centre for Europe, and from there Greco-Roman medical knowledge spread to the Arab world, while, in the West, health problems were still dealt with largely in terms of magic and religion. During medieval times some of the larger monasteries had proper water supplies and heating and ventilation facilities, but smaller buildings did not; and most medieval cities had a chronic problem in trying to provide sufficient supplies of nonpolluted water and in protecting the public from disease. During the Middle Ages in England the common-law concept of public nuisance provided a means for relief against some flagrant cases of pollution of water supply, as did the common-law crime of malicious mischief; and, in both England and on the Continent, ordinances and statutes were enacted to outlaw some of the most prevalent threats to public health. For example, Douai, France, in 1271, Augsburg, Germany, in 1453, and Rome, in 1468, forbade tanners to wash animal skins on the banks of streams, dyers to pour dye residue therein, and the public from washing clothes at a river that served as a source of water supply. Early

in the 15th century several German cities forbade the construction of hogpens facing the street. In 1185 the streets of Paris were paved to help keep the city clean. London from 1309 had ordinances dealing with the disposal of sewage and offal, and Milan from the 14th century had statutes regulating cesspools and sewers. Most of these ordinances and laws, however, were only sporadically enforced and tended to become dead letters.

Pure-food laws of Middle Ages

The health measures that received the most conscientious enforcement during the Middle Ages were those pertaining to the sale of adulterated or contaminated food and those providing for quarantine in the case of epidemics. Municipal authorities in many places policed the fairs and marketplace to protect purchasers of food. Inspection was practiced, and detailed regulations were enforced. For example, Augsburg in 1276 ordered meat that was not freshly slaughtered to be sold at a special stand, and the Florentines forbade the sale on Monday of meat that had been on sale the previous Saturday.

Although epidemics were regarded by many as a punishment for man's sin, the transmissibility of certain diseases was well-known during the Middle Ages and had been for centuries. The historian Thucydides had vividly portrayed an epidemic during the second year of the Peloponnesian War. Between the plague during the reign of the emperor Justinian (AD 543) and the Black Death (1348), there were a large number of plagues, involving such diseases as leprosy, bubonic plague, smallpox, diphtheria, measles, influenza, sweating sickness, and lesser diseases. The medieval urban dweller lived in constant fear of epidemics, and, when leprosy reached serious proportions in the 13th and 14th centuries, quarantine laws were rigidly enforced. At the beginning of the 13th century, in France alone there were about 2,000 leprosariums, and in all of Europe the number probably exceeded 19,000, although many sick persons were admitted who did not have leprosy.

Regulations for leprosy and bubonic plague

The third Lateran Council, in 1179, promulgated detailed rules to govern the commitment of lepers. The experience with lepers was carried over when the Black Death (bubonic plague) killed thousands of workers in the 14th century. Persons suspected of having the disease were required to report to local authorities, who examined them; if the patients were found to have the plague, they were isolated, as were all those who came into contact with them; food and necessities were provided; the dead with their personal effects were buried outside the city, and the place where they had died was fumigated. Venice, the chief port for trade with the Orient, usually is credited with originating the pattern of quarantine procedure that was in effect during the Middle Ages, and other seaports also established observation stations and isolation hospitals.

In England from the 12th to the 15th century, more than 720 hospitals were established, 217 of which were for lepers. At the beginning of the 14th century, Paris had about 40 hospitals and at least an equal number of leper houses. During the latter part of the Middle Ages, cities and guilds took an active part in founding and maintaining hospitals.

Later history. *Europe.* By the close of the Middle Ages, medicine and public health had fallen under local governmental control. Regulations protecting the water supply from pollution were supplemented by provisions dealing with street cleaning, garbage disposal, and the like, all of which collectively might be called sanitary codes. Physicians and surgeons were required to follow rigid codes and set fee schedules. In addition, writings on nutrition began to appear, and public bathhouses were licensed. When syphilis became widespread, early in the 16th century, the first control measures were directed against prostitutes.

The attempt to deal with public-health problems on a municipal basis was far from successful. Inland cities were all but defenseless against the spread of disease from the seaports. In the larger cities, moreover, the administration of particular ordinances was delegated to various boards, commissions, and officials, a situation that resulted in a crazy quilt of sanitary committees.

Colonial America. Colonial America followed the pattern of utilizing inspection and quarantine for communicable diseases and enacting sanitation legislation. An account by George Percy, John Smith's successor as governor of Virginia, regarding the settlement of Jamestown, Virginia, describes the "cruell diseases, Swellings, Fluxes, Burning Fevers" and famine that afflicted that colony. Almost half of the 100 or more Pilgrims who landed at Plymouth, Massachusetts, in 1620, died within three months after arrival. During the winter before the Pilgrims' arrival, an epidemic had reduced the numbers of Indian warriors of New England from 9,000 to a few hundred. Inoculation was brought to the Colonies by Zabdiel Boylston and Cotton Mather and later was regulated by colonial legislatures. As early as 1647, ships from the West Indies were quarantined in Boston Harbor; and in 1663, during a smallpox epidemic, New York regulated entry into its city by travellers from contaminated regions. Abatement of public nuisances seems to have been in the hands of local authorities throughout most of the 17th century, but a South Carolina statute of 1692 forbade swine running at large in the city of Charleston and directed all persons to cut noisome weeds in and about the lots and streets. During the same year the Massachusetts General Assembly empowered selectmen and justices to assign locations for slaughterhouses and places for the testing of tallow and the currying of leather. Charters granted by William Penn to Philadelphia and Chester, Pennsylvania, in 1701 conferred power upon mayors and magistrates to abate nuisances. Detailed quarantine regulations were enacted in Massachusetts in 1701. The first American health board was appointed in Baltimore in 1793, and by 1797 there were similar boards in New York and Massachusetts.

Early quarantine measures in Colonies

Influence of the Industrial Revolution. The Industrial Revolution occasioned a tremendous increase in urban population, and slum conditions became acute in many manufacturing areas. The first English attempt at slum clearance was in London in the 1760s, when many timbered buildings were pulled down and replaced by brick structures, and streets were paved, drained, and lighted. In addition to disease and slum conditions, the urban population faced poverty and frequent unemployment. The Elizabethan Poor Laws imposed on the individual parish the duty to provide relief, including medical relief, for the indigent.

SOCIAL REFORM

By the beginning of the 19th century in England and on the Continent, the interrelationship between poverty, disease, physical environment, and crime had become apparent, and the view that such conditions were neither inevitable nor irremediable was gaining ground. To the Rationalists of the Age of Reason, a scientific approach to social problems would provide the answers. Industrialists such as Sir Robert Peel and Robert Owen and physicians such as Sir John Simon and Thomas Southwood Smith joined the social-reform movement under the leadership of Jeremy Bentham and Edwin Chadwick. Royal commissions investigated and reported, studies were made, statistics were compiled, and, beginning in the 1830s, social legislation was enacted.

The gathering of statistics

An integral part of the social-reform movement was the compilation and publication of statistics that afforded a factual basis for legislative and other investigations and reports that were an essential prelude to legislation and served to create a demand for reform by informing the public of existing conditions. William Farr was perhaps the greatest medical statistician in the 19th century, but the father of political arithmetic was William Petty, a 17th-century physician, scientist, and economist, who urged the collection of numerical data in such areas as population, education, diseases, and revenue. Petty's friend John Graunt wrote the first important study of health statistics, in 1662. In Germany the philosopher Gottfried Wilhelm Leibniz recommended statistical investigations of health problems and in the 1680s published several essays on the urgent need for

vital statistics. Leibniz' work was carried on and furthered by his contemporary Veit Ludwig von Seckendorff, who is credited with having developed for Germany the public-health philosophy that later was systematized by Johann Peter Frank and enacted into law by Bismarck. Sweden was the first country, however, to require nationwide reporting of vital statistics, legislation having been passed in 1748, in accord with the recommendations of Pehr Elvius, mathematician and secretary of the Swedish Academy of Science.

In England, the first national census took place in 1801. In the United States, Massachusetts was the first state to establish a bureau of labour statistics. The first important study of public health was the 1845 report of John Griscom, city inspector for the New York Board of Health. During the same year the National Institute of Washington undertook a survey of the nation's health. The first state health department was created in Massachusetts in 1869. Several other states followed the Massachusetts example within the next few years.

In the second half of the 20th century, the compilation of vital statistics was regarded as a matter of general public importance. Almost all of the health, safety, and welfare laws that were passed in the 19th and in the first half of the 20th century were based on factual studies and statistical reports.

EARLY LEGISLATION

Europe. There were but few examples of social or health legislation on a national scale before the 19th century. In order to check the high rate of infant mortality and for reasons of public morality the British Parliament passed a series of gin acts, which, in an act of 1751, gave control of licensing to magistrates; and perhaps the first example of social insurance was an act of 1757, "for the relief of coal-heavers working upon the river Thames," that required the employer to deduct sums from employees' wages for a fund from which benefits would be paid in case of sickness, invalidity, old age, or death. Because of its abuse by employers, however, this early insurance plan was abolished in 1770; it was revived in 1792, when a similar act was passed providing for trustees to manage the fund. The Health and Morals of Apprentices Act, sponsored by Sir Robert Peel, was enacted in 1802 to improve the conditions of child labour in the cotton mills. Although this legislation was largely ineffective, it did establish the government's interest in industrial conditions and led to agitation for child-labour laws in other industries and countries.

A reformer, Edwin Chadwick, who was convinced that people were poor because they were sick, was a prominent figure in both the commission that reported on the Poor Laws in 1834 and the commission that reported on public health in 1844 and 1845. As secretary of the poor-law commission he wrote the famous *Report . . . on an Inquiry Into the Sanitary Conditions of the Labouring Population of Great Britain*, which was published in 1842. These reports eventually were embodied in legislation and established national supervision of health, safety, and social problems. The Factory Act of 1833 was the immediate forerunner of modern industrial legislation, and the Liverpool Sanitary Act of 1846 provided for health officers and borough engineers for that city. The Nuisances Removal and Diseases Prevention Act and the Baths and Washhouses Act, passed the same year, were preludes to the Public Health Act of 1848. The first general housing acts were passed under the sponsorship of Lord Shaftesbury in 1851.

19th-century social legislation in England

Movements similar to that which brought about the rise of social legislation in England gained momentum in Germany and France. Johann Peter Frank, German pioneer in public health, ambitiously undertook to systematize all existing knowledge on the subject and to devise detailed codes of hygiene for enactment. Frank, because of his broad scholarship, was one of the first to urge international regulation of health problems and creation of a national health authority to coordinate matters within the country. A proposal made at this time was the creation of "medical police" to make and enforce health and safety regulations. In France, during the Revolution, the Constituent Assembly, on the motion of Joseph-Ignace Guillotin, created a Comité de Salubrité (Health Committee). In 1793 and 1794 a national system of social assistance, including medical care, was passed. Both France and Germany thus became committed to the proposition that government has a positive duty to promote the health, safety, and welfare of workers. In 1883–84, when Bismarck's national social insurance program for workers was enacted (including sickness insurance, industrial-accident insurance, maternity-care benefits, and benefits or pensions for disability, old age, and death), Germany provided the inspiration and established the model for subsequent social legislation elsewhere in Europe and in America.

United States. In the United States the same patterns that led to English legislation are also evident, though English social legislation generally antedated its U.S. counterpart by at least a generation, and the U.S. doctrine of judicial review slowed down development. Early in the 19th century there was considerable support for maximum-hour laws to protect women and for restrictions upon child labour. A few industrial states passed such laws, but there was no effective enforcement because of fear of competitive disadvantage. Massachusetts, for example, in 1842 enacted a child-labour law, and in 1848 Pennsylvania prohibited the labour of children under the age of 12 in certain mills and limited the hours of work to ten unless there was a special contract. The latter provision was nullified when special contracts became customary, and, since in any event the statutory sanction was only a $50 fine, the statute had little effect. As early as the 1830s some industrialists suggested national legislation to regulate child labour and hours of work. Until after the U.S. Civil War, state social legislation was regarded as constitutional, and a judicial position that social legislation violates the due-process clause or impairs the obligations of contracts did not gain currency until about the 1890s.

Early child-labour laws in United States

Congress enacted only a few pieces of social legislation before the 20th century. In 1798 a federal plan of state insurance for disabled seamen based on a 20-cent-per-month deduction from wages was established, and in 1870 the deduction was increased to 40 cents. In 1878 the National Quarantine Act was enacted; it empowered the surgeon general to enforce port quarantines. In the following year the National Board of Health was established.

RISE OF INDUSTRIAL MEDICINE

In the past, most regulation on behalf of the worker has been an expression of religious conviction, such as the duties imposed upon the master in Deuteronomy, or a reflection of the ancient concept that a ruler is supposed to be a father to his people. During the Middle Ages the craft guilds regulated working conditions in detail. Rothari, king of the Lombards, in 643 ordered that bodily harm caused by accidents to builder's labourers should be compensated. The Ordinance of Labourers, in 1349, and the Statute of Labourers, in 1351, enacted in order to negate the economic bargaining power of the survivors of the Black Death, impaired the mobility of the workers, as did the settlement laws that were part of the Poor Laws after 1601. Since workers in England were not free to travel about to seek other jobs, humanitarian impulses alone furnished an incentive to improve their lot. Of greater direct importance were decrees issued by Queen Elizabeth I in England and Jean-Baptiste Colbert and Sébastien Le Prestre de Vauban in France to regulate working conditions.

14th-century regulation of labourers

Industrial medicine, the enactment of factory laws and safety laws, and, eventually, the expansion of workmen's compensation acts to include occupational diseases in their coverage were a direct application of the new Enlightenment. The founder of industrial medicine was Bernardino Ramazzini, professor at Modena and Padua, Italy, who in 1700 published his book *De morbis artificum diatriba* ("Treatise on the Diseases of Artisans"), which examined the relation between disease and pov-

erty. Before and after Ramazzini there were reports on industrial health, such as a pamphlet by one Ulrich Ellenbog in 1473 "on the poisonous wicked fumes and smoke" that afflicted goldsmiths, a treatise on the diseases of miners by Georgius Agricola in 1556, one by Percivall Pott in 1775 on the diseases of chimney sweeps, and a comprehensive study of industrial disease and poverty by Charles Turner Thackrah in 1831. The latter study was relied upon by social reformers.

Liability in industrial accidents. Common-law principles in England and the United States generally placed the cost of industrial accidents on the injured worker in the vast number of cases. To recover in a suit against his employer, the employee had to prove that the master either was personally at fault or had violated a nondelegable duty. Moreover, the injured worker had to overcome the defenses of contributory negligence, the legal presumption that the worker knew and had assumed the risks of his work, and the fellow-servants rule, which meant that if the injured man had himself been careless for his own safety or knew or should have known of the danger or was hurt because one of his fellow workers was negligent, then the employer was not liable for damages. Although the courts recognized that certain nondelegable duties were assumed by the employer, such as an obligation to provide a reasonably safe place to work and reasonably safe machines, tools, and appliances, kept in a reasonably good state of repair, and the issuing of safety rules and warnings of dangerous conditions, nonetheless a violation of such duties did not result in employer liability if one of the common-law defenses was applicable. Most significant, the common-law defenses frequently barred employer liability even though the employer had violated a health or safety statute or ordinance, as when the employee knew of the existing conditions and was said to have assumed the risk.

Difficulty in obtaining compensation for injury

In England in 1880 Parliament enacted the Employers' Liability Act, which modified the defense accorded by the fellow-servants rule but left untouched the defenses of contributory negligence and assumption of risk and required that there be negligence by the employer or a supervisory employee for the worker to maintain suit against his employer. This act was ineffective because English courts construed it so as to permit employees to contract away their rights under the act. In other words, employees were forced to assume the risks of their work as a condition of their employment. German compensation-insurance laws enacted under the sponsorship of Bismarck in 1884 set the pattern for future legislation, and Parliament's Workmen's Compensation Act of 1897 was modelled on its provisions, as were the later United States statutes (see below).

In the United States before 1880, five states had enacted statutes that made railroads liable to employees upon the same basis as they were liable to strangers. Georgia enacted such a law in 1855 and Montana Territory in 1873.

Between 1885 and 1910 most states enacted some kind of employer-liability law, either trying to abrogate some of the common-law defenses or to impose safety inspection and regulation upon hazardous industries. In 1906 Congress enacted the first federal Employers' Liability Act, which covered railroad workers, but that act was held unconstitutional because employees not engaged in interstate commerce were included within its purview. The second Employers' Liability Act was passed by Congress in 1908 to remedy this defect and subsequently was upheld as constitutional. In effect, these railroad statutes abrogated the common-law defenses but required the existence of negligence for recovery. For example, contributory negligence by the employee diminished the amount of recovery had against his negligent employer, and assumption of risk remained a good defense until abolished by amendment in 1939.

Employers' Liability acts of 1906 and 1908

Workmen's compensation. In the United States, despite temporary setbacks and, for a time, strict judicial construction, workmen's compensation acts came to be enacted by each state. Additional congressional legislation covering occupations related to federal commerce, such as those of stevedores and longshoremen, were passed. All but a few states came to include occupational diseases, as well as accidents, as a basis for recovery when either arose out of and in the course of employment. Also, such acts have been extended beyond so-called hazardous industries to include most industrial and commercial enterprises.

PUBLIC WELFARE

Public assistance, unemployment compensation, and social insurance of various kinds became the subjects of national legislation in Europe in the 19th century and in the United States in the 20th. Until Poor Laws were enacted in England at the start of the 17th century, care of the indigent usually had been a matter for church charity or mutual aid by people of the same occupation. The Jews made and enforced regulations for the relief of the poor and regarded charity as a duty. Charlemagne ordered his counts to care for the poor at their own expense and made them *advocati* ("protectors") of the distressed. This action marked the beginning of secularization and regulation of charity in Europe; gradually, poor relief was abandoned as a general function of the church and was assumed by the parish. From ancient times people in the same social class or occupation banded together in friendly societies that arranged burials, financial help where needed, and benefits in case of sickness or widowhood. In time these societies expanded beyond a particular calling, and in England at the beginning of the 19th century they had a membership of 1,000,000 out of a total population of 9,000,000. They were almost as numerous in Holland and Belgium. Neither public nor private assistance, however, was adequate for the problems created by great depressions and chronic unemployment, and a national responsibility for the care of the sick and the poor was repeatedly urged. Sir Thomas More in his *Utopia*, which appeared at the beginning of the 16th century, had advocated public-health measures, social insurance, social security, and public housing. "Freeborn John" Lilburne, pamphleteer for the so-called Leveller movement during the Puritan revolution, called for a comprehensive program of government action to cope with social problems. Daniel Defoe, in his *Essay Upon Projects*, at the end of the 17th century proposed many reforms, including social insurance; and Denis Diderot, in his *Encyclopédie*, in the article on hospital, outlined a public-assistance scheme, including old-age insurance and medical care.

Mutual benefit societies

England. England finally eradicated the last traces of its Elizabethan Poor Laws in 1948 by the enactment of the National Assistance Act, which eliminated the old concepts of parish responsibility, primary family responsibility, and rules as to settlement and removal. Even before World War I, the National Health Insurance Act of 1911 had introduced the principle of state responsibility for the payment of sickness benefits to workers during illness and for the provision of some forms of medical care on a basis of compulsory insurance. Near the end of World War II and immediately after its close, Parliament also passed the National Insurance Act (1944), the Family Allowances Act (1945), the National Health Service Act (1946), and the National Insurance Act (1946). A report by Sir William Beveridge in 1942, experience with prior legislation, and postwar conditions in general combined to provide a favourable political climate for a comprehensive national welfare system.

United States. *Federal legislation.* Although the Department of Health, Education, and Welfare was created in 1953, most social legislation continued to be based on federal–state cooperation rather than on national action alone. For example, such acts as the Sheppard-Towner Act of 1921 (Maternity and Infancy Act), the National School Lunch Act of 1946, and Title V of the Social Security Act are based on federal–state cooperation, the federal government extending grants-in-aid and setting standards and the states administering the programs. There has been a steady trend toward adoption of the insurance principle as a basis for social legislation.

Federal–state cooperation

It was not until the Depression of the 1930s that public assistance, unemployment compensation, social-security benefits, and an integrated program of welfare legislation had been enacted for the nation as a whole. The Social Security Act of 1935 and the Fair Labor Standards Act of 1938 were perhaps the most significant pieces of social legislation enacted by Congress under its authority to regulate interstate commerce and matters affecting such commerce.

Beginning in 1937 the Supreme Court upheld the validity of such legislation, overruling prior decisions that had held that Congress has no authority to abolish child labour or to regulate wages and hours. The separation-of-powers doctrine was viewed by the court as exacting due deference for the legislative and executive judgment and determination of public policy, and the court declined to substitute its economic and social predilections for those of Congress. Most important, certain powers reserved to the states or to the people under the 10th Amendment were held to be no limitation whatsoever upon federal power granted by the Constitution. The court also permitted state legislatures to experiment with social legislation unless a federal statute or power had pre-empted the field. Sociological jurisprudence had replaced the laissez-faire commitment that had dominated the Supreme Court from the 1890s to 1937.

Although health, safety, and welfare legislation, whether enacted by Congress or by the states, was, in the second half of the 20th century, regarded as a legitimate exercise of the police power of government, there yet remained many issues as to substance and procedure that might arise under such laws. The principal limitation upon an attempted exercise of federal power was that some basis therefore must be found in the Constitution.

Powers of local boards

States' powers. State courts determine the constitutionality of state legislation under state constitutions. Historically, health regulations have presented numerous issues, including problems as to the delegation of authority by the legislature to health officials, the specification of sufficient standards to guide administrators so that an arbitrary discretion is not assigned, the reasonableness of rules and regulations promulgated by officials, and the reasonableness of legislative classification. Usually, courts have not reviewed the judgment of health officials with legal authority to exercise discretion, and private persons have not ordinarily been able to hold officials personally liable for action taken in good faith within the scope of their authority.

State health officials utilize many methods of regulation, including inspection, destruction of private property to prevent the spread of contagion or to abate a nuisance, closing of public places, disinfection and sterilization, vaccination, and quarantine. Compliance with orders may be sought by injunctions or the imposition of fines. Some states do not require health officials to give notice and to provide a hearing before taking summary action if the statute does not specify such procedure, a condition that places a heavy burden on the citizen who cares to challenge the validity of the action in a subsequent legal proceeding. In 1967 the United States Supreme Court, reversing a prior decision, held that health inspectors must first obtain a warrant before forcing their way into a private dwelling or place of business. The state may act under either its power of eminent domain or its police power. Generally, if it exercises eminent domain, just compensation is due; such is not the case for an expression of police power. A person who negligently or deliberately infects another with a contagious disease may be held accountable for damages, at least if there was no consent to such contact.

LEGISLATION

Food and drug laws. One of the most important areas of health legislation is that pertaining to adulteration. Both Athens and Rome had laws to prevent the adulteration of wine, and England from the time of Henry III—the 13th century—prohibited the adulteration of certain foods. Parliament from time to time passed statutes prohibiting the adulteration of such commodities as tea, cocoa, and beer, and in 1872 it enacted the Adulteration of Food or Drink Act, which had general application to foodstuffs; it provided for inspection and analysis of samples and for a £50 fine for the first violation and six months imprisonment at hard labour for a second offense. After that time, additional statutes were passed to plug loopholes and to correct judicial construction, culminating in the comprehensive Food and Drug Act, 1955.

Laws against adulteration of foods and drugs

In the United States an act prohibiting the adulteration of drugs was passed by Congress in 1848, and in 1890 a similar statute regarding food was passed. Between 1880 and 1906 there were 103 bills introduced in Congress to control interstate traffic in food and drugs, but none were enacted until the Pure Food and Drugs Act of 1906 was signed by Pres. Theodore Roosevelt.

Difficulty was encountered in the enforcement of the act because of the definitions given to such terms as adulteration, the narrow construction given by the courts, and the "distinctive-name" exemption that permitted a manufacturer who adopted a distinctive name for his article to ship what he pleased. Moreover, the sanctions of a maximum fine of $200 for the first offense and $300 or a year's imprisonment or both for subsequent offenses, as well as the seizure provisions, did not prove to be successful deterrents.

Except for minor amendments, the 1906 act and similar state legislation remained in effect until the Federal Food, Drug and Cosmetic Act of 1938 was passed after the tragic death, in 1937, of at least 73 persons who had taken a drug known as Elixir Sulfanilamide. (Existing legislation did not prohibit the distribution of poisonous or dangerous drugs.) The 1938 act, within the limits of the commerce power, prohibits foods dangerous to health and also prohibits foods, drugs, or cosmetics in insanitary or contaminated containers. Enforcement of the act was given to a Food and Drug Administration, and inspection stations were set up in several large industrial cities.

In addition to the 1938 Federal Food, Drug and Cosmetic Act, state legislation was important in covering food and drugs outside the scope of the federal act because of its limitation to interstate commerce. Of jurisprudential significance is the fact that under both the federal law and the statutes of some states it is not essential that there be a specific criminal intent for conviction. It may be enough that the manufacturer or processor intended to ship the articles in question even though he neither knew nor had reason to know that there was a violation.

The international tragedies that occurred in 1961 when thalidomide, taken as a sedative by pregnant women, caused gross malformations in their offspring led to the 1962 amendment of the Food and Drug Act and a 1966 memorandum of the Surgeon General that drastically altered the law as it pertains to clinical investigations and human experimentation. Earlier, the Nuremberg Code on the use of human subjects for medical research had some moral influence, but there was a dearth of law. The change, in effect, added a requirement to the law that subjects or patients must be informed that they are to receive an experimental drug.

Safety laws. The enactment of safety legislation has been contemporaneous with the passage of workmen's compensation acts and public-health laws. In this area, however, education and the activities of voluntary groups have almost completely overshadowed legal developments. There has been an intensive safety movement in the 20th century, and safety codes have been devised by engineers for various industries. At the outset humane industrialists who realized that safe working conditions are an asset to the business, particularly if the factory comes under workmen's compensation laws, promoted accident-prevention campaigns by imposing safety rules in plants. In the United States, cities and states at an early date required inspections of steam-power boilers, and, later, building inspectors, fire marshals, and elevator inspectors exercised municipal authority. Mine safety was a subject for legislation in many coal-producing

states; the Safety Appliances act for railroads, originally passed in 1887 and expanded and amended in 1893, 1903, 1910, and 1920, was the first significant federal statute. The Walsh–Healey Public Contracts Act of 1936 required government contractors to comply with the health and safety laws of the state in which the contract is being fulfilled. During and since World War II it has been one of the most important statutes because it reinforces state laws that otherwise might be ignored. For the most part, state safety laws have been strictly enforced only in a few industrial states that have provided a sufficient number of inspectors. The United States Department of Labor and, since 1953, the Department of Health, Education, and Welfare are both concerned with certain aspects of the general problem of industrial safety. Federal Coal Mine Inspection acts, passed in 1941 and 1952, require inspection and reports, empower the director of the Bureau of Mines to forbid miners to enter dangerous portions of a mine, and specify detailed rules concerning roof support, ventilation, equipment, fire protection, and other safety matters.

Walsh–Healey Public Contracts Act

Before workmen's compensation acts were securely established as constitutional around 1920, most state safety acts relied upon inspection and indirect enforcement through civil litigation in damage suits brought by injured workmen. After 1920, when damage suits all but ceased because most workmen were limited to compensation claims, administrative boards with power to make rules, investigations, and reports and to seek injunctions became the mode of enforcement. In a few states—*e.g.*, Washington—educational standards of safety, as well as rules dealing with safe working conditions, were, in the early 1960s, issued by the supervisor of safety. Statutes that provide for apprenticeship systems, limit the hours of work, or require rest periods may be classified as safety laws because of their bearing on accident prevention. The constitutionality of safety laws in the second half of the 20th century was usually upheld unless there was a showing of arbitrary action in a particular situation, and such regulations typically were enforced by inspection plus criminal or administrative sanctions. In addition to laws pertaining to industrial safety, there are statutes concerning regulation of traffic; inspection of theatres, schools, and other public places; highways and sidewalks; public carriers; bathing beaches; and fire prevention.

INTERNATIONAL EFFORTS

International concern about health and safety has been expressed in the writings of public-health pioneers such as Johann Peter Frank; in international conventions, such as that held in Paris in 1851; and in health organizations within the framework of the League of Nations and the United Nations (UN). The preamble to the constitution of the World Health Organization (WHO) provides that "an informed opinion and active co-operation on the part of the public are of the utmost importance in the improvement of the health of the people"; it further declares that every individual has a fundamental right to the highest attainable standard of health. The World Health Organization is committed to the principle of mutual aid in dealing with social and health problems, and its work is supplemented by and correlated with the activities of such other United Nations agencies as the United Nations Children's Fund, the Food and Agriculture Organization, the International Labour Organisation, and the United Nations Educational, Scientific and Cultural Organization (UNESCO).

Private organizations have also laboured in the field of international health and safety problems. The International Association of Industrial Accident Boards and Commissions has promoted safety codes, and the Rockefeller Foundation has sponsored major international health work. Numerous international conferences have been held to promote health and safety. Mohammed Ali, ruler of Egypt, called a conference in 1833 to discuss quarantines and international hygiene; the International Association for the Legal Protection of the Poor was organized in Paris in 1900; and the International Labour office was organized at Basel, Switzerland, in 1901. The Bern (Switzerland) conferences of 1905, 1906, and 1913 drew up international conventions prohibiting the use of white phosphorus in the manufacture of matches and the employment of women or of children under 16 on night work and limiting their hours of work. President Truman's Point Four program included technical assistance on health matters, and many underdeveloped nations received help from the United States and other governments in an effort to eliminate various preventable diseases that had been eradicated in Europe and the United States in the 19th century. Other nongovernmental agencies prominent in the international health and welfare fields include the International Red Cross and League of Red Cross Societies, the Save the Children Fund, and the Society of Friends.

International conferences

International trade in radioactive materials led to treaty regulations governing consignments. The first convention on the subject, signed by 16 nations belonging to the Organization for European Economic Co-operation (later the Organisation for Economic Co-operation and Development) in 1959, permitted the level of financial protection to be set by any nation at a minimum amount of $5,000,000. In 1961 the International Atomic Energy Agency, composed of 81 nations, issued Regulations for the Safe Transport of Radioactive Materials, and these were amended in 1964. The basic purpose of the regulations was to control and limit the risks of irradiation and radioactive contamination that radioactive materials may present in the course of transport; there are specifications as to shielding, distance and duration of exposure, proper containment, adequate labelling and marking, and precautions against a fissile chain reaction.

The Brussels convention of 1962 established licensing for nuclear-propelled ships, and the Inter-American Nuclear Energy Commission later promulgated regulations for the Pan-American trade in radioactive materials. Also in the 1960s the first steps were taken in Europe toward regional cooperation in protecting the human environment, and meetings were held between the United States and Canada to abate and control transboundary pollution. A specific matter of international concern was the conference called by the Inter-Governmental Maritime Consultative Organization (IMCO), which met at Brussels in 1969 and prepared international conventions relating to oil-pollution casualties on the high seas and civil liability for such damages. In these and other meetings, there was increasing reliance upon the sanction of imposing liability under international law for environmental damage caused by industry and commerce.

RECENT DEVELOPMENTS

During the late 1960s and early 1970s, widespread concern about radiation hazards, air and water pollution, and poisons used in control of plant and animal pests led to legislation in these areas.

In the Union of Soviet Socialist Republics, for example, a law enacted in 1969 established safety standards with respect to production, processing, application, storage, and transport of natural or artificial radioactive substances and of other sources of ionizing radiation (with the exception of ultraviolet radiation). In the United Kingdom, legislation concerning radiation included establishment of a National Radiological Protection Board. In Belgium a law was enacted against assigning workers declared unfit for exposure to ionizing radiations to jobs involving such exposures. In the United States the Radiation Control for Health and Safety Act of 1968 authorized the formulation of standards to control the emission of radiation by electronic products and compelled compliance with these standards by manufacturers who produce such devices and ship them in interstate or foreign commerce. Moreover, the Bureau of Radiological Health of the Food and Drug Administration promulgated regulations covering performance standards for television sets and medical X-ray equipment. Other countries that enacted laws on radiation hazards included Canada, Iceland, the Republic of the Philippines, Ceylon, the German Democratic Republic, Hungary, and Italy.

U.S.S.R. law on radioactive materials

Laws designed to prevent pollution of water or air were enacted in many countries, including Canada, Denmark, France, the Dominican Republic, and Malta. The Bulgarian measure, for example, enacted in April 1969, prescribes that projects for the utilization of water must include safeguards against contamination of surface water and groundwater, must protect mineral and medicinal waters from contamination, and must not do injury to agriculture, forestry, or fisheries. In the United States the 1955 Air Pollution Control Act had left the responsibility of pollution control to the state and local governments, and the Clear Air Act of 1963 had repeated that states' right principle but had authorized interstate compacts, such as that established for New York and New Jersey. In 1967 the federal government intervened more actively into the air-pollution problem by the enactment of the 1967 Air Quality Act, which directed the Department of Health, Education, and Welfare to establish air-quality-control regions throughout the country, to devise standards, and to notify affected areas. By 1970 some 40 states had enacted air-pollution statutes, usually setting up boards or commissions to establish air quality and emission standards. The federal Environmental Protection Agency also was established to enforce and implement the Clean Air Act of 1970.

Fear of contamination of food and fodder by pesticides elicited laws on the subject in a number of countries. In Algeria such a law was specifically directed at the use of mercury compounds on plants. The Australian law concerned a wider range of substances—insecticides, fungicides, herbicides, and vermin destroyers.

Other individual health measures established during this period included acts concerned with food hygiene, control of communicable disease, training of health workers, prevention or control of drug dependence, and control of the manufacture of pharmaceutical products and the testing of new drugs. In 1966 in the United States, for example, a regulation was established by the Food and Drug Administration that directed physicians, except in unusual circumstances, to obtain the written consent of patients before using investigational drugs.

National Traffic and Motor Vehicle Safety Act

One piece of controversial safety legislation enacted during this period was the United States National Traffic and Motor Vehicle Safety Act of 1966, which set mandatory standards for all new cars beginning with the 1968 models and gave the Secretary of Commerce the responsibility of carrying out research, testing, development, and training as to safety features in automobiles. The manufacture or sale of substandard motor vehicles or equipment was proscribed. The act applied, however, only to new cars and equipment.

In December 1969 a fundamental health code was enacted in the Union of Soviet Socialist Republics. The first of the nine main divisions of this law is concerned with the principles of organization of public health in the Soviet Union and with the jurisdictions of the Union and of its constituent republics in the field of health. Other main divisions of the act deal with (1) the practice of medicine and of pharmacology; (2) providing protection to the people with respect to sanitation and the control of communicable disease; (3) therapeutic and prophylactic care; (4) maternal and child health; (5) treatment at sanatoriums and health resorts and organization of leisure activities, tourism, and physical culture; (6) expert medical evaluation of working capacity and of matters concerning forensic medicine and forensic psychiatry; and (7) supplying of drugs and of prosthetic devices. The final section of the law specifies that, when provisions of treaties into which the Soviet Union has entered conflict with Soviet law or the laws of the constitutional republics of the Soviet Union, the provisions of the treaties shall take precedence.

BIBLIOGRAPHY

Public health in general: GEORGE ROSEN, *A History of Public Health* (1958), a classic volume on the subject that has been extensively used and widely quoted; W.M. FRAZER, *A History of English Public Health*, 1834–1939 (1950), the standard English work on the public-health movement of the 19th century; FRANK P. GRAD, *Public Health Law Manual* (1965), a handbook on the legal aspects of public-health administration, principally designed for health officers and administrators; D.W. CLARK and B. MACMAHON (eds.), *Preventive Medicine* (1967), a textbook with chapters written by 39 authors—lawyers, doctors, and health administrators; COLIN FRASER BROCKINGTON, *World Health*, 2nd ed. (1968), both a history and a synthesis of current views on health questions; *Public Health in the Nineteenth Century* (1965); E. HAROLD HINMAN, *World Eradication of Infectious Diseases* (1966), a monograph that considers the origins of eradication efforts, the nature of such programs, their progress to date, and the future outlook for their success; W. HOBSON (ed.), *The Theory and Practice of Public Health*, 3rd ed. (1969), a textbook designed to cover the community aspects of medicine and to give basic information to those working in the field; L.S. GOERKE and E.L. STEBBINS, *Mustard's Introduction to Public Health*, 5th ed. (1968), a textbook designed to introduce students to concepts and principles of public health and to contemporary modes of community health organization and practice.

Health care: ANNE R. SOMERS, *Health Care in Transition: Directions for the Future* (1971), a significant survey of the trends in health care and the direction it is taking; *Hospital Regulation: The Dilemma of Public Policy* (1969); NATHAN HERSHEY, "Compulsory Personal Health Measure Legislation," *Public Health Report 341* (1969), a discussion of a variety of legislative enactments (of four states), including a review of a number of court decisions.

Special health and safety problems: PETER D. LOWES, *The Genesis of International Narcotics Control* (1966); RUTH FOX (ed.), *Alcoholism: Behavioral Research, Therapeutic Approaches* (1967); LAWRENCE LADER, *Abortion* (1966); "Cigarette Advertising and the Public Health," *Columbia Journal of Law and Social Problems*, 6:99 (1970); "Automobile Design Liability; Larsen v. General Motors (391 f 2d 495) and Its Aftermath," *University of Pennsylvania Law Review*, 118: 299–312 (1969).

Food, drugs, and cosmetics: Food, Drug, Cosmetic Law Journal, a monthly publication covering statutes, decisions, and developments in this field; H.A. TOULMIN, JR., *A Treatise on the Law of Foods, Drugs, and Cosmetics*, 2nd ed., 4 vol. (1963), a treatise covering statutes and amendments and their legislative history; RICHARD HARRIS, *The Real Voice* (1964), a journalistic account of politics and the drug industry with emphasis on the Kefauver committee investigation; DAVID CAVERS, "Administering That Ounce of Prevention: New Drugs and Nuclear Reactors," *West Virginia Law Review*, 68:109–135, 233–262 (1966).

Pollution and environmental control: SIDNEY EDELMAN, *Law of Air Pollution Control* (1970), a comprehensive discussion of law up to date of publication; "Symposium: The Environment and the Law," *Hastings Law Journal*, 22:467–830 (1971), see especially J. STEVENS, "Air Pollution and the Federal System: Responses to Felt Necessities"; "Toward a Constitutionally Protected Environment," *Virginia Law Review*, 56:458–486 (1970), an examination of theories that could justify judicial recognition of the constitutional right to healthful environment.

Regulation of medical practice: ROBERT C. DERBYSHIRE, *Medical Licensure and Discipline in the United States* (1969), an analysis of the practical aspects of medical licensure in the United States today; RICHARD H. SHRYOCK, *Medical Licensing in America, 1650–1965* (1967), on the evolution of the medical licensing system as we know it today; IRVING LADIMER and ROGER W. NEWMAN (eds.), *Clinical Investigation in Medicine: Legal, Ethical and Moral Aspects, an Anthology and Bibliography* (1963), a full treatment of legal and ethical problems involved in human experimentation and the use of novel medical procedures; A.M. and B.L. SADLER, "Transplantation and the Law: The Need for Organized Sensitivity," *Georgetown Law Journal*, 57:5–54 (1968).

Protection from radiation: GERALD L. HUTTON, *Legal Considerations on Ionizing Radiation* (1966).

International problems: COLIN FRASER BROCKINGTON, *World Health*, 2nd ed. (1968); WORLD HEALTH ORGANIZATION (WHO), *International Digest of Health Legislation* (quarterly), in English and French; ARTHUR LARSON, *The Law of Workmen's Compensation*, 2 vol. (1952, 1971 suppl.); REED DICKERSON, *Products Liability and the Food Consumer* (1951); WILLIAM L. PROSSER, *Handbook of the Law of Torts*, 4th ed., ch. 17 (1971).

(H.H.F.)

Heat

Heat is energy that is transferred from one body to another as the result of a difference in temperature. If two

bodies at different temperatures are brought together, energy is transferred—*i.e.*, heat flows—from the hotter body to the colder. The effect of this transfer of energy usually, but not always, is an increase in the temperature of the colder body and a decrease in the temperature of the hotter body. A substance may absorb heat without an increase in temperature by changing from one physical state (or phase) to another, as from a solid to a liquid (melting), from a solid to a vapour (sublimation), from a liquid to a vapour (boiling), or from one solid form to another (usually called a crystalline transition; see PHASE CHANGES AND EQUILIBRIA). The important distinction between heat and temperature (heat being a form of energy and temperature a measure of the amount of that energy present in a body) was clarified during the 18th and 19th centuries.

HEAT AS A FORM OF ENERGY

Because all of the many forms of energy, including heat, can be converted into work, amounts of energy are expressed in units of work, such as joules, foot-pounds, kilowatt-hours, or calories. Exact relationships exist between the amounts of heat added to or removed from a body and the magnitude of the effects on the state of the body. The two units of heat most commonly used are the calorie and the British thermal unit (BTU). The calorie (or gram-calorie) is the amount of energy required to raise the temperature of one gram of water from 14.5° to 15.5° C; the BTU is the amount of energy required to raise the temperature of one pound of water from 63° to 64° F. One BTU is approximately 252 calories. Both definitions specify that the temperature changes are to be measured at a constant pressure of one atmosphere, because the amounts of energy involved depend in part on pressure. The calorie used in measuring the energy content of foods is the large calorie, or kilogram-calorie, equal to 1,000 gram-calories.

Heat capacity

In general, the amount of energy required to raise a unit mass of a substance through a specified temperature interval is called the heat capacity or the specific heat of that substance. The quantity of energy necessary to raise the temperature of a body one degree varies depending upon the restraints imposed. If heat is added to a gas confined at constant volume, the amount of heat needed to cause a one-degree temperature rise is less than if the heat is added to the same gas free to expand (as in a cylinder fitted with a movable piston) and so do work. In the first case, all the energy goes into raising the temperature of the gas, but in the second case, the energy not only contributes to the temperature increase of the gas but also provides the energy necessary for the work done by the gas on the piston. Consequently, the specific heat of a substance depends on these conditions. The most commonly determined specific heats are the specific heat at constant volume and the specific heat at constant pressure. Relationships between heat capacities and molecular structure of gases are covered in the article GASEOUS STATE. The heat capacities of many solid elements were shown to be closely related to their atomic weights by the French scientists Pierre-Louis Dulong and Alexis-Thérèse Petit in 1819. The so-called law of Dulong and Petit was useful in determining the atomic weights of certain metallic elements, but there are many exceptions to it; the deviations were later found to be explainable on the basis of quantum mechanics (*q.v.*).

It is incorrect to speak of the heat in a body, because heat is restricted to energy being transferred. Energy stored in a body is not heat (nor is it work, as work is also energy in transit). It is customary, however, to speak of sensible and latent heat. The latent heat, also called the heat of vaporization, is the amount of energy necessary to change a liquid to a vapour at constant temperature and pressure. The energy required to melt a solid to a liquid is called the heat of fusion, and the heat of sublimation is the energy necessary to change a solid directly to a vapour, these changes also taking place under conditions of constant temperature and pressure.

Air is a mixture of gases and water vapour, and it is possible for the water present in the air to change phase; *i.e.*, it may become liquid (rain) or solid (snow). To distinguish between the energy associated with the phase change (the latent heat) and the energy required for a temperature change, the concept of sensible heat was introduced. In a mixture of water vapour and air, the sensible heat is the energy necessary to produce a particular temperature change excluding any energy required for a phase change.

EARLY THEORIES AND OBSERVATIONS

Despite the fact that early man made use of the effects of heat and surely recognized differences in temperature, it has been only in relatively recent times that man has understood that heat is energy and, in particular, energy being transferred. The earliest evidence of man's use of fire comes from caves occupied by Peking man about 500,000 years ago. Early man's fire served two main purposes: warming the body and giving light during hours of darkness. It is not known whether Peking man was able to make fire himself, but later man produced fire by percussion (flint), friction (two pieces of wood), and even by gas compression, although the details of these phenomena were not understood.

Before the 5th century BC, much progress was made in science and technology in Egypt and Mesopotamia, but heat and temperature remained undefined. Despite the lack of understanding of its nature, heat has been used to produce and work metals suitable for tools since the beginning of civilization. Copper, tin, and lead were the first metals to be smelted from natural ores, probably before 4000 BC. Silver and bronze (an alloy of copper and tin) came into use later (3000–2200 BC), followed by brass in the Roman period (500–50 BC). Iron was used from about 1200 BC. Glass was also produced, with transparent glass being made from about 1500 BC. The Egyptians produced a turquoise-blue glaze that required delicate temperature regulation in firing.

Greek philosophers. Much of early philosophy was directed toward explaining the universe and its origin. Aristotle believed in four elements proposed by earlier philosophers: Earth, dry and cold; Water, moist and cold; Air, moist and hot; Fire, dry and hot—and believed that they were different forms of the same substance. In Aristotle's view each element occupied its own sphere in the universe, and one element might be transformed into another: the formation of steam by boiling water could be described as Water (cold moist) + Fire (hot dry) → Air (hot moist) + Earth (cold dry). Aristotle finally added a fifth element representing the heavens—ether, from a word meaning "to glow."

Aristotle's elements

Philon of Byzantium (*c.* 100 BC) and Heron of Alexandria (*c.* AD 100) studied the flow of liquids and gases. Philon described an instrument for the demonstration of the expansion of air. This device may have been used as a thermometer, one of the earliest known.

Heron invented the first steam turbine as a toy. Water was heated to boiling in a caldron by an external fire. The steam was fed to a pivoted globe having two exhaust jets. As the steam left the jets it caused the globe to rotate.

It is apparent that the study and use of heat had not progressed greatly. As previously stated, Peking man had used fire for warmth and light about 500,000 years ago; the use of fire for smelting copper occurred prior to 4000 BC; Empedocles of Agrigentum proved the corporeality of air (*c.* 450 BC); Philon demonstrated the expansion of air (*c.* 100 BC); and Heron built a toy reaction steam turbine (*c.* 100 AD); yet, the real nature of heat and combustion processes remained unknown, and the use of heat to produce work was not yet achieved.

The alchemists. The alchemists appeared about AD 100–300, seeking ways of turning other metals into gold. The pursuit of methods for changing one substance into another came to be called "khymeia" probably from the Greek word *khymea*, meaning "to fuse or cast a metal." In the 8th century, Arabic scholars turned their attentions to *al khymea*. Alchemists began the first systematic investigation of metals and developed chemical analysis in a limited fashion. A woman alchemist, Maria the Jewess, is believed to have invented the first distillation apparatus in the 1st or 2nd century AD, in Alexandria. The discovery

of a number of new materials constituted the major contribution of the alchemists.

The investigations of Roger Bacon

One notable exception to the alchemists was the English scholar Roger Bacon, who was a proponent of experimental investigations together with the use of mathematics as a method of interpreting nature. Bacon made an extensive study of lenses and mirrors and in 1269 predicted the automobile, the airplane, and motorized ships, but he apparently did not work directly in the area of heat. Alchemy persisted into the 17th century, with chemistry being a medical art or mystical science. The name alchemy began to fall into disrepute, even though some of the mystical concepts of alchemy were retained by more serious investigators, who began to call themselves "chymists."

Early chemists. In 1661 the English natural philosopher Robert Boyle published *The Sceptical Chymist*, a book credited with making chemistry an empirical science. Boyle rejected the Aristotelian concept of the elements and defined an element as a substance that cannot be broken up into two or more apparently simpler substances or built up from simpler substances. Copper was therefore an element, but bronze was not. In 1805 another English scientist, John Dalton, in advancing the modern atomic theory, defined an element by suggesting that all matter is made up of atoms and that each element is composed of atoms of one characteristic type. He worked out the first table of atomic weights based on experimental data. Boyle's contributions were not confined to the recognition of the elements. He studied air at ordinary temperatures and concluded that the pressure of the air was inversely proportional to the volume if the temperature was held constant. In 1787 a French physicist, Jacques-Alexander-César Charles, studied several gases and stated that at a fixed pressure, raising the temperature by one Celsius degree (1.8 Fahrenheit degrees) increases its volume by 1/273 of the volume it occupies at 0° C (32° F). Combined, the laws of Boyle and Charles become the perfect gas law, relating pressure, volume, and temperature of a gas. In 1802 another French chemist, Joseph-Louis Gay-Lussac, showed that air and other available gases (nitrogen, oxygen, hydrogen) "expand equally with the same degree of heat"; *i.e.*, that their coefficients of expansion were the same. This fact, which is true only of perfect gases, was used in the design of gas thermometers.

THERMOMETRY

Galileo's air thermoscope

The concept of temperature as hot or cold was long known, but a method for measuring it accurately was not available during Boyle's life. Galileo is credited with making an air thermoscope that consisted of a glass bulb containing air, connected to a glass tube of small bore, which dipped into a bowl of coloured liquid. When cooled, the air in the bulb contracted, and the liquid moved up the glass tube; when heated, the air expanded and the liquid moved down. The stem was marked in gradations, but the readings were affected by changes in atmospheric pressure as well as temperature. The principle of operation of this instrument was similar to that of Philon. Ferdinand II, grand duke of Tuscany, is said to have developed the first sealed liquid-in-glass thermometer, probably in 1641; the liquid used was alcohol. The idea of representing hot and cold numerically was familiar to physicians. In the 2nd century AD, the Greek physician Galen suggested a temperature scale based on boiling water and ice. Arab and Latin physicians developed a scale of 0–4 degrees of hot or cold, but they had only the senses for measurement. In the 17th century, thermometers based on the expansion and contraction of air or of water were used for measuring body temperature. The temperature of the human body was taken as one of the fixed points on these temperature scales. In 1688, in France, a physicist named Guillaume Amontons proposed to measure temperature by the variations in pressure of a fixed mass of gas contained in a constant volume. He defined absolute zero as the temperature at which the gas pressure goes to zero. He used a tube of mercury to measure pressure variations. In 1701 Sir Isaac Newton suggested that the temperature scale should be zero for the freezing point of water and 12 for the human body. In 1714 a German–Dutch physicist, Gabriel Daniel Fahrenheit, proposed that a freezing mixture of ice and salt should be the zero point and that the human body temperature be made 96 to give a greater number of degrees on the scale. Fahrenheit showed that the freezing point and the boiling point of water at a fixed pressure are constant. In 1724 Fahrenheit suggested that these two fixed points be used for defining the temperature scale (zero was still the ice–salt mixture and human blood 96), and the fixed points were established as 32° F (freezing point) and 212° F (boiling point) on the scale that bears Fahrenheit's name. With these fixed points, blood temperature was found to be 98.6° F. Fahrenheit greatly improved the accuracy of the thermometer by using mercury as the fluid in a sealed glass tube. The centigrade ("hundred step")—now called the Celsius—scale was suggested by a Swedish astronomer, Anders Celsius, in 1742. On this scale, the freezing point of water is zero and the boiling point of water is 100°.

All the thermometers mentioned so far depend upon the fact that gases and liquids generally tend to expand upon heating. Other properties that change with temperature can be employed in the construction of thermometers. Electrical resistivity is such a property, used in platinum resistance thermometers for accurate determinations of temperatures between −183° and 630° C (−297° and 1,170° F). The emission of electromagnetic radiation from hot objects is the basis of the optical pyrometer, a device used for measuring high temperatures (above about 1,000° C).

Modern thermometers

The German physicist Thomas Johann Seebeck discovered in 1821 that an electromotive force exists in a circuit composed of two dissimilar metals when the junctions of the metals are at different temperatures. Such devices, called thermocouples, have been thoroughly developed for use in measuring temperatures and are the basis of an internationally recognized scale of temperatures between 630.5° and 1,060° C (1,167° and 1,940° F; see THERMOMETRY.)

DEVELOPMENT OF THE PRINCIPLES OF THERMODYNAMICS

Although other sciences had made considerable progress as early as the 7th century, physics and chemistry, which form the foundations of thermodynamics (the study of the relations between heat and other forms of energy), were still in a primitive stage. Jan Baptista van Helmont, a Flemish physician and alchemist, performed the following experiment: he watered the shoot of a young willow tree with rain for a period of five years; he then weighed the shoot and concluded that the increase in weight was due to water alone, since that was all he had fed it. It appeared to him that water could transform itself into wood and, hence, into fire and ashes. Van Helmont also believed that water could be turned into metal with the help of a "seminal spirit" or alkahest. In his studies on combustion and fermentation, van Helmont was led to make a distinction between air and the products of these processes, which he called "gas" after the Greek word *chaos*, although he still used the name *spiritus sylvestris* as well. He made a distinction between flammable and nonflammable gases. He called the gas from burning charcoal gas "carbonum" (carbon dioxide or carbon monoxide). Only the names gas and alkahest were adopted by his contemporaries. During the 17th century, the chymists were primarily interested in studies involving the use of salts as part of the search for the universal solvent. They made experiments using black powder (antimony sulfide) or nitre (potassium nitrate) with combustible substances. The influence of these studies seems to have led to the belief that nitre was the universal solvent and that combustion and respiration were due to the actions of a "nitro-aerial spirit." Although the English physicist Robert Hooke adhered to this barren concept, he showed a better understanding when, in 1665, he wrote, "Heat is a property of a body arising from the motion or agitation of its parts; and therefore whatever body is thereby touched must necessarily receive some

part of that motion, whereby its parts will be shaken." This idea is similar to that which later formed the basis of the kinetic theory. Boyle was one of the first experimentalists to examine the role of air in combustion; he invented a pneumatic pump, which he and Hooke constructed and with which they were able to obtain a vacuum in which a flame was extinguished and no animal could survive. Boyle believed that there was a special agent called volatile nitre in the air that gave air its special qualities.

Phlogiston theory. The studies of chemists continued in the field of reactions of acids and bases, in particular, and led to the corpuscular concept of matter—*i.e.*, that matter is composed of indivisible and immutable particles. In 1697 a German physician and chemist, Georg Ernst Stahl, proposed the phlogiston theory. Stahl was influenced by the work of another German physician-chemist, Johann Joachim Becher, who had proposed earlier that all minerals are made up of three components in various amounts—*terra pinguis* (sulfur), *terra mercurialis* (mercury), and *terra lapida* (saline constituent)—and that all natural substances contained *terra pinguis*, which was lost on combustion. It was this concept that Stahl built upon: he postulated a weightless element, phlogiston, present in all combustible material. Substances that burned well were rich in phlogiston, and those that did not burn did not because they contained no phlogiston. Stahl explained rust in much the same way: the greater the amount of phlogiston in a given metal, the more readily it will react to form the metal oxide. Stahl visualized the phlogiston particles as escaping in the presence of air. Most interesting was Stahl's explanation of why an increase in weight occurred on the formation of a metal oxide although the phlogiston had left the metal, not entered it. Stahl argued that phlogiston was weightless and that the metals, in losing the phlogiston, became more concentrated and dense. Later, he suggested that the weight increased because air filled the vacuum left by the escaping phlogiston. The phlogiston theory was accepted universally and became not only the explanation of combustion and oxidation in general but also the principle describing the chemical and physical properties of all bodies.

The caloric theory. In 1760, 63 years after Stahl proposed the phlogiston theory, Joseph Black, a Scottish chemist, introduced the caloric theory. He stated ". . . heat is evidently not passive; it is an expansive fluid, which dilates in consequence of the repulsion subsisting among its own particles." Black considered this fluid, caloric, to be indestructible and believed that when a body was heated, it accumulated caloric, and when the body cooled, it lost caloric. Black defined the unit of quantity of heat as the amount of caloric fluid needed to raise the temperature of a given body, a unit of temperature or the amount of caloric lost when cooled one degree of temperature. This quantity of heat was called the "capacity for heat" of the body. Black also introduced the concept of latent heat, the amount of energy needed to melt ice at constant pressure and temperature. He demonstrated that the melting of a fixed mass of ice requires a constant quantity of heat. His contribution was an important one because he distinguished between heat and temperature. The caloric theory was believed as late as 1824. Radiant heat was also studied but was not readily adapted to the caloric theory and so was thought to be related to light as invisible light. The notion of phlogiston died out after 1783, when Antoine-Laurent Lavoisier explained the essentials of respiration and combustion. The concept of electricity as a fluid lasted the longest, with Benjamin Franklin's support. What was still missing was a concept of energy that would include heat, light, magnetism, electricity, chemical energy, and work. The conservation of matter had been suggested by Newton's views, and Lavoisier assumed it for all chemical reactions: "nothing is created in the operations either of art or of nature, and it can be taken as an axiom that in every operation an equal quantity of matter exists both before and after the operation." The principle of the conservation of energy was longer in coming. A Dutch physicist and astronomer, Christiaan Huygens, developed the concept of kinetic energy. That mechanical energy is always conserved was proved by several workers. Hermann von Helmholtz, a German physiologist and physicist, in 1847 showed that the sum of the kinetic and potential energies in a system of particles is constant. In 1843 an English physicist, James Prescott Joule, calculated the numerical relationship between a quantity of electricity and the quantity of heat produced by it.

Heat as a fluid

Conservation of energy

Gases and combustion. Important to the progress in understanding combustion was an understanding of gases. In 1772 Daniel Rutherford, a Scottish chemist, formed "fixed air" and "phlogisticated air" by burning a candle in a closed container until the candle went out. He removed the fixed air by bubbling the air through a solution of lime and noted that the portion of air that remained was essentially the same as the original air in its physical properties but that it would not support combustion. This "phlogisticated air" was nitrogen, although Rutherford did not realize it. The English chemist Joseph Priestley discovered oxygen in 1774: he put mercury calx (mercuric oxide) in a glass container and heated it with sunlight, using a lens. Mercury was formed in the upper portions of the glass container. Priestley thrust a smoldering splinter of wood into the container, and it burned brilliantly, more so than in ordinary air. Priestley called this gas "dephlogisticated air"; Lavoisier named it "vital air" or "oxygène."

Lavoisier and another French scientist, Pierre-Simon, marquis de Laplace, repeated earlier work and stated that water was composed of equal weights of inflammable air and vital air. Lavoisier had already heard of Priestley's results and was led to heat mercury in a limited supply of air. The mercury and container were covered with red spots and scales of mercuric oxide, and the original volume of air had decreased, leaving what was called mofette (mephitic air, mainly nitrogen). The mercuric oxide was collected and heated, and the volume of air was recovered as vital air or oxygen. Lavoisier concluded that air was a mixture of vital air and mephitic air and that phlogiston was not necessary in combustion, because combustion was a chemical combination with vital air. In 1756 a Russian chemist, Mikhail Vasilyevich Lomonosov, also rejected the theory of phlogiston, because metals increase in weight on oxidizing.

In 1852 Joule and Thomson made another important contribution to thermodynamics by showing that, on expansion, certain gases cooled, an effect that has since been applied in liquefaction of the so-called permanent gases and in design of refrigerating and air-conditioning systems. Nitrogen and oxygen were liquefied in 1883. In 1892 the Scottish physicist and chemist Sir James Dewar developed a vacuum-insulated vessel for storing fluids at very low temperatures, and in 1898 he produced liquid hydrogen in bulk quantities. In 1908 helium was liquefied. In 1783 Lavoisier and Laplace designed the first practical ice calorimeter, a device for measuring a quantity of heat by determining the weight of ice it caused to melt, which was perfected by a German chemist, Robert Bunsen, in 1870. This was replaced by Marcellin Berthelot's calorimeter, which measured the quantity of heat by the rise in temperature of a given mass of water.

Mechanical equivalent of heat. An American–British physicist, Sir Benjamin Thompson, count von Rumford, using a horse-turned machine for boring cannons in the military arsenal at Munich, was "struck with the very considerable degree of heat which a brass gun acquires, in a short time, in being bored; and with the still more intense heat (much greater than that of boiling water, as I found by experiment), of the metallic chips separated from it by the borer. . . ." Thompson designed an apparatus in which a blunt borer was forced against the bottom of a hollow metal cylinder and was rotated by machinery operated by horses. The cylinder was enclosed in a box containing about two gallons of water. The temperature had risen 47° F at the end of an hour, and after two and a half hours the water boiled. In a report to the Royal Society of London in 1798, he said,

It is hardly necessary to add, that any thing which any in-

> sulated body, or system of bodies can continue to furnish, without limitation, cannot possibly be a material substance; and it appeared to me to be extremely difficult, if not quite impossible, to form any distinct idea of anything, capable of being excited, and communicated, in the manner the heat was excited and communicated in the experiments, except it be motion.

From the result of this experiment, Thompson calculated an equivalence between the heat generated and the mechanical energy expended. Between 1840 and 1849, Joule refined these measurements by measuring the temperature rise in water churned by a paddle driven by a descending weight.

In 1799 in England, a chemist named Sir Humphry Davy, interested in friction, used a clockwork to make a metal wheel turn against a wax-coated metal plate. The wax was melted even though the entire system was held below freezing. In addition, he rubbed two pieces of ice together, again keeping the ice below freezing, yet the ice melted. The caloric theory failed to explain these results, and Davy concluded that heat was kinetic in nature, a form of motion. When Dalton suggested the atomic theory in 1803, he still supported the caloric theory. According to his theory, atoms were in contact in liquids and solids but far apart, surrounded by caloric, in gases.

The first law. A German physicist, Julius Robert Mayer, also measured the mechanical equivalent of heat; more importantly, he suggested that energy is perfectly preserved if all its forms are included. This "law of conservation of energy," essentially the first law of thermodynamics, was not immediately accepted. In 1847 Helmholtz presented the conservation of energy in much the same terms as Mayer.

The conversions of work into heat shown by Thompson, Joule, and Davy had been preceded by demonstrations that heat could be converted into work. A French physicist, Denis Papin, demonstrated in 1681 that, when water is heated in a closed vessel, enough pressure is generated to raise the lid of the container. Using this principle, Papin designed the first steam engine in 1687 and a second, much improved version in 1707. About the same time, the English engineers Thomas Savery and Thomas Newcomen built steam-operated water pumps for use in coal mines. The Scottish engineer James Watt improved this engine and developed the much superior expansion engine in 1763.

Invention of the steam engine

Carnot's heat engine. Another proponent of the equivalence of mechanical energy and heat, although he still accepted the caloric theory, was the French physicist Sadi Carnot, who published his classic paper, *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance*, in 1824; in it he defined work (as weight lifted through a height) and set forth the concepts that later became known as the second law of thermodynamics. Carnot made a study of a theoretical heat engine (a mechanical device that transforms heat into work) and concluded that every such engine requires a source of heat and a receiver (sink), at a lower temperature, to which part of the heat is rejected. Carnot's ideal heat engine operates in cycles of four processes. In the first process, heat is taken from the source at constant temperature while a fluid in the engine expands and does work on the surroundings; in the second process, the fluid expands further, doing additional work without absorbing additional heat, and cooling to the temperature of the sink. In the third process, the fluid in the engine contracts while heat flows from it to the sink at constant temperature and work is done upon the fluid by the surroundings; in the fourth process, no heat is transferred while further work is done upon the fluid, compressing it back to its original volume and pressure and heating it back to its original temperature, that of the heat source. The engine is now in the same state as it was at the beginning of the first process. The amount of work done by the engine during the first and second processes is greater than the amount of work done upon the engine during the third and fourth, and the amount of heat rejected to the sink is less than the amount taken from the source. The difference between the amounts of heat taken from the source and rejected to the sink is equal to the net amount of work done by the engine. Carnot likened his heat engine to a waterwheel, in which work is obtained from a fall of water from a high level to a low level: in his heat engine there is a "fall" of heat from a high-temperature source to a low-temperature sink. Most important, Carnot's theory showed that for a fixed amount of heat added to the engine, there was a fixed amount of work the engine could do and that for optimum conditions this amount of work depended only on the temperatures of the source and receiver—*i.e.*, the ratio of the amounts of heat rejected and added can be mathematically expressed in terms of the temperatures of the sink and source. (Lord Kelvin later used this relationship to devise the absolute temperature scale that bears his name.) In addition, Carnot showed that a reversible engine operating between a source at a given high temperature and a sink at a low temperature has an efficiency that cannot be exceeded by any other engine working under the same temperature conditions. His heat engine remains the model of the most efficient engine to this day.

The second law. The work of Carnot was continued by (among others) a German theoretical physicist named Rudolf Clausius and, in Scotland, a mathematician and physicist, William Thomson, Lord Kelvin. In 1850 Clausius stated the second law of thermodynamics in essentially the following manner: "Heat cannot pass spontaneously from a body of lower temperature to a body of higher temperature." In 1851 Lord Kelvin stated the second law in a different but, as was subsequently proved, equivalent fashion: "It is impossible to construct an engine which would extract heat from a given source and transform it into mechanical energy, without bringing about some additional changes in the bodies taking part." In 1865 Clausius introduced the concept of entropy (a thermodynamic quantity related to the tendency of a process to occur spontaneously) and developed many of the important consequences of the second law using this quantity; in particular, he proposed that the energy of the universe is constant but the total entropy of the universe always increases. By this time, then, the foundations of classical thermodynamics had been formulated in the first and second laws.

Entropy

In 1882 Helmholtz observed that the maximum work that can be done by chemical forces is expressed by the decrease in what he called the "free energy," now called the Helmholtz function. He concluded that if the free energy of a system is maintained at constant temperature and volume is a minimum, then the system is in thermodynamic equilibrium. A Dutch physicist, Heike Kamerlingh Onnes, defined the thermodynamic function enthalpy, which Josiah Willard Gibbs, a physicist in the United States, used in defining what he called the "free enthalpy," now called the Gibbs function. Gibbs showed that his function is a minimum for a system at equilibrium when the temperature and pressure are held constant. In 1906 in Germany, the physical chemist Walther Nernst stated that theorem now called the third law of thermodynamics: if a chemical change occurs between pure crystalline solids at absolute zero, there is no change in entropy.

The molecular basis of heat. The fact that heat is energy that is transferred from one body to another by virtue of a temperature difference, with that heat flowing from hot to cold, is now universally accepted. Attempts to explain thermal energy in terms of molecular motion can be traced to a Swiss mathematician, Daniel Bernoulli, who in 1738 explained Boyle's law by the assumption that at constant temperature the mean velocity of the particles in a gas remains constant. The kinetic theory of gases was developed in the 1850s. In 1859 a Scottish mathematician and physicist, James Clerk Maxwell, amplified Bernoulli's concept, derived a law of distribution of velocities, and found an expression for the viscosity of gases using the concept of the mean free path, the distance between two consecutive collisions of a molecule. In Austria, Ludwig Eduard Boltzmann, a physicist, further extended this work and in 1877 redefined entropy in terms of probability. In 1902 Gibbs published a study

relating thermodynamic properties to statistical principles, using a very general approach, not restricted to gases. Earlier, in 1900, the German physicist Max Planck had presented his concept of the quantum of energy, and modern statistical thermodynamics was begun (see THERMODYNAMICS, PRINCIPLES OF).

Heat transfer. Since heat is energy in transition, some discussion of the mechanisms involved is pertinent. There are three modes of heat transfer, which can be described as (1) the transfer of heat by conduction in solids or fluids at rest, (2) the transfer of heat by convection in liquids or gases in a state of motion, combining conduction with fluid flow, and (3) the transfer of heat by radiation, which takes place with no material carrier. The flow of heat in metal bars was studied analytically by the French mathematician Jean-Baptiste-Joseph Fourier and measured by the French physicist Jean-Baptiste Biot in 1816. The conductivity of water was first determined in 1839; the conductivity of gases was not measured until after 1860. Biot formulated the laws of conduction in 1804, and Fourier published a mathematical description of this phenomenon in 1822. In 1803 it was found that infrared rays are reflected and refracted as visible light is, and, thenceforth, the study of thermal radiation became part of the study of radiation in general. In 1859 a physicist in Germany, Gustav Robert Kirchhoff, presented his law of radiation, relating emissive power to absorptivity (see SPECTROSCOPY, PRINCIPLES OF). An Austrian, Josef Stefan, established the relationship (now called the Stefan–Boltzmann law) between the energy radiated by a blackbody and the fourth power of its temperature. Boltzmann established the mathematical basis for this law of radiation in 1884. It was in the study of radiation that Planck arrived at the concept of the quantum. Understanding of heat transfer by convection was developed during the period 1880–1920, although an equation describing such processes had been suggested by Sir Isaac Newton in 1701.

Conduction, convection and radiation

APPLICATIONS

Refrigerators and air conditioners. Modern air conditioning and refrigeration use the cooling effect of the expansion of gases but also use the effect of latent heat (*i.e.*, the heat rejected in condensation and the heat added in evaporation of the refrigerant). The basic components of a refrigeration system are a compressor followed by a condenser; then an expansion device, which can be a valve, a capillary tube, an expansion engine or an expansion turbine; and finally an evaporator. (For details see HEATING, VENTILATING, AND AIR CONDITIONING.)

Reciprocating steam engines. As was previously mentioned, the first steam engine was developed by Denis Papin in 1690, made practical for pumping water by Savery in 1698, and improved by Newcomen in approximately 1712. The principle was to lead steam into a cylinder, condense it with an outside coolant, and let the resulting vacuum suck a piston into the cylinder. Newcomen's engine, modified somewhat, was unrivalled until 1765, when Watt introduced the separate condenser. In 1782 Watt patented two other important changes, making the piston double-acting (the piston is driven alternately in each direction) and letting the steam expand against the piston by cutting off the flow of steam instead of letting the steam flow continuously. Many improvements of the basic design of the steam engine were introduced later, including the use of higher pressures (Watt's engine used steam at only two to three pounds per square inch [0.1 to 0.2 kilogram per square centimetre] above atmospheric pressure), higher speeds, and lighter parts. By 1900 the more efficient steam turbine had made the reciprocating steam engine obsolete in stationary power plants, but, in the third quarter of the 20th century, reciprocating engines using steam or other fluids were receiving renewed interest for possible use in automobiles. (For details see STEAM POWER.)

Turbines. If the steam engine is replaced by a steam turbine, the cycle of changes in the pressure, volume, and temperature of the steam is called a Rankine cycle, after the Scottish engineer William J.M. Rankine. Two major modifications have been made in the use of the steam turbine to improve the efficiency. The first of these is to have two turbines, one operating at high pressure and the second at low pressure, returning the steam to the generator for reheating after expanding in the high-pressure turbine. The reheated steam is then sent to the low-pressure turbine. The second modification involves the concept of regeneration (*i.e.*, an exchange of heat within the system). In a regenerative cycle, steam is bled from the turbine or turbines and led to a feedwater heater to preheat the water going to the steam generator. The function of the steam generator is the same whether the fuel is a fossil fuel or a nuclear one. In order that the entire plant not be radioactive, a nuclear power plant must have an intermediate fluid, heated by the nuclear reactor, that produces the steam in the generator.

Another type of turbine power plant, used in aircraft, employs an open cycle, the exhaust gases leaving the turbine not being returned to the compressor. Air is partially compressed in the inlet to the compressor and fully compressed by the compressor itself. Fuel is added, and combustion takes place before the gases enter the turbine. The exhaust gases expand through a nozzle after leaving the turbine. In a jet engine, the turbine provides power to drive the compressor, and the difference between the exit and inlet velocities provides the thrust for the aircraft. In a turbo-prop engine, the turbine also provides power to drive a propeller (for details see TURBINE and JET ENGINE).

Jet propulsion

Internal combustion engines. When Papin was investigating the use of steam for power, he also tried using the explosive power of gunpowder in a piston–cylinder device. This concept, suggested by Christiaan Huygens, was the predecessor of the internal combustion engine. The first such engines using the burning of gas as the power source were much like steam engines. The gas was introduced into the cylinder and then ignited by some external mechanism. The pressure resulting from the combustion acted on the piston, causing it to move. One such engine used an electric spark to ignite the combustible mixture of gas and air. In 1862 an engine designed to operate on four strokes was patented but was not built before the patent lapsed, and the cycle is known as the Otto cycle, after Nikolaus August Otto, a German inventor, who in 1878 revived the four-stroke engine with its important idea of compressing the combustible mixture before ignition and produced a gas engine using this principle.

The Otto four-stroke engine

Since gas was not available everywhere, efforts were made to develop an engine that would operate using oil. Many of the first oil-fuelled engines operated on the Otto cycle. The oil was first atomized, mixed with preheated air to vaporize the oil, and then this mixture was drawn into the cylinder. In 1892, in Germany, Rudolf Diesel patented a cycle in which fuel is added after the air has been compressed. The combustion occurs spontaneously because the air is hot enough after compression to ignite the fuel.

In 1885 Gottlieb Daimler, another German inventor, developed the surface carburetor, which permitted the use of gasoline as a fuel in the Otto cycle. The Otto cycle is still the most frequently used cycle for gasoline engines, the Diesel cycle for oil engines.

In 1878 the Scottish engineer Sir Dugald Clerk developed the two-stroke cycle, in which the intake and compression strokes are combined and the expansion and exhaust strokes are combined. The cycle is used primarily for fuel-injection or Diesel engines. Another internal combustion engine of interest is the Wankel, or rotating combustion, engine. In the Wankel engine, three cyclic processes proceed simultaneously in three separate chambers formed by an essentially triangular rotor in a nearly elliptical stator. The steps occurring are intake and compression, combustion and expansion, and exhaust. (For details see DIESEL ENGINE and GASOLINE ENGINE.)

Nuclear power plants. The power-producing portion of a nuclear plant is the same as in any modern high-pressure, high-temperature steam power plant. The major difference lies in the source of heat added to the feedwater to produce the high-temperature steam. In general, the nuclear reaction heats an intermediate fluid, which in

turn heats the feedwater. In experimental reactors the fluid may be liquid sodium, but commercial units use boiling water or pressurized water. (For details see NUCLEAR REACTOR.)

Other power sources. The desire to find new and lightweight sources of power has led to the investigation of many direct energy conversion devices. Three such devices are briefly described.

The fuel cell may be considered an electric battery in which both the fuel and oxidizer are continuously replaced. In the fuel cell, the chemical energy is not converted to heat but directly into electrical energy.

With the development of semiconductors, thermoelectric generators and refrigerators became possible. In the generator, heat is added to the system at a high temperature and removed by a cooling system at a lower temperature. Because of the temperature difference across the semiconductors, a current flows. The opposite effect is observed in the refrigerator, in which power is supplied, causing a current to flow that in turn produces the temperature difference. The effects upon which the operation is based are called the Seebeck effect (see above *Thermometry*) and the Peltier effect. In 1834 the French physicist Jean Peltier observed the reverse of the Seebeck effect: the passage of current through a junction of two dissimilar conductors produced a heating or cooling effect depending on the direction of the current. Lord Kelvin found a third effect, namely, that the flow of current in a conductor with a temperature gradient produces heating or cooling depending on the direction of flow. He also developed the relationship between the three effects (for details see BATTERIES AND FUEL CELLS; ELECTRIC POWER; THERMOELECTRIC DEVICES).

BIBLIOGRAPHY. F. SHERWOOD TAYLOR, *A Short History of Science and Scientific Thought, with Readings from the Great Scientists from the Babylonians to Einstein* (1963), places the development of ideas of heat in the context of the evolution of science; H.C. VAN NESS, *Understanding Thermodynamics* (1969), is a readable treatment requiring little mathematical background; B.D. WOOD, *Applications of Thermodynamics* (1969), requires some acquaintance with thermodynamics, and provides excellent coverage of applications. Classic works available in English include J.-B.-J. FOURIER, *Théorie analytique de la chaleur* (1822; Eng. trans., *The Analytical Theory of Heat*, 1878); and MAX PLANCK, *Vorlesungen über Thermodynamik*, 7th ed. (1921; Eng. trans., *Treatise on Thermodynamics*, 3rd ed., 1927).

(L.Gr.)

Heat Exchanger

The heat exchanger is a device in which heat from a hot fluid is transferred to a cold fluid; the temperature of the hot fluid decreases and that of the cold fluid increases. Heat exchangers are manufactured in many different designs and are used extensively in various technologies—for instance, in steam and nuclear power plants, in gas turbines, in heating and air-conditioning, in refrigeration, and in the chemical industry. More recently, special types of heat exchangers have been developed for artificial satellites and space vehicles. Heat exchangers are given different names when they serve a special purpose. Thus boilers, evaporators, superheaters, condensers, and coolers may all be considered heat exchangers.

PRINCIPLES OF OPERATION

The basic operation of a heat exchanger may be discussed with the help of a simple type shown in Figure 1. This exchanger is constructed from two pipes in a concentric arrangement. Inlet and exit ducts are provided for the two fluids. In the sketch the cold fluid flows through the inner tube and the warm fluid in the same direction through the annular space between the outer and the inner tube. This flow arrangement is called parallel flow. In it heat is transferred from the warm fluid through the wall of the inner tube (the so-called heating surface) to the cold fluid. The temperature in both fluids varies as shown in the lower part of Figure 1. The temperature of the warm fluid decreases from t_{w1} to t_{w2}, and the temperature of the cold fluid increases from t_{c1} to t_{c2}. The amount of heat Q that is transferred from one fluid to the other per unit

Parallel flow arrangement

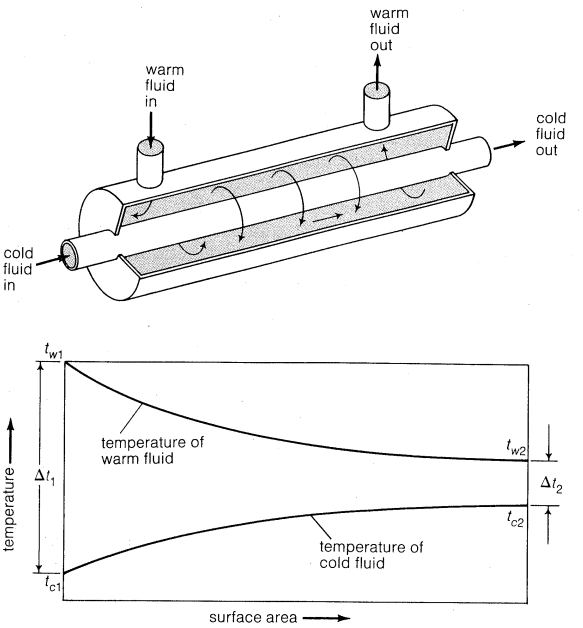


Figure 1: (Top) Operating principle of a parallel-flow heat exchanger and (bottom) changes in fluid temperatures that take place therein (see text).

of time, called heat flow, can be calculated from the following equation:

$$Q = mc\,(t_2 - t_1) \qquad (1)$$

This equation states that the heat flow Q (in kilowatts [kW]) can be obtained by multiplying the mass per unit time of fluid m (in kilograms per second [kg/sec]) by the specific heat c of the fluid and by the temperature increase $t_2 - t_1$ of the fluid from the entrance to the exit of the heat exchanger (in degrees Centigrade). The specific heat is a property of the fluid involved. For example, it has to be entered with the value 1 for air and 4.18 for water into equation (1). The amount of heat leaving the warm fluid must be the same as the amount of heat received by the cold fluid. The mass flow and the temperature increase for the cold or the decrease for the warm fluid can therefore be entered into equation (1). The heat exchanger may have to be designed, for example, to increase the temperature of a prescribed mass per unit time m_c of the cold fluid from t_{c1} to t_{c2}. Entering these values into equation (1) then determines the heat flux Q which has to be transferred in the heat exchanger. This value will be needed in the following discussion to calculate the heating surface of the exchanger. It is sometimes preferred to express the heat flux in thousands of calories per hour (kcal/hour). The corresponding value can be obtained by multiplying the heat flux Q in kW, obtained from equation (1), by 860.

Counterflow arrangement

The temperature difference Δt_1 between the fluids at the entrance of the heat exchanger decreases to the value Δt_2 at the exit, as illustrated in Figure 1. A heat exchanger is operated in counterflow when the flow direction of one of the fluids is reversed. The counterflow arrangement has the advantage that the exit temperature t_{c1} of the colder fluid can be increased beyond the exit temperature t_{w2} of the warm fluid. In addition, a smaller surface area is required in counterflow than in parallel flow to transfer the same amount of heat. This is so because the mean temperature difference Δt_m in the counterflow heat exchanger, for a given heat flux and prescribed inlet temperatures, is larger than in the parallel-flow exchanger.

The heating surface of the heat exchanger can be obtained from the equation:

$$A = \frac{Q}{U\Delta t_m} \qquad (2)$$

The equation indicates that the required surface area A (in square metres [m²]) is obtained by dividing the heat

flux Q obtained with equation (1) by the overall heat transfer coefficient U (to be discussed later) and the mean temperature difference Δt_m (in degrees Centigrade).

Larger heat exchangers utilize a bundle of tubes through which one of the fluids flows. The tubes are enclosed in a shell with provisions for the other fluid to flow through the spaces between the tubes. Fluid flowing outside the tubes can be directed either in the same direction as or counter to the effective flow in the tube bundles (see Figure 2). In the latter arrangement, parallel or counter flow

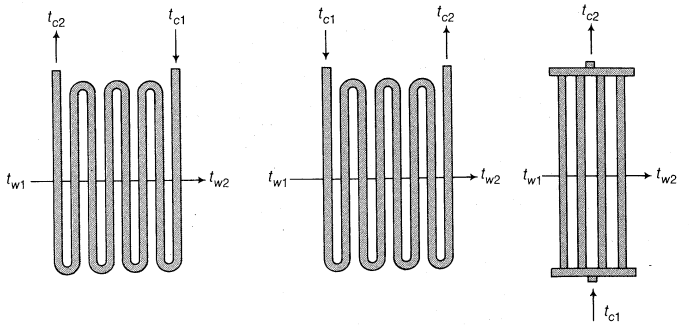


Figure 2: (Left) Counterflow, (centre) parallel flow, and (right) crossflow in tubular heat exchangers (see text).

can be approximated in the way shown in Figure 2. In another arrangement, the cold fluid is distributed in such a manner that it flows in parallel through the tubes forming the heating surface and is then collected by a header. This arrangement creates a cross flow, as shown schematically in Figure 2. In nuclear reactors fuel rods may replace the tubes, and the cooling fluid flowing around the rods removes the heat generated by the fission process. In a similar way, rods containing electric resistance heaters may supply heat to the fluid passing through the exchanger between the rods.

MECHANISM OF HEAT TRANSFER

The mechanism of heat transfer from the hot to the cold fluid must be considered. For this purpose, attention may be focussed on local conditions determining transfer from the warm to the cold fluid. Several physical processes are involved in this transfer. One of the basic laws of thermodynamics states that heat by itself always flows in the direction of decreasing temperature. In the heat exchanger, a temperature drop Δt_f is required to transport the heat from the bulk of the warm fluid to the heating surface. Another temperature drop Δt_w is needed to transport the heat through the wall between the warm and cold fluids, and yet another Δt_f to move the heat from the surface of this wall into the bulk of the cold fluid. Figure 3 illustrates the temperature variation just described.

Heat conduction

The physical process that causes the heat to flow through the solid wall from one surface to another is called heat conduction. It has to be visualized in a specific way: each atom of the wall passes on its energy to the neighbouring atoms of lower temperature, handing heat energy from the hot side of the wall to the cold side. The temperature drop required to produce this heat flux depends on the thickness of the wall, its surface area, and its thermal conductivity. The heat flux produced by a temperature difference Δt_w between the two surfaces of the wall can be calculated by the following equation:

$$Q = \frac{k}{s} A \Delta t_w \qquad (3)$$

The equation states that the heat flux Q (in kW) is obtained by multiplying the thermal conductivity k (a measure of the ability of a material to conduct heat) by the area of the heating surface A of the heat exchanger (in m^2), by the temperature difference Δt_w (in degrees Centigrade), and then dividing by the thickness s of the wall (in m). In a heat exchanger it is desirable to decrease the temperature drop so that a smaller overall temperature difference results. This can be achieved by reducing the wall thickness and by selecting a material with a high thermal conductivity. Metals have high values of k, and the magnitude of the thermal conductivity is roughly proportional to the electrical conductivity. In other words, metals that are good electrical conductors are also good conductors of heat. The following values of k may be used in equation (3) for materials that are frequently used in heat exchangers: 0.38 for copper, 0.2 for aluminum, and 0.04 for steel.

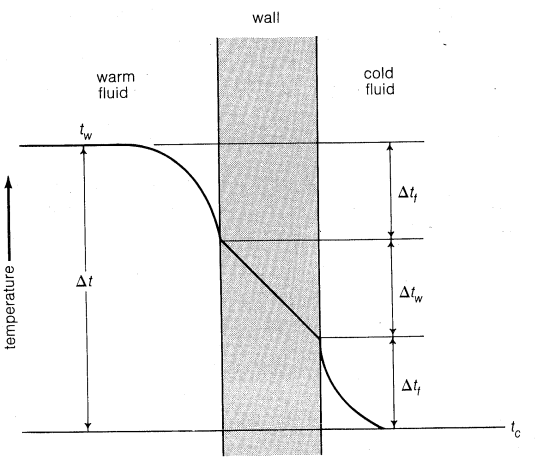


Figure 3: Heat flow from a warm fluid to a cold fluid through a conducting wall (see text).

Heat transfer from a fluid to a solid surface takes place partially by conduction. It is supplemented, however, by a process in which the fluid carries heat along as it moves through the heat exchanger. This process is called convection and depends on the nature of the flow. There are two types of flow that occur in nature. In one type, called laminar, the fluid particles move along in a smooth way side by side. In turbulent flow conditions, waves and vortices are continuously formed and destroyed again. Superimposed on the mean flow, these vortices cause a continuous mixing of fluid particles. Whether the flow through a tube is laminar or turbulent depends on the tube diameter, on the velocity of the fluid, and on fluid viscosity. The flow tends to be laminar when the velocity is low, the tube diameter is small, and the viscosity is large. Laminar flow is encountered in oil coolers, for example, because oil has a high viscosity. Heat exchangers for liquids like water or for gases usually operate with turbulent flow. The heat transfer from the bulk of the fluid to the tube wall in laminar flow results mainly from conduction and is then described by an equation similar to equation (3). Consequently, it is determined by the conductivity of the fluid and by the tube diameter. Liquids have considerably larger heat conductivity values than gases and thus transfer heat more readily. Liquid metals have especially large heat conductivity values and are for this reason used in engineering applications in which large amounts of heat have to be transferred at small temperature differences. Some new designs of nuclear reactors use liquid metals as heat transfer media.

In turbulent flow, the mixing process described previously constitutes a third mechanism that transfers the heat from the bulk of the fluid to the wall. It is understandable that this process depends strongly on the velocity of the fluid. Heat transfer in turbulent flow is therefore larger than in laminar flow.

The following equation is used to calculate the heat transfer in a fluid for either laminar or turbulent flow:

$$Q = h A \Delta t_f \qquad (4)$$

The heat transfer coefficient

According to this equation, the heat flux Q (in kW) is obtained by multiplying a parameter h, called heat transfer coefficient, by the heat transfer area A (in m^2), and by the temperature difference Δt_f between the bulk of the fluid and the wall in either the warm or the cold fluid. Values for the heat transfer coefficient depend on the parameters that have been mentioned above and can be determined by fairly involved calculations. Ranges commonly encountered can be mentioned here. In air, for

instance, values that have to be introduced into equation (4) range between 0.01 and 0.1 and in water between 0.5 and 5. These values indicate that much larger heat transfer surfaces are required to transfer the same amount of heat in a gas than in a liquid. Correspondingly, it is advantageous to provide a larger heat transfer surface on the gas side of a heat exchanger than on the liquid side. This is done by equipping the tubes with fins on the side of the surface that is exposed to the gas flow. Much effort has been directed toward improving manufacturing processes for such finned surfaces.

The overall heat transfer coefficient U required in equation (2) is obtained from the equation:

$$\frac{1}{U} = \frac{1}{h_w} + \frac{s}{k} + \frac{1}{h_c} \qquad (5)$$

in which h_w and h_c denote the heat transfer coefficients for the warm and cold fluids, respectively, and s and k have the same meaning as in equation (3).

The temperature drop in heat exchangers increases occasionally with time when some material is deposited from the fluids on the surfaces of the heat transfer walls. This "fouling process" degrades the performance of the heat exchanger and may necessitate periodic cleaning to restore it to the original performance.

TYPES OF HEAT EXCHANGERS

Specific heat exchangers had to be developed for satellites and space vehicles. These vehicles operate in a vacuum; therefore heat cannot be transferred by conduction or convection. Here, heat transfer has to be effected by another process called thermal radiation. This mechanism of heat transfer, used in such heat exchangers, is the same as the process by which the Earth receives energy from the Sun. Energy transfer by radiation is especially effective at high temperatures. In engineering devices operating at high temperatures, like steam boilers or gas turbines, heat transfer is caused by radiation added to the conductive and convective heat transfer.

Optimizing heat exchangers

In designing a heat exchanger for a specific purpose, the initial cost resulting from the weight of the heat exchanger, the material used, and the required labour must be balanced against the operating cost of the power required. Power, usually in the form of electrical energy, has to be expended to overcome the pressure drop of the fluids in the heat exchanger. Much thought has been given to the problem of optimizing heat exchangers. In general, it has been found that small tube diameters and small clearances between the tubes reduce the initial and the operating costs. Designers of heat exchangers are constantly striving to develop more effective surface configurations. A wavy shape, for instance, was found to be more effective than straight fins. In this way, compact heat exchangers have been created.

Regenerative heat exchangers. A different type of heat exchanger, conventionally referred to as a regenerative heat exchanger, is constructed from a porous material through which the warm and the cold fluids are ducted alternately. The heat is stored in the porous material—the matrix—during the time the warm fluid passes through and is released from the porous material to the cold fluid when the latter fluid passes through. The matrix can be formed in a multitude of ways from screens, fibres, or other elements. Nonmetallic materials, like ceramic and glass, have also been used. The surface area per unit volume can be made very large by providing many passages with small dimensions. This leads to very compact heat exchangers. The alternate flow of the two fluids can be controlled by valves that at the proper time open and close the inlet and exit ducts for the two fluids, or the heat exchanger can be designed in such a way that the matrix rotates through the passages of the two fluids. Such heat exchangers, introduced by a Swedish engineer, F. Ljungstrom, were used frequently, beginning around 1930, in steam power plants. The compact size of this heat exchanger makes it desirable for gas turbines that power automobiles, airplanes, or other vehicles. Still to be overcome in its development is the need for effective and durable seals between the rotating matrix and the ducts for the two fluids.

Heat pipes. A heat exchanger of a different kind has recently stimulated much attention. Called the heat pipe, it is used to transport heat over relatively large distances with the temperature difference kept as small as possible. The heat pipe consists of a hollow tube closed at both ends and partially filled with a liquid that boils at a desired temperature. One end of the tube is immersed in the warm region, and the other in the cold region. The objective is to transfer heat through the pipe from the warmer to the colder region. The process may be visualized by assuming that the tube is in a vertical position with the lower end immersed in the warm region. The liquid fills the lower end of the tube and starts boiling when the temperature of the warmer region exceeds the evaporation (or boiling) temperature of the liquid. The accumulation of vapour increases the pressure at the lower end of the tube. This forces the vapour upward, where it condenses since the temperature of the colder region is below the evaporation temperature. The condensed liquid runs down along the inner surface of the tube, due to gravity. In this way a steady circulation of the fluid is maintained. Heat is required to evaporate the liquid. This heat of evaporation is removed in the boiling process from the hot region. The same amount of heat is released when the fluid is condensed and is transferred to the cold region. Consequently, circulation of fluid causes transport of heat from the warm to the cold region.

From *The Heat Pipe* by G. Eastman. Copyright © (1968) by Scientific American, Inc. All rights reserved.

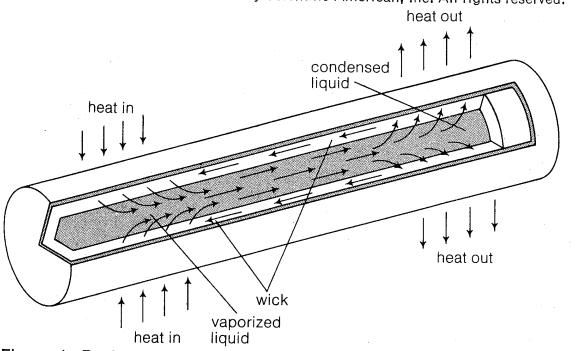


Figure 4: Basic operation of a heat pipe in which condensed liquid (at right) is returned to evaporator section (at left) by means of a wick.

Schmidt's experiments

Development of the heat pipe. Experiments conducted by a German engineer, E. Schmidt, in 1939 demonstrated that this mechanism for transporting heat is especially effective if the working fluid is close to its critical point. In 1960 he reported that a tube filled with ammonia or carbon dioxide near its critical point transfers an amount of heat per unit time that is more than 4,000 times larger than the amount of heat transferred by a solid rod of copper with the same dimensions and at the same temperature difference between the hot and cold regions.

A device of this kind can only operate in a vertical position, with the hot region located below the cold region. This limitation was eliminated through an improvement conceived by R. Gaugler at General Motors Corporation in 1942. He incorporated a wick or porous matrix covering the inside wall of the tube into the design, as shown in Figure 4. Because the wick contained many capillary passages, capillary action transported the liquid through the wick from the cold to the hot end of the tube, regardless of the direction in which it was oriented. This capillary action may be illustrated by an experiment in which a small diameter glass tube is held in a vertical position with its lower end immersed in water. The water rises in the tube to a level that is higher than that of the water surrounding the tube. The forces of attraction between the inner tube surface and the water surface—a capillary action—are able to overcome the force of gravity and lift the water in the tube above the surrounding liquid level. The height to which the water rises is greater, the smaller

the diameter of the tube. When the tube is heated in such a way that water evaporates from its surface, the water level in the tube is maintained by a continuous inflow through the lower tube end. In this way, a continuous flow of liquid through the tube is generated. Exactly the same process occurs in the wick of the heat pipe where the working fluid continuously evaporates at the hot end of the wick. Capillary forces continuously replace the evaporated liquid. The circulation of the fluid in the pipe occurs in such a way that a pressure difference in the vapour causes the vapour to flow from the warm to the cold end of the tube. Capillary forces return the liquid through the wick.

Heat pipe in space power plants

Since the device conceived by Gaugler did not become widely known, G.M. Grover, at Los Alamos Scientific Laboratory, described it again in 1963. He gave it the name "heat pipe," suggesting that it be used to transport heat in power plants for space vehicles. The effectiveness of the device derives from the fact that only small temperature differences are required to move the heat through the tube wall. Additionally, the transport of heat with the vapour requires almost no temperature difference. In this way, very large quantities of heat can be transported over considerable distances with very small temperature differences. In one such device, for instance, a heat flux in the amount of 11 kW was transported by a heat pipe with 2.5 centimetres diameter and 69 centimetres length, with a temperature difference so small that it was difficult to measure. By comparison, a temperature difference of 40,000° C would be required if a rod of copper (one of the best thermal conductors) with the same dimensions were used to transfer the same amount of heat. The hot end of the rod would thus have had a temperature much greater than the surface temperature of the sun, even though the other end was at room temperature. Obviously the copper rod would have evaporated long before the high temperature could have been reached.

This ability to transport very large quantities of heat with small temperature differences is the main feature characterizing the heat pipe. The device has other desirable features as well. For instance, the heat pipe operates whether the heat at the warm or cold end is transferred to a large or small surface area. When heat is concentrated at one location, it can be removed through a heat pipe and distributed over a large region. This feature has become important in the effort to miniaturize electronic equipment. In such equipment heat is generated as a by-product; the heat pipe makes it possible to remove a large amount of heat from the interior and transport it to some place where it can easily be removed by a cooling air stream. Consequently, electronic equipment can be constructed in very compact form. The fact that heat can be transported through the pipe with very small temperature differences suggests its application in cases where a uniform temperature of an engineering device is desirable. The heat pipe has thus been found useful in space flight. It eventually operates at a nearly constant temperature regardless of the amount of heat, large or small, it has to transfer per unit time. It can be used, therefore, to keep temperatures in various devices constant in time even when the amount of heat that has to be removed varies.

Limitations of heat pipes. On the other hand, a specific heat pipe has certain limitations for its operation. Each wick has a maximum amount of liquid flow that it can transport. This prescribes an upper limit for the amount of heat that such a pipe can transport per unit time. When the liquid at the warm end of the pipe evaporates, vapour bubbles are created at the tube surface and separate when they have reached a certain size. The density of the vapour bubbles at the surface increases as the amount of heat transferred per unit area and time increases. It finally reaches a condition in which the bubbles combine to form a continuous vapour film blanketing the surface. This condition, called the boiling crisis or burn-out condition, has to be avoided since vapour is a very poor conductor of heat. As a consequence, the temperature of the tube material may increase to values that cannot be tolerated. The boiling crisis then determines an upper limit for the amount of heat that can be transferred from the warm region to the heat pipe per unit area and time. The maximum temperature at which the heat pipe can operate is determined by the fact that the pressure of the fluid inside the pipe increases with temperature, because the evaporation temperature depends on pressure. An upper limit is then given by the pressure that the walls of the tube can withstand. Nevertheless, heat pipes can be designed to operate in a wide temperature range from below 0° C up to around 2,000° C. This is accomplished by proper choice of the liquid that the pipe contains and the material of which the pipe is manufactured. Methanol, acetone, water, fluoridated hydrocarbons, mercury, lithium, lead, bismuth, and a range of inorganic salts have been used as working fluids; the tube itself has been made of glass, ceramic, copper, stainless steel, nickel, tungsten, molybdenum, tantalum, and various alloys. Wicks have been constructed of sintered, porous materials, woven mesh, fibre glass, longitudinal slots, and combinations of these structures. An interesting example of a specialized application is a heat pipe in which both the pipe and the working fluid are electrically nonconducting. With such construction, a heat pipe can be used in an electrical device to transfer heat from locations where the electric potential exceeds 5,000 volts. The physical size and configuration of heat pipes can vary widely.

The heat pipe has become a unique and versatile heat transfer device with its range of applications seemingly unlimited.

BIBLIOGRAPHY. A.P. FRAAS and M.N. OZISIK, *Heat Exchanger Design* (1965), is intended to help practicing engineers design heat exchangers but should also be useful for the layman. The book describes heat-exchanger types for various engineering applications, fabrication techniques, fluid flow, heat transfer, stress considerations, testing of heat exchangers, and design calculations. Extensive tables of required properties and relations for design analysis are appended. W.M. KAYS and A.L. LONDON, *Compact Heat Exchangers* (1958), presents an introduction to heat-transfer and pressure-drop analysis, but is useful for engineers only. Tables and diagrams present test results for a large number of surface configurations of compact heat exchangers. G.Y. EASTMAN, "*The Heat Pipe*," Sci. Am. 218:38–46 (1968), discusses physical processes in a heat pipe and its operation in terms understandable to the layman. The advantages as well as the limitations are described and many applications are mentioned. Useful illustrations supplement the text.

(E.R.G.E.)

Heating, Ventilating, and Air Conditioning

Heating, ventilation, and air conditioning are all aspects of environmental engineering, a recent concept embracing all aspects of the engineering of buildings, such as structure, drainage, acoustics, and internal transportation. The concept of environmental engineering takes cognizance of the fact that all elements of building are interrelated. The heat produced by lighting, for instance, affects the need for air conditioning, calling for ducts, which in turn affect the design of the structure.

The present article is confined to buildings, but the subject may be extended to cover mobile vehicles, such as motor coaches, aircraft, ships, railway trains, spacecraft, and submarines, each having highly specialized requirements. Heating, ventilating, and air conditioning in the present context may be defined in general terms as the control of the environment within enclosed spaces. Apart from comfort, many industrial processes depend on a controlled atmosphere; thus, this branch of engineering covers a considerable range of applications. Heating is concerned with raising the temperature of the thermal environment. Ventilating is concerned with the supply of fresh air and removal of air that is vitiated, polluted, or contaminated. Air conditioning may include the functions of heating and ventilating but in addition is concerned with lowering of temperature of the thermal environment, raising or lowering of humidity, and purifying

the air by removal of dust, bacteria, and other airborne matter.

History. Making fire was among man's earliest achievements, and doubtless wood formed the earliest of fuels, first in caves, as remains of Stone Age hearths show, and later in mud and turf enclosures.

The discovery that charcoal could be made from wood to produce a fuel without smoke seems to have been an early step toward progress in places where only moderate warmth was needed, such as China, Japan, and the shores of the Mediterranean.

The invention of the fireplace

Another evolution was the flue or chimney, first as a simple aperture in the centre of the hut roof and later rising from the fireplace, invented in Europe in the 13th century. Smoke and fumes no longer pervaded the living space.

Stoves, far less wasteful of heat than fireplaces, appear to have been used first by the Chinese about 600 BC. They can be traced through Russia and into Germany and in the European countries generally, where they are still used today, often as a focus of family life. The stove crossed the Atlantic to the United States, where Benjamin Franklin in 1744 invented an improved design, the forerunner of the traditional potbellied stove.

The first scientific refinement of the fireplace seems to be attributable to the physicist Benjamin Thompson, count Rumford, toward the end of the 18th century. His aim was to improve the efficiency of the open fire by the use of lumps of fireclay and by a canopy, both designed to increase the amount of radiant heat. Because his ideas were not accepted, however, a gross fuel waste has continued up to the present day.

Heating by a fire outside the space to be heated, now described as central heating, appears to have been invented by the Lacedaemonians of Greece, who first used heated floors. The Great Temple in the city of Ephesus (350 BC) is believed to have been heated by flues laid in the floors, using lignite as fuel.

The Greeks perceived the advantages of central heating, but it was the Romans who became the supreme heating engineers of the ancient world with their hypocaust system. The floor was raised on pedestals, or *pilae*, and the hot gases from a furnace were guided into the underfloor space, from which they rose through hollow terra-cotta tubes embedded in the walls. Such systems are to be found throughout Europe wherever Roman culture flourished. In Italy the hypocaust is found only in baths, but in cooler climates, such as Britain's, not only baths but also the living room and sometimes other rooms were heated. The form of ducting and mosaic floor of one Roman room are shown in Figure 1 (left); the *pilae* are illustrated in Figure 1 (right). The fuels used in hypocausts were charcoal, brushwood, and, occasionally, coal. Coal has been recorded on 20 sites in ancient Britain. At Ely remnants of unburnt fuel indicate that the coal came from the Forest of Dean, 130 miles away.

Such scientific development and refinement of life came to an end with the fall of the Roman Empire and the ruin of the cities that followed. The Dark Ages saw a return to a less civilized form of life; castles and homesteads employed the crude methods of heating used by primitive man. Drafty halls were warmed by a log fire in the centre of the stone floor. Enormously heavy fur-lined cloaks were worn to keep warm. It has taken 1,500 years for the comfort of the Roman system of floor heating to be rediscovered by modern civilization.

The advent of steam as a source of power in the Industrial Revolution of the 18th and early 19th centuries offered a new way of heating, first used in factories and mills. Steam conveyed in pipes was extensively used for heating not only for industrial needs but for schools, churches, courts of justice, assembly halls, and even homes and horticultural greenhouses.

Using steam for heat

The very hot surfaces of steam heating cause a parching effect on the air, often accompanied by a disagreeable odour of burnt dust. The advantages of hot water, with a lower surface temperature and milder general effect than that of steam, began to be recognized about 1830.

One of the first such systems was installed at the New Westminster Hospital, London. Hot water at low pressure has continued from that time to occupy a principal place in methods of heating; it is used in radiators, convectors, embedded floor and ceiling heating systems, and as a means of warming air for distribution by fans in a variety of ways.

In 1831 Jacob Perkins of England patented a revolutionary method of high-pressure hot-water heating in which a continuous circuit of very strong piping, receiving heat from a coil in the furnace, conveyed water in a closed system to coils disposed about the building. Circulation was by thermo syphon. The system achieved considerable popularity in spite of its high temperatures; its tubing was small and neat compared with the massive cast-iron pipes used with low-pressure apparatus. The high-pressure hot-water principle has been revived in a modern system developed first in continental Europe and later spread elsewhere, chiefly for industrial heating.

The need for some form of induced ventilation to enclosed spaces probably did not arise until the 19th cen-

Reprinted with permission from Sir Mortimer and Mrs. Wheeler, "Verulamium," (1936); Society of Antiquaries of London

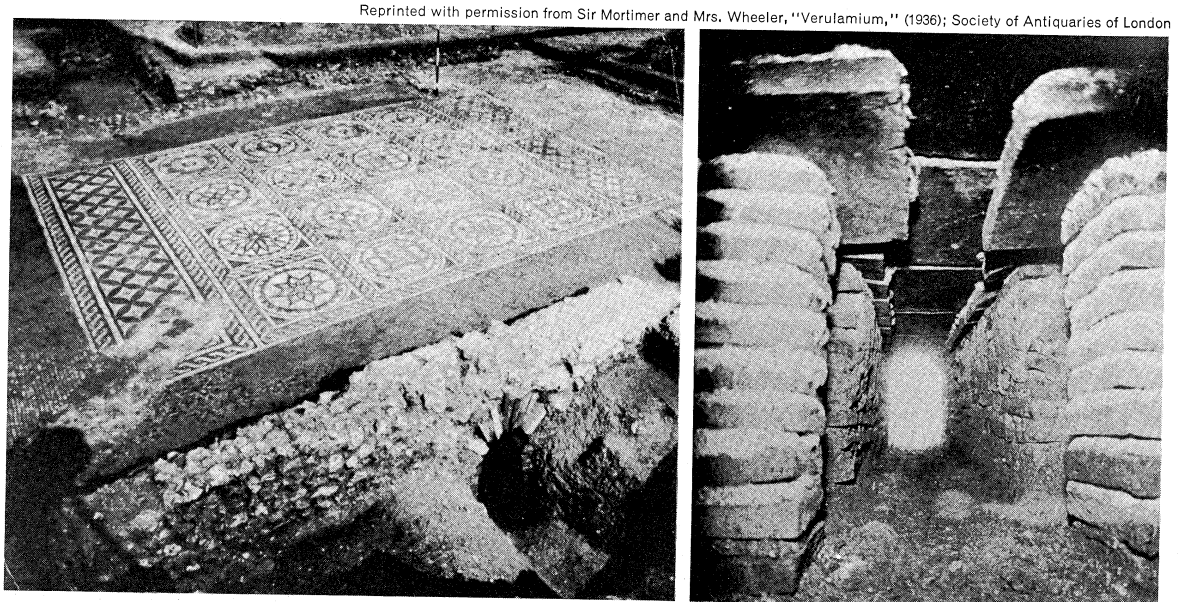


Figure 1: *Roman hypocaust system of heating.*
(Left) Mosaic floor showing underground ducts from furnace. (Right) View below mosaic floor showing pilae.

tury, by which time assembly halls, theatres, and churches were being built to hold hundreds and even thousands of people. In industry some amelioration of the atmosphere in the steamy, gas-lighted workrooms was forced on mill owners to offset an increasing sickness rate.

Mines had been ventilated from earliest times by furnaces kept burning at the base of vent shafts. Some of the early ventilating systems in buildings followed this method, for example, the New Houses of Parliament in London, rebuilt in 1837. There air was exhausted from the chambers by means of coke fires, kept burning day and night in the roof spaces beneath tall shafts, whose Gothic outlines remain a feature of the architecture, though the shafts are no longer used. The air was drawn in by the suction of these extraction shafts through gratings in the floor of the chambers, which were connected in turn to tunnels leading to the Thames River, from the banks of which the fresh air was drawn.

Another method of ventilating by heat was by the use of gas jets in large arenas and in assembly rooms and town halls. Fresh air was admitted by Tobin tubes connected to the outside, as shown in Figure 2. The ring of gas jets beneath an exhaust vent through the roof produced the necessary updraft.

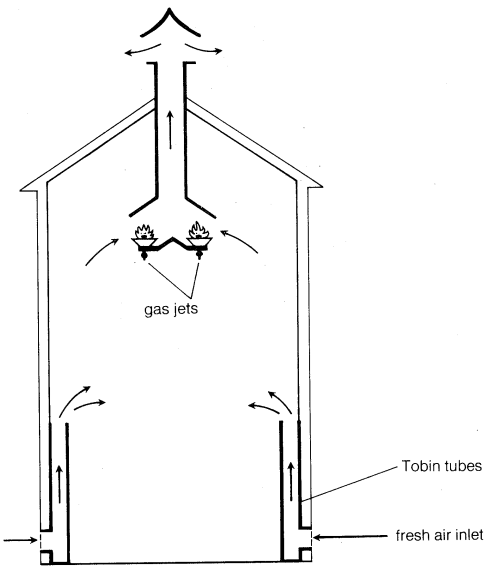


Figure 2: Early form of building ventilation. Gas jets produce updraft, drawing in fresh outside air through the Tobin tubes.

Although a rotary fan had been invented in the 16th century, a convenient power supply was not available until the days of ubiquitous electric current. From the 18th century steam drove fans in industrial installations.

The plenum system was an early method of combining heating with ventilation by using steam in tubes in an air duct to heat the air, the driving force being a fan, also steam driven. The heated air was delivered through ducts under pressure (or plenum), conveyed to all parts of the building. Schools, hospitals, and factories were often heated by this means.

India's wetted mat cooling system

Cooling by evaporation may have originated in India, where mats of wetted grass hung over openings on the windward side resulted in evaporative cooling by as much as 20 to 30° F (11 to 17° C). The mats were kept wet by hand or by a perforated trough above. Such a system employs one of the fundamental principles of air conditioning.

The term "air conditioning" is credited to Stuart W. Cramer, who in 1907 presented a paper on humidity control for textile mills before the American Cotton Manufacturers Association. Control of moisture content in textiles by addition of steam to the atmosphere had long been known as "conditioning." It was not, however, until 1911, when Willis Carrier published the results of many years of research, that the scientific basis of air conditioning was laid.

From air conditioning in industry, for the quality of the product, to air conditioning for comfort was but a step. Comfort air conditioning was developed in the 1920s for the increasingly large theatres, stores, and office blocks, in which the economic value was evident. The development of air conditioning as a major industry, however, awaited new refrigerating techniques.

Until the end of the 19th century, design of heating and ventilating systems was largely empirical. A change to scientific methods began to evolve in the 1890s. Research work was undertaken at the University of Berlin-Charlottenburg and in a few other places before World War I. After the war, research was begun in several countries, notably in the United Kingdom, the United States, Germany, Sweden, and France. Out of this international research effort modern air conditioning has grown.

This article is divided into the following sections:

I. Environmental factors
Man and his environment
Body heat losses
Human comfort
Other influences affecting comfort
Environmental effects on buildings
Heat losses in winter
Heat gains in summer
II. Heating
Generation of heat
Solar energy
Heat pump
Boilers
Media for heat transmission
Air
Hot water
Heat exchange liquids
Steam
Heat emitters
Central, or indirect, heating
Local, or direct, heating
District heating
Control equipment
Hot-water supply (HWS)
III. Ventilation
Mechanical ventilation: applications
Mechanical ventilation: equipment
IV. Air conditioning
Temperature and humidity control
Cooling
Warming and humidifying
Equipment
Noise and vibration control
Associated comfort factors

I. Environmental factors

MAN AND HIS ENVIRONMENT

Body heat losses. The human body, like any other heat engine, consumes fuel (food) to generate heat. The heat leaves the body in the form of radiation and convection from the skin, evaporation of moisture from pores in the skin, and in the breath. Radiation and convection occur only if the air temperature or the surroundings are below blood heat. If approaching or above this condition, the sweat glands expand and heat is rejected solely by evaporation. High relative humidity coupled with high temperature renders life increasingly difficult.

At low temperatures, the blood vessels near the surface of the skin close up, leaving the epidermis dry and acting as an insulator. Shivering is a provision of nature to cause muscular activity and hence increase blood circulation.

Factors in human comfort

Human comfort. Four factors affect comfort: air temperature, radiation, relative humidity, and air movement. Most thermometers measure air temperature but are not influenced by infrared (long-wave) radiation, because their glass envelopes are opaque to such radiation. The human body, however, may receive heat from a radiant source; if so, air temperature must be reduced if body heat release is to remain normal. Conversely, if the body is losing heat by radiation to its surroundings, such as to the cold walls of a room, the air temperature must be raised for an equal degree of comfort.

The study of radiant effects is complex. The mean radiant temperature in an enclosed space is the mean of all the surfaces bounding the space in proportion to their areas and temperatures, the latter being calculable from the known conductivity of the materials and an assumed temperature difference between inside and outside. This calculation gives only an approximation; such factors as the effect of a large area of window, which may be cold in winter and hot in summer, produce variations in different parts of a room. Moreover, the body itself is sizeable, and its response varies according to differences of clothing. The matter is further complicated by the introduction of heated surfaces to warm the space and by the entrance of heat through windows via sunshine.

A close approximation to mean radiant temperature may be obtained by the use of the globe thermometer (Figure 3), consisting of an ordinary thermometer in a

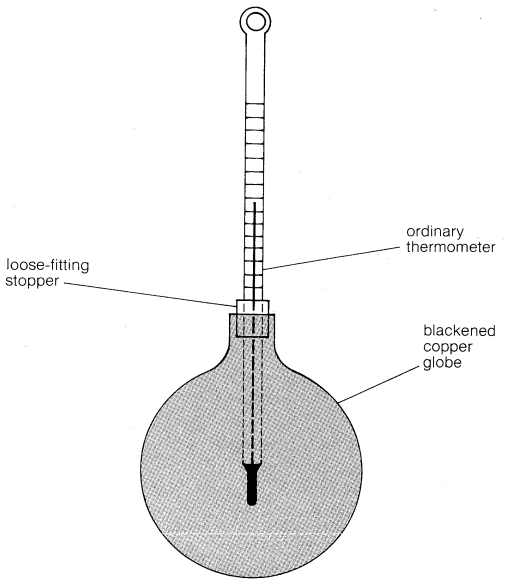


Figure 3: Globe thermometer for determining mean radiant temperature.

blackened, hollow metal sphere about six inches (15 centimetres) in diameter. If suspended in an enclosure in which all surfaces are at the same temperature as the air, the globe thermometer reads the same as an ordinary thermometer, regardless of air movement; but, if the surface and air temperatures differ, the globe becomes warmer or cooler in proportion to the positive or negative radiation falling on it. Given the globe temperature and air temperature, mean radiant temperature may be calculated.

Research in Great Britain on factory workers doing light sedentary work established that globe thermometer temperatures of 62°–68° F (17°–20° C) and air temperatures of 60°–68° F (16°–20° C) were regarded as comfortable. Persons seated at rest required temperatures a few degrees higher, as did those persons who were used to warmer climates.

The above criteria take no account of humidity. In colder climates, such as in northern Europe, variations in relative humidity within the range 40–70 percent have little effect on comfort. In warmer climates, in which sweating plays an important part in heat release, humidity must be taken into account. The effective temperature scale was developed in the United States in a long series of subjective tests in carefully conditioned rooms, from which lines of equal comfort or discomfort—effective temperature lines—were derived. Limits within which individuals feel comfortable may also be charted, establishing what is known as the "comfort zone," which differs for summer and winter.

How humidity affects heat release

The effective temperature scale takes no account of radiation. This can be included simply by using a globe thermometer instead of an ordinary thermometer. The scale so derived is the corrected effective temperature scale, which is no longer widely used. Another scale, resultant temperature, takes account of air temperature, radiation, and air speed. It is measured by a small blackened globe and is more sensitive to air temperature and movement and less to radiation than the globe thermometer. The wet resultant temperature takes account of humidity by using a wet-bulb thermometer—a useful device for determining relative humidity. It consists of a thermometer with the bulb wrapped in cloth kept moistened by a small reservoir of water. Evaporation causes cooling, the amount of cooling depending on the relative humidity.

Another scale is known as environmental temperature; it is weighted in that it is equivalent to two-thirds mean radiant temperature and one-third air temperature. It takes no account of air movement or humidity. Its application is in the study of the thermal properties of buildings.

Other influences affecting comfort. It is possible to enumerate here only a few of the other factors affecting comfort.

Quality of heat radiation. It has long been known that the human body derives greater comfort when the air temperature is lower than that of the surrounding wall and other surfaces. This probably accounts for the comfort of the panelled room common in the 18th and 19th centuries; even though the source of heat was no more than an open fire, the surfaces of the timber panelling were quickly warmed by radiation, while the air drawn in by the large flue ensured a low reading of the thermometer.

Radiant heating makes use of extended surfaces of wall or ceiling heated to some 100° F (38° C) or floor to 75° F (24° C). Air temperature may then be some 5° F (3° C) below that required with a convective system. Unfortunately, the effect of low-temperature radiation on the skin, coupled with the lack of air movement from convection currents, produces an enervating condition. Consequently, it has been suggested that a ventilation system accompany this form of heating to stimulate a freshening atmosphere.

Radiant heat sources employing higher temperatures, such as infrared panels or luminous reflector elements electrically heated, present the problem of directional effects; a certain intensity acceptable on the feet may be uncomfortable if directed at the head. At an air temperature of 50° F (10° C) or under, it is believed impossible to create a feeling of comfort, no matter how great the supply of radiant heat. These two considerations are bound up in one, namely, that as a means of distribution of warmth, very high-temperature radiant heat sources are inherently unsatisfactory.

The effect of radiant heat sources on the human body requires special attention. The great merit of the open coal fire was its wide range and constant variation of radiant wavelength, from the high incandescent parts of the fuel and flames to the dull, low-temperature radiation from the surrounding firebrick and cast-iron or tile decoration. Although the fireplace is decried as an inefficient item of heating equipment, it does have beneficial effects, on a limited scale, room by room. It is open to question, however, whether the effects of various wavelengths of radiation on the human skin and on the stimulation of the senses as a whole have been thoroughly explored.

The advantage of the open fire

The unrelieved monotony of a certain limited range of wavelengths, as from a gas fire or electric heater, can cause dryness and intense irritation of the skin. The human system seems to need stimulation from constantly varying conditions.

Temperature gradients. Because hot air rises, a system of heating relying on warming of air for heat distribution results in a higher temperature at the ceiling than near the floor. An exception is a floor-heating system, in which convection from the floor is balanced by radiation at the floor level.

In the case of a large open space, such as a factory, a system that minimizes temperature gradient is more effi-

cient than one in which the effect is disregarded. Thus, the once common use of overhead pipes produced inordinately high temperature gradients and high roof losses.

Radiant-heating systems in general minimize temperature gradients. Forced-convection systems are preferable to natural convection. Radiators employing extended surfaces below windows give lower air temperature rise than those concentrated on inner walls.

Uniformity of conditions. In an air-conditioned building, air temperatures differ little from those of surrounding surfaces. Thus, a state of thermometric uniformity may exist, except insofar as air movement is concerned. Comfort, as defined by a certain range of effective temperature, is likely to be achieved, but it is possible that the human body continuously so enveloped may suffer from not only physiological but also psychological enervation.

Air motion as a determinant of effective temperature

Air movement. The effective temperature depends on air velocity. The higher the air motion, the lower the effective temperature for a given temperature and relative humidity. An effective temperature of 70° F (21° C) with an air speed of 30 feet (nine metres) per minute, for instance, falls to 60° F (16° C) at an air speed of 600 feet (180 metres) per minute. The variation diminishes with increasing temperature until, at blood heat, air motion ceases to have any effect. It is for this reason that in hot, crowded atmospheres fanning seems to give little alleviation.

Air movement up to 50 feet (15 metres) per minute in an occupied space is usually considered acceptable and desirable, rising locally perhaps to 60 feet (18 metres) per minute and producing some degree of turbulence. There is a much stronger risk of complaints of drafts at higher air speeds.

The problem for the air-conditioning engineer is often to find ways of introducing large quantities of air into a space while keeping velocities in the occupied region within acceptable limits. Raising the relative humidity brings a higher effective temperature, which may allow a slightly higher air movement.

To summarize, it will be seen that the four factors involved in comfort, namely, air temperature, radiation, humidity, and air motion, are largely interdependent. There is no single instrument by which they may be measured collectively, yet the human body can quickly assess whether or not it is in a state of comfort. That state may be the result of any of the possible combinations of these factors.

ENVIRONMENTAL EFFECTS ON BUILDINGS

Heat losses in winter. *Desirable temperatures.* Given the fact that the human body loses heat to its surroundings, it is clear that the prime purpose of heating is to prevent that loss from becoming excessive. Thus, heating involves controlling air temperature or mean radiant temperature or both, within fairly narrow limits depending on activity. A temperature of 55° F (13° C) (lower limit) is appropriate for heavy manual work and 75° F (24° C) (upper limit) for sedentary workers and domestic living rooms. Acceptable temperatures also vary in different countries.

Apart from comfort, many industrial processes demand a strictly controlled temperature for the sake of the product. Printing machinery, for instance, requires a high temperature of 75 or even 77° F (24°–25° C) for the proper flow of ink.

Heat lost by transmittance

Outside temperature. The maintenance of such temperatures internally when outside air temperature is low results in a flow of heat through the fabric of the building. This heat flow outwards is termed the transmittance loss.

For design purposes, it is important to know the general pattern of outside temperature. It is not economical to design for the lowest temperature ever recorded. On the other hand, the average winter temperature is not suitable, because there may be frequent spells of several days or weeks when lower temperatures obtain.

Another factor that affects the choice of design temperature is the "fly-wheel" effect of the building. A massive structure may be expected to experience sudden cold spells of short duration with small effect internally, whereas a light construction rapidly transmits any drop of outside temperature.

Overload capacity of the system is also important under conditions of low external temperature; a system of heating with an inherent overload capacity may meet extremes of cold by drawing on this extra supply. Systems such as those relying on electrical resistances of fixed rating, however, have no such reserve.

Several countries have established basic design temperatures, generally following some law of probability but excluding extremes. In many countries in which prevailing conditions vary considerably from one latitude to another, regional zones of temperature have been established. A few basic design temperatures accepted in various parts of the world are given in Table 1.

Table 1: Design Temperatures for Heating Buildings in Cities of the World

	temperature	
	°F	°C
Auckland, N.Z.	45	7
Berlin	5	−15
Chicago	−10	−23
Cologne	10	−12
Hong Kong	45	7
London	30	−1
Madrid	25	−4
New York City	0	−18
Oslo	−6	−21
Ottawa	−15	−26
Paris	21	−6
Rome	30	−1
Sydney	45	7
Tokyo	28	−2
Washington, D.C.	0	−18

Because of the influence of the Gulf Stream on the climate of the British Isles and western Europe, temperatures rarely drop below 20° F (−7° C), even though the latitude of London is the same as parts of northern Canada in which subzero temperatures regularly obtain. The influence of the great land masses is also significant, accounting for lower design temperature, for example in Berlin. The low temperatures of the Eastern seaboard of the United States are accounted for by the absence of the Gulf Stream and the effect of the land mass to the westward.

Effect of building materials on heat loss

Transmittance through building materials. The rate of heat loss through the structure per degree of temperature difference depends on the conductivity of the materials of roof, walls, windows, and floors. There is also a skin effect, by which the air immediately in contact with the surface constitutes an inert layer offering resistance to heat transfer. This is known as surface resistance and is greater for the still air of the interior than for the exterior, where the effect of wind comes into play. The outside surface resistance varies with different aspects and exposures.

The overall conductance of building material is the transmittance factor, U, and is expressed in British thermal units (BTU) per square foot per degree difference per hour or in watts per square metre per degree Celsius (W/m^2–° C). Researches carried out in various countries have established U factors for a great variety of common forms of construction, or they may be calculated from a formula. Some of the more common U factors, expressed in English and SI (revised metric) units, respectively, are given in Table 2.

Insulation. In the interests of economy it is now common practice to include some form of insulation in building construction. For walls, an inner layer of foamed slag or lightweight concrete block is often used, but such porous materials are generally unsuitable for load bearing, so that the structure must be framed or given load-bearing exterior walls. Curtain-wall construction used in

Table 2: Heat Loss of Building Materials as U Factor

	type	BTU* °F difference	W/m²–°C
Windows	single glazing	1.0	5.7
	double glazing	0.5	2.9
Wall	brick, 9 in. plastered, normal exposure	0.43	2.5
	brick, 11 in. cavity plastered, unventilated	0.30	1.7
	concrete, 6 in. plastered	0.58	3.3
Roof	flat, 6 in. concrete, lined insulation board	0.30	1.7
	pitched, tile-lined boards, plaster ceiling, attic boards	0.2	1.2
Floor	concrete on earth, hard finish	0.2	1.2
	boards on joists and cavity	0.32	1.8

* BTU per square foot per hour.

tall buildings for lightness often incorporates glass fibre slabs or similar material between the inner and outer skins. Similarly, roofs may be insulated in a variety of ways by linings or insulating materials. Ample data exist on the properties of insulating materials, such as avoidance of condensation. Condensation occurs when the temperature of the surface of window, wall, roof, or floor drops below the dew point of the air in the space. In hollow-wall, or cavity, construction, the temperature inside the wall may be such as to cause condensation within the construction. Certain porous materials produce the same effects. Condensation has become a serious problem in blocks of apartments of concrete construction, some of which are afflicted with mold growth on walls and on floors under carpets. The insertion of a vapour barrier is one way of preventing migration of moisture in cavity constructions.

It is usually possible to achieve a U factor of about 0.2 (1.2 in SI units) or less without prohibitive cost and with greatly reduced fuel consumption throughout the life of the building. In many countries insulation is mandatory for industrial and domestic premises, limits being set for the U factor.

Air change. In addition to the transmittance loss through the fabric of walls, roof, or elsewhere, there is an additional loss resulting from infiltration of air through cracks around windows and doors. If the roof is made of some form of sheeting, cracks allow the escape of warm air that collects near the ceiling. Air change also occurs through the opening of doors, particularly in a factory in which goods are constantly entering or leaving. In a tall building there is a natural inflow of cold air at ground floor and an outflow of warmed air on upper floors—the "chimney effect."

Air-change rate

The assessment of air change-rate cannot be scientifically accurate, because it depends on many factors. They include strength and direction of wind, size and type of windows, proportion of window opening area, temperature gradient within the space, and frequency of door opening. Air-change rates usually vary between one-half and three complete changes per hour.

Another approach uses research on windows of various types, with and without weather stripping, and at various wind speeds. The inflow rate particularly on tall buildings may by this means be related to what is the principal means of ingress.

By whatever means the air-change rate is assessed, the heat necessary to warm that air from outside temperature to inside temperature may then be calculated.

Total heat loss. The total heat loss for the room or space in question is then the sum of the transmittance losses and the air-change loss. Some designers add a height factor for lofty structures to allow for temperature gradient and an exposure factor for especially severe conditions.

Intermittent heating. The conventional approach to heat losses discussed above assumes steady state conditions. Heating is, however, frequently intermittent because of night shut-down or limited hours of occupancy, as in a school or church. Unsteady states may be simulated by computer analog techniques but these are of more academic than practical interest.

An allowance for intermittent heating may be made by a percentage addition to the capacity of the entire system, varying between 20 and 25 percent in systems that can only operate at fixed output. If overload capacity is available, however, as by raising temperature above normal in a water system, a lower addition is permissible.

In extremely cold weather, intermittent heating is impracticable and continuous running is essential if a specified temperature is to be maintained. This is because the pre-heat boost necessary to overcome the night loss of the building is beyond the capacity of an economically designed system.

Heat gains in summer. *Fabric gains.* Heat gains are the opposite of heat losses. In summer, external temperature is above internal temperature, and therefore heat flow is from outside to inside. The parts of the surface skin of the walls and roof that are exposed to sun are warmed above air temperature, thus adding to the rate of transmission.

Various methods have been devised to allow for this effect. One is the "sol-air" concept, by which the temperature increment because of solar heat may be estimated according to the nature and colour of the surface, the latitude, and the sun angle at different times of year.

While this factor applies throughout the day on a flat roof, walls of various aspect are subject to solar gain only for limited periods. Furthermore, mass has a delaying effect, so that solar gains undergo a shift in time depending on thickness and material of construction. In a building with many floors and many rooms per floor, it is necessary to treat each compartment separately to arrive at the heat load in each case, but the coincident total load will not be the sum of the individual room loads.

Heat gains through the fabric are calculated in the same manner as heat losses, using U factors and assumed temperature differences. Insulation is even more important for economy of cooling than of heating because the cost of cooling is usually much higher.

Air-change gains. In cooling as in heating, allowance must be made for air change, because warm air enters by infiltration and by door opening. The chimney effect is negative; that is, with the temperature cooler inside than it is outside, ingress is greatest at higher levels and the entrances at ground level act as outlets.

Glazing. Solar gains through glass in windows and roof lights demand separate consideration. They may constitute by far the heaviest cooling load because glass is transparent to the shorter wave lengths of solar radiation.

Shading of glazed areas

Such gains are so great that shading is generally essential for economy as well as to reduce glare. Such shading takes a variety of forms. One form is architectural, comprising louvred blades fitted externally and sometimes arranged to swivel according to time of day. Alternatively they may form a distinctive pattern resembling an egg crate. In certain latitudes in which the sun is mainly overhead, projections at each floor in the form of balconies may be adopted. Alternatively, external roller blinds may be used. Heat intercepted by any form of external protection is either re-radiated or lost by convection to the atmosphere.

Internal shades are less efficient as they become heated themselves and convert shortwave solar radiation into long-wave dull radiation, to which glass is opaque. In addition, convection currents carry heat into the room. Venetian blinds behind single glazing offer no reduction in solar gain; in fact they increase it slightly. A venetian blind in the interspace of double glazing is more effective.

Another method of reducing solar gain through windows is the use of heat-absorbing glass. This may take a variety of forms, some giving a greenish tint and some incorporating finely divided metallic particles, such as of copper or gold. This has the object of reflecting much of the incident radiation back into space but may be a source of complaint from neighbouring buildings because of the glare of high reflectivity.

Solar heat gains through glass vary throughout the day and for different seasons, as mentioned earlier in relation to walls. There is no time lag with glass as there is with walls.

Sun-path diagrams. The study of solar gains is assisted by use of a sun-path diagram drawn for the particular latitude. This diagram presents in graphic form the motion of the sun, its altitude and azimuth for various times of the year and hours of the day. When a plan of the building, correctly oriented, is superimposed, the angle of incidence on the various faces as well as the periods when they are subject to radiation can be noted.

Incidental heat gains. Throughout the year buildings have incidental sources of heat within their enclosed spaces. These derive from occupants, lighting, and machinery. In residential property, stoves, television sets, washing machines, and other appliances provide a source of heat gain.

In winter these gains are beneficial in reducing the amount of heating required from other sources, but in summer they raise temperatures internally and may require increased ventilation; in air-conditioned buildings they add to the cooling load.

Heat gains from machinery

In offices the greater use of accounting machines, card-punch systems, computers, and other electrical equipment, together with the present high standard of lighting, is responsible for rendering maintenance of satisfactory working conditions increasingly important.

Electrical equipment of all types generates heat. In effect all electrical energy supplied within a space is converted into heat except potential energy that is conveyed outside, as in pumping water or in a conveyor. Machine tools generate heat at the cutting head, and the product becomes warm and gives up heat to its surroundings. In all motors and machines, there is friction that generates heat.

All heat released within a space constitutes heat gain. Means of removing it are discussed below (see below *Air conditioning*).

II. Heating

Early methods of heating, as has been shown, consisted chiefly of burning fuel within the habitable space, the exception being the Roman hypocaust. In the home the fireplace was the central feature of the room. In other buildings braziers, and later enclosed stoves, were used. In cold regions, such as parts of Europe and North America, the stove continues in use to this day.

The early 19th century, with its increased population in the industrial countries, brought a demand for buildings of much greater size, such as assembly halls, factories, churches, schools, chapels, and hospitals, for which some method of heating other than local means was a necessity. Out of this need evolved what is now called central heating.

The essential components of a system of central heating are an appliance in which fuel may be burned for the generation of heat; a medium conveyed in pipes or ducts for transferring the heat to the space or spaces to be heated; and an emitting apparatus in the spaces for releasing the heat either by convection or radiation or both.

Steam heating has long predominated in the North American continent because of its very cold winters. It is here that the vacuum and vapour systems originated. For the home, ducted warm air has been by far the most common system. Hot water is, however, beginning to find a place.

In Great Britain and much of the European continent, hot water has become the favoured method of heating; steam heating is now rare except in industry. Ducted warm air, except for recent developments such as gas-warm air, has never been popular.

Most other countries have adopted the pattern of either the American or the European method, generally according to the cultural influence that has predominated.

While heating in winter is the sole need of many types of building, there is a tendency for air conditioning to supplant it, giving year-round control of temperature and humidity as well as ventilation.

GENERATION OF HEAT

The combustion of fuel is a molecular process in which the principal elements, carbon and hydrogen, react with oxygen from the atmosphere. The heat produced in this way arrives at the walls of a combustion chamber to be transferred to a medium such as air or water. The equipment is so arranged that the heated medium is constantly removed and replaced by a cooler supply—*i.e.*, by circulation.

If air is the medium, the equipment is called a furnace, and if water is the medium, a boiler or water heater. The term boiler more correctly refers to a vessel in which steam is produced, and water heater to one in which water is heated and circulated below its boiling point.

Coal or coke, oil, or manufactured or natural gas may be used as fuel. Solid fuel is stoked either by hand at intervals or mechanically by an automatic stoker. Ash and clinker resulting from the process are removed by hand or, in large plants, by mechanical means. The dust and dirt produced by solid fuel, and the labour it requires, have led to its decline.

Advantages of oil and gas

Oil and gas require no labour except for occasional cleaning, and they are handled by completely automatic burners, which may be thermostatically controlled. There is no residual ash product for disposal. Gas requires no storage. Oil is pumped into storage tanks, which may be located at a distance from the heating equipment.

Another source of heat for central heating is electricity. The relatively high cost can be reduced by off-peak use of current—that is, at night, when normal demand decreases. One method is by thermal storage: water in large insulated cylinders is raised to a high temperature at night, and the heat is dissipated throughout the following day in the heating apparatus in the building. The boiler takes the form of heating electrodes in a cylindrical vessel. Other methods of using electricity for heating are described below in the sections on direct heating and the heat pump.

Solar energy. The sun is the ultimate source of all energy on earth, whether as stored energy in fuels fossilized ages ago, as water power, or as nuclear energy. The sun's radiation, of which a minute portion is received on Earth, if converted at 100 percent efficiency into work at the Earth's surface, would yield about one horse power for each square yard or metre of ground. But in winter, solar energy is at minimum intensity in temperate climates because of the low angle of the sun. Furthermore, clouds interfere. Nevertheless, solar heating for warming and hot-water supply is used on a domestic scale in Israel, southern France, parts of the United States, and elsewhere.

In one form of solar heating the collector comprises a black metallic surface with piping attached, covered by double or triple glazing and backed by insulation. The collector is placed on the ground or on a roof, angled to be normal to the sun at midday or rotated to preserve normality. Water circulated through the piping of the collector is kept in a storage vessel with as little mixing as possible to preserve a high temperature. The stored water is then circulated in the heating system throughout the day and night as required.

Heat pump. The principle of the heat pump, or reverse heat cycle, was first enunciated by Lord Kelvin in 1851, but little note was taken of it until much later. The heat pump is in effect a refrigerator in reverse: a refrigerator produces cooling in an evaporator and rejects heat by means of a condenser, but in the heat pump low-temperature heat is supplied to the evaporator, and energy is added by the compressor, thus making the heat available at a higher temperature. Heat may be removed by air or water from the condenser for warming or water heating. The heat so derived is made up of the low-grade heat received by the evaporator plus the heat equivalent of the power input to the compressor. The ratio of heat input to heat output is termed the advantage and may be 3:1 or 4:1 depending on temperature difference. The closer the output temperature is to that of the low-grade source, the greater is the advantage. Thus the heat pump performs best when heating, say, a swimming pool.

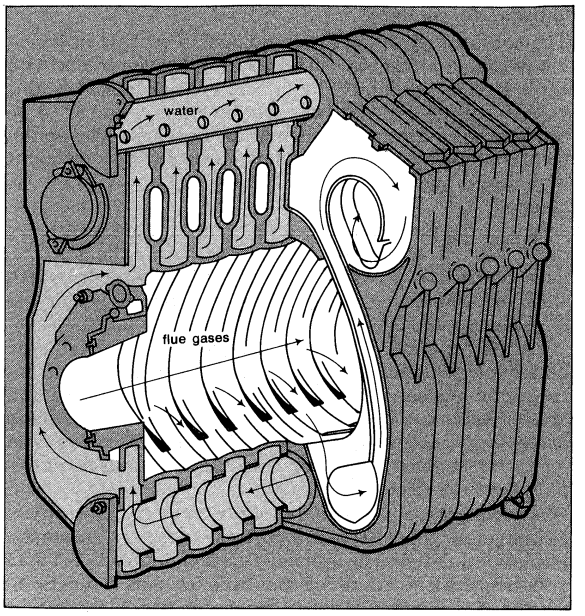


Figure 4: Modern oil-fired, cast-iron sectional boiler. Arrows show direction of flow of gases and water (see text).
By courtesy of Ideal-Standard

Air conditioners that contain a compressor for cooling in summer may, by reversal valves, become suppliers of warm air in winter. Many commercial heat-pump units employ this method. In Switzerland heat-pump installations deriving electric power from hydroelectric sources were extensively adopted during World War II because of the shortage of fuel. In some U.S. cities heat pumps have been combined with cold-storage plants for heating buildings and city districts.

Other sources of low-grade heat that have been used with heat pumps are pipe coils buried in the earth, sewage, rivers, and cooling water from industrial processes. Unfortunately the high cost of the compressor plant and ancillary equipment usually renders the economics of the heat pump unattractive. Furthermore, the apparent advantage of using electrical power derived from a heat engine is largely cancelled out by heat rejected at the power station so that from a national or regional fuel-economy point of view little has been gained.

Boilers. Boilers solely for heating, as distinct from steam boilers for power, date from the early years of the 19th century. The earliest forms were simply made of wrought-iron plates rivetted together and set in brickwork. The tubular boiler for high pressure was followed by tubular heating surfaces for low pressures. All of these types of boiler were of low efficiency, probably not much over 50 percent, but coal was so cheap that this was of little account.

The cast-iron sectional boiler, first developed in the latter part of the 19th century in the United States, was cheaper and more efficient than the earlier forms. It could be added to or replaced in part, if required, and could be handled inside a confined space more easily than a one-piece boiler. Sectional boilers have survived to this day with little alteration; a modern form is shown in Figure 4. Larger duties may be met by using several in battery form.

The shell type of boiler

Because there is a limit in pressure and size for cast-iron boilers, the trend has been to the shell type boiler for high pressures and duties of greater magnitude. This type is formed of a cylindrical shell with furnace tube and fire tubes or a completely tubular form. While the principle of this design is old, modern applications achieve a high efficiency. Other types in common use include the firebox and smoke-tube form, similar to a locomotive boiler.

One of the hazards of boiler operation is fireside corrosion resulting from burning sulfur-bearing fuels such as coal and oil, which, with condensation, produce sulfuric acid. Precautions with regard to operating temperatures are necessary to prevent excessive corrosion and early failure.

Another hazard is internal furring, or deposition of scale-forming salts from the boiler water. This is more troublesome in steam-generating plants than in water heating on a closed circuit but may be met by suitable water treatment.

MEDIA FOR HEAT TRANSMISSION

Air. Air as a means of conveying heat over a distance suffers from the disadvantage of low density and low specific heat (a measure of the amount of heat that can be carried by a cubic foot of air), as compared with water, and is therefore confined to local distribution covering only short distances.

The so-called pipeless heater at one time common in the cold countries of Europe and Canada used air for heating houses and apartment buildings. It consisted of a coal-fired furnace in the basement; a cast-iron combustion chamber and flue surface contained in an enclosure or casing admitted return air from a grille in the floor or hall or staircase and delivered the heated air by natural buoyancy through a second grille, in such a way that it permeated the whole space. Its effect, so far as the rooms off the central space was concerned, depended on interior doors being left open.

In more recent developments, the air flow is forced by a fan, and the burner in the furnace is oil- or gas-fired. Air from the heater is conveyed through ducts to the various rooms (see Figure 5).

Another method using air for conveyance of heat is the plenum system already mentioned; once widely used for the heating of factories, it has been largely superseded by other methods. In this system air is, in effect, a secondary medium, the primary being steam or hot water. The system comprises a fan and a heater battery consisting of a series of pipe coils. The heated air is discharged by the fan through a system of ducts. The disadvantage of the system is its cumbersome ductwork and lack of flexibility.

Heated air has been used in more recent developments as a medium to provide radiant heating for industrial use. This consists of a closed system of overhead ducts with bare surfaces arranged for emission of radiant heat, while the transmitting ducts are insulated. A disadvantage is that there is a limit to the area that may be covered by such an arrangement because of the low volume-to-heat ratio.

Adapted from Honeywell, Inc., *How to Get the Most from Your Heating and Cooling Dollar*

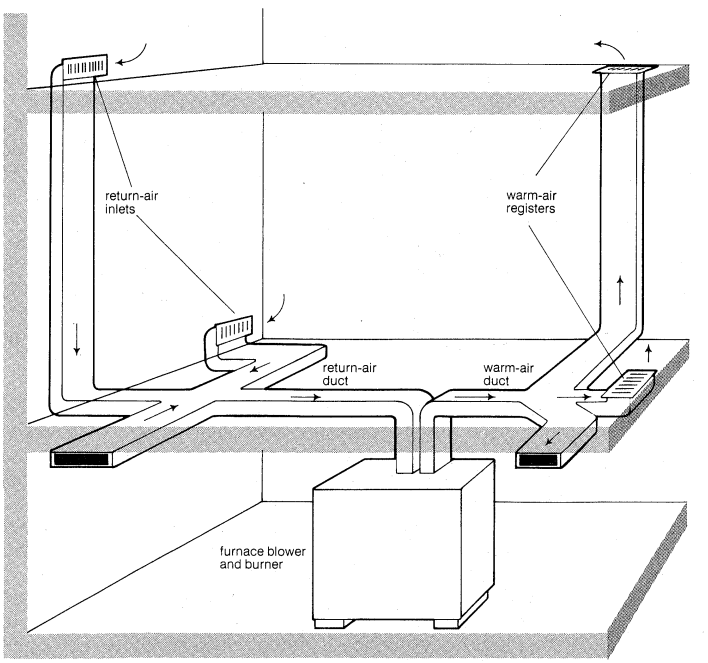


Figure 5: Perimeter forced warm-air system.

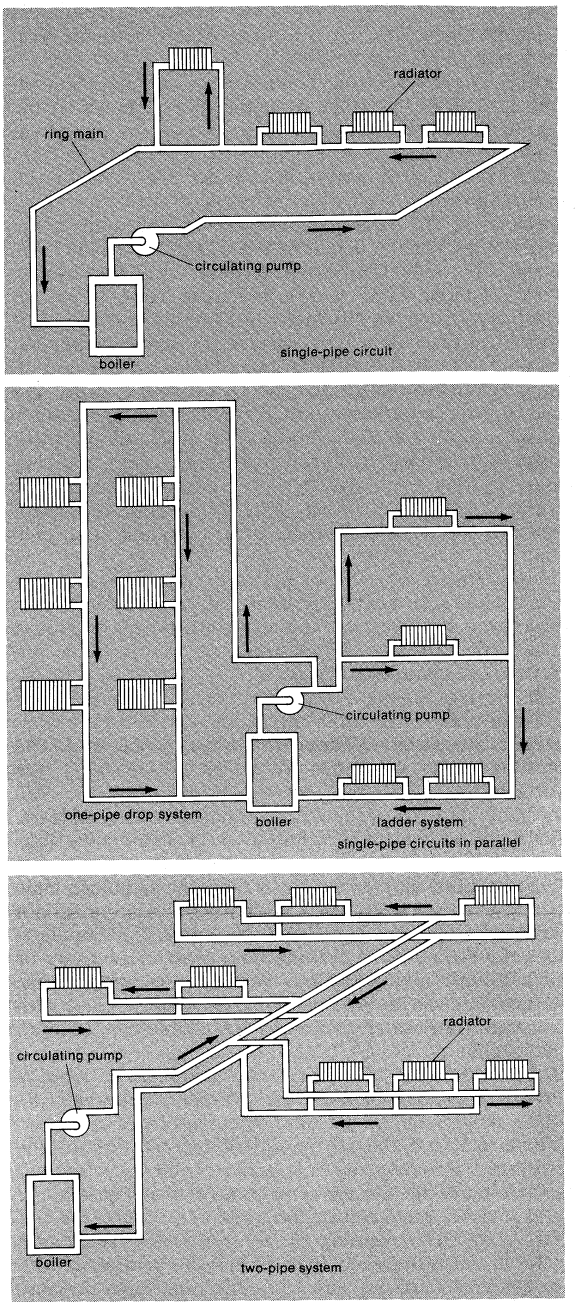


Figure 6: Hot-water systems with pump circulation (feed and vent pipes not shown).

The direct-fired air heater

Another recent development, chiefly for the heating of large factory spaces, is the direct-fired air heater. This comprises a combustion chamber fired by gas or oil, with a flue to the outside, and a casing with a fan to draw air from the working space. This air flows over the heated surface of the combustion chamber and then is discharged at considerable velocity back into the space. These units are made with large outputs and have the merit of flexibility to meet changes of plan.

Air is used as a heat-conveying medium in the new Anglican Cathedral, Liverpool, in which ducts below the floor convey warm air from a heater; after transferring heat to the floor the air is returned for reheating. In effect this is the Roman hypocaust system with air in place of the hot flue gases.

Hot water. Hot water is the most versatile of all the practicable media. It has the greatest specific heat at normal temperatures of all known liquids; it is stable after limitless cycles of heating and cooling; and it is cheap. Its one disadvantage is its low boiling point of 212° F (100° C) at atmospheric pressure; to achieve higher temperatures, the system must be sealed so that the water can be placed under pressure. In this condition temperatures up to 400° F (200° C) can be achieved.

For heating a building, the temperature of emitting surfaces must be adjusted to suit the weather. Hot water is most flexible in this respect, being easily controllable from minimum to maximum either by regulation of boiler temperature or by mixing return with flow water.

The hot water is circulated through pipes by thermosiphon (gravity) or by mechanical pump. The former was normal in earlier systems because an electrical supply was not often available; the pipes were necessarily large, and circulation was often sluggish. Various self-accelerating systems were devised to improve circulation. One, the Reck, made use of steam from a boiler injected into the rising column of hot water, which was thereby lifted, creating a considerable imbalance with the cooler return water columns. The differences in weight of the two columns was the force creating circulation.

Electricity has now virtually monopolized the pumping function. A small home pump with submerged rotor (rotating part of the motor) relies on a nonmagnetic shield such that the windings of the stator (stationary part of the motor) are in air, while the rotor is submerged in the water being circulated. The pump is silent and applicable to the smallest systems. On a larger scale circulating pumps take the normal form of standard centrifugal pumps. Figure 6 illustrates single and two-pipe, low-pressure, hot-water systems with pump circulation.

The low-pressure system

In a low-pressure system the water expands on heating and is allowed to rise through a pipe into a tank above the level of the highest part of the system. On cooling, the same water is returned.

An alternative method of accommodating expansion water takes the form of a closed vessel connected to the system near the boiler; this vessel contains a flexible diaphragm, above which is nitrogen under pressure. As the water in the system expands, the diaphragm rises, compressing the nitrogen still further and increasing the pressure (see Figure 7).

At low pressures, below the boiling point of water, the temperature drop from flow to return in a system is limited to some 30° or 40° F (17°–22° C); if the drop were larger, the average temperature of the radiator would be too low for economical heating. For more extensive systems, water under pressure may be used. High pressures (up to 10 atmospheres with temperatures to 366° F [186° C]) are used for industrial purposes.

If water at a temperature above atmospheric boiling point (212° F or 100° C) is used, pressure may be produced by steam, as in a steam boiler; the water, which is circulated by pump, is taken from and returned to the water supply below the steam cushion in the boiler. Normal water gauge fittings are used on the boiler, which is only partly filled with water, the level rising with expansion.

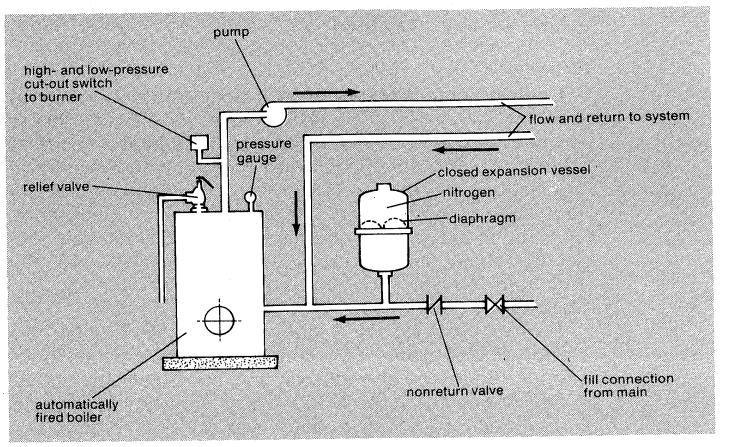


Figure 7: Low-pressure, hot-water system with closed expansion vessel.

Alternatively, in medium- and high-pressure systems, pressurization may be by a gas cushion using an external pressure vessel. Because higher pressures are usually associated with larger systems, the water of expansion is allowed to flow off into a separate tank, from which it is returned to the system on cooling by pump.

Heat-exchange liquids. Besides having a low boiling point unless under pressure, water is subject to expansion on freezing, with the danger of bursting pipes and boilers. Various heat-exchange liquids that avoid the problem have been developed for use in the chemical-engineering industry. These liquids have a boiling point in the region of 500° to 600° F (250°–300° C) at atmospheric pressure, but have the disadvantages of high cost, strong odour, and toxicity on leakage. So far, water has not been supplanted in spite of its disadvantages.

Steam. The latent heat of steam (the heat required to vaporize water to steam, or conversely, the heat given up by steam when it condenses to water) is used for heating by condensation within the heat-emitting surface. Latent heat of steam at atmospheric pressure is high, and steam may be conveyed in pipes at considerable velocity. Much less heat is stored in a steam system than in a hot water system, so a steam system heats up and cools down faster than a water system.

Inflexibility of steam

The disadvantage of steam lies in its inflexibility to meet variations of weather conditions. Hot water may easily be varied from about 100° F (38° C) upwards, whereas steam is constant at the temperature corresponding to pressure. Attempts to make steam more flexible have made use of sub-atmospheric or vapour systems, forming a vacuum within the system so that temperatures well below atmospheric boiling point may be achieved in the heat-emitting apparatus. Such systems, however, depend on a variety of devices calling for regular maintenance, and they are prone to be noisy.

When steam (or vapour) condenses in heat-emitting apparatus, it returns to the liquid state and must be removed quickly. For this purpose steam traps, allowing water but not steam to pass, are fitted on the outlets. The water, or condensate, is then returned by gravity to a collecting point, from which it is pumped back to the boiler. Some of the condensate re-evaporates when it is released from the traps at atmospheric pressure, forming flash-steam. This constitutes a loss, and means to employ flash-steam usefully are desirable. Steam thus has certain inherent losses, which may amount to as much as 10 percent.

Steam as a medium of heat transmission is now mostly confined to industry, where its use is required for other purposes than heating. Even so, high-pressure hot water in certain industrial processes is supplanting steam. Steam is also used in hospitals and similar institutions in which laundry, kitchen, and sterilizing require it. Heating in these cases is usually provided by the use of calorifiers, or heat exchangers, in which steam, the primary medium of distribution, heats water for distribution at lower temperature in the heating system. Similar arrangements supply hot water for baths, basins, sinks, and other facilities.

HEAT EMITTERS

Central, or indirect, heating. Heat emitters in central heating systems may be classified in two broad groups: primarily convective and primarily radiant. The first group may be sub-divided into those relying on natural convection and those relying on forced convection. Radiant systems may be divided according to temperature. Assuming hot water is the medium, the classification becomes: high temperature, 250°–350° F (120°–175° C); medium temperature, 140°–180° F (60°–80° C); low temperature, 100°–130° F (40°–55° C).

Primarily convective heating. The most familiar example in this group is the common radiator, which emits about 20 percent by radiation and 80 percent by natural convection. The term radiator is obviously a misnomer.

The earliest form of radiator was the box coil, consisting of a series of cast-iron pipes in a perforated casing. Later in the 19th century cast-iron sectional radiators came into use, often elaborately decorated. Radiators are now much lighter, with small water content; steel has largely replaced cast iron.

The natural convector comprises a heating element, with fins, firmly attached to a pipe through which the heating medium circulates. The element is placed at the bottom of a casing having an opening at the base and an outlet at the top. A flue effect is created by the rising column of warm air, drawing in air from along the floor of the room and discharging it heated at the top. The greater the height of the casing, the greater the flue effect, and hence the greater heat output for a given size of element. There is little radiation from the casing, which is warmed only by the air it contains.

Another form of heating surface in which convection predominates is the gilled (finned) pipe, often used in the past for heating in factories. No casing is used, and heat transfer is the result only of the convection currents set up in open air. Plain pipe surface also falls into this category. The large cast-iron pipes now used for greenhouse heating were at one time a common method of overhead factory heating. The convection-to-radiant component varies according to temperature and pipe size.

Forced convectors. The application of a fan to finned heating surfaces greatly increases the heat emission rate for a given surface area. This feature has been used in a great variety of ways, all of which are 100 percent convective with no radiant component.

Positioning of units

Unit heaters ranging in size from 100,000 BTU per hour (30 kilowatts) to 1,000,000 BTU per hour (300 kilowatts) or over, are commonly used for industrial heating. The units may be suspended overhead, discharging the warmed air towards the floor. In another form they are floor mounted, discharging horizontally above head level. Thermostatic control is achieved by switching the fans on and off.

Incorporating slow-speed silent fans, the same principle is used extensively for such purposes as heating school classrooms, entrance halls, and canteens. In such applications the element and fan are contained in a casing of cabinet type, giving a large output in a small space. Miniature forms are also available for domestic use with ratings on hot water as low as 6,000 BTU per hour (2 kilowatts).

Radiant heating. Emission of radiant heat from a surface is proportional to the difference between the fourth powers of the absolute temperature of the emitting surface and the mean radiant temperature of the enclosure. Emission by convection follows a simpler law; as emitting surface temperature is increased, the ratio of radiation to convection rises.

For economical heating of a large space, such as a factory, temperatures as high as may be consistent with reasonable cost of pressure parts are used. High-pressure hot water is particularly suitable in the range 300° to 350° F (150°–175° C), giving surface temperatures of 250° to 300° F (120°–150° C).

A radiant heating surface may consist of flat steel plates with a gridiron form of pipe coil welded to the back. The back may be insulated to give heat emission on one side only, or may be bare giving emission on both sides. In industrial installations, panels are slung well above head level, and sometimes inclined at an angle towards the floor. A bay width of 80 feet (24 metres) may easily be covered by such an arrangement. The advantage of radiant heating over unit heaters is that no moving parts are involved, and maintenance is easier.

Another form, strip heating, makes use of a continuous horizontal strip, with piping attached, which extends the length of the space to be heated, supported from the roof members. In some systems this strip is combined with lighting.

The surface colours of radiant heaters are unimportant except that they should not be metallic.

The medium temperature range is served by low-pressure hot water, usually between 140° and 180° F (60°–80° C). Because radiation is less intense, the system may be applied to buildings of many types.

Panel heating surfaces

Heating surfaces in the form of flat panels of steel or cast iron are used on walls and ceilings. They have the merit of cleanliness, though wall staining may result from the convection currents. There are numerous variants that increase convection, for example, finned surfaces at the back.

The panel system is commonly used to heat the ceiling, the radiant heat coming downward. Walls and floors have also been used. The system has the merit of being invisible and of avoiding convection current staining.

Water temperatures in the range 100° to 130° F (40°–55° C) are used. Coils of piping embedded in the structure warm the floor, wall, or ceiling (see Figure 8). In

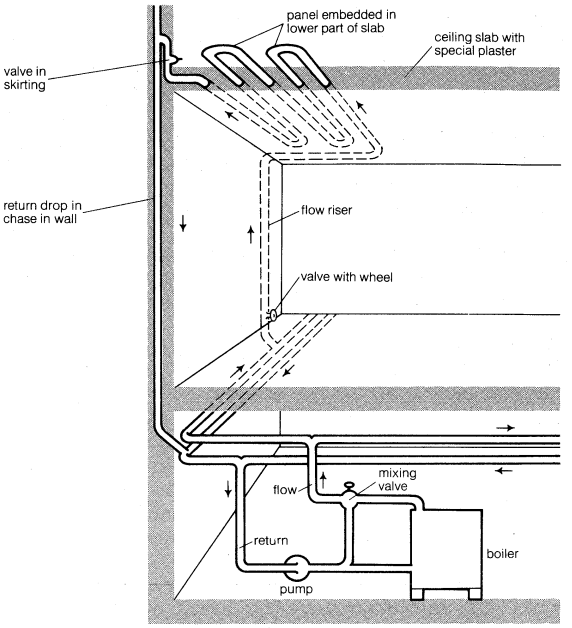


Figure 8: The panel system, consisting of hot-water pipes in ceiling.

the floor, a surface temperature of 75° F (24° C) is considered the maximum desirable for comfort, but in the ceiling, closer spacing of coils may be used giving surface temperatures of the order of 90°–100° F (30°–40° C).

For acoustical reasons, perforated tiles of metal or plaster, heated from above by coils fixed in the void space and covered by a thermal insulating blanket, have come into use. Higher temperature water is used than with the panel system. This system responds to control faster than one in which coils are embedded in the structure.

For an equal state of comfort, a system that is mainly or entirely convective must provide a higher air temperature in the heated space than one that relies mainly on radiation. This is because the cold surfaces of walls and windows are removing from the body heat that in the absence of radiation can only be replaced from the surrounding air.

Lower air temperatures, possible with radiant heating, mean lower air infiltration losses and lower temperature gradients from floor to ceiling. Thus radiant systems save fuel, unless heating is intermittent, as in a school. In that case the fully convective system is appropriate because of the rapidity with which air temperature may be raised in the morning and lowered at night.

Local, or direct, heating. The methods described above constitute central heating, which may otherwise be described as indirect. Local heating is direct—that is, heat is released within the space to be heated with no intermediary fluid for transmission. Such systems include electric heating, gas fires, gas convectors, gas-warm air, oil-warm air, oil stoves, and solid-fuel fires and stoves.

Electric heating. Electricity used at peak hours is generally expensive and confined to spot heating or to small buildings in which a central system is not practicable. In domestic applications the electric heater is a substitute for the now fast-disappearing coal fire.

Electric heating appliances in all cases, except one that will be discussed later, depend on resistance wire elements in one form or another, either bare or incorporated in a nonmetallic sheath that becomes heated. Elements are enclosed in tubes to give a lower surface temperature, or in cabinets to form convectors. Forced convectors include a fan to deliver air over the elements. Radiators containing oil are heated by elements in the base operating at temperatures similar to their hot water counterparts.

Radiant electric heating is either luminous or dull. The former includes the electric fire with various forms of reflector, and overhead units projecting a beam to floor level over a limited area. Dull emitters comprise panels of ceramic or other material, in which the elements are embedded, working at a surface temperature in the region of 160° F (70° C). In another form wire elements are embedded in wallpaper applied to walls or ceilings.

A recent development departs from wire resistances by using finely divided carbon molded into plastic sheet in such a way that, through electrical contact strips at the edges, the whole sheet becomes the heating element.

Because off-peak electricity is generally available at much lower rates, thermal storage has been developed for heating buildings during the day. In one method, use is made of the floor as the storage medium by embedding resistance elements in it. The surface temperature is limited for comfort to about 75° F (24° C), which in practice declines during the day; to be successful this method requires a midday boost.

The night-store heater

Another method of storage is the night-store heater. The storage material (ceramic or concrete) is raised by embedded elements to some 400° F (200° C). The outer casing, which is made of steel and insulated, gives a surface temperature of about 180° F (80° C) at the end of the charging period, declining to 100° F (40° C) at the end of the day. A variant of this includes a fan that delivers air through passages in the storage block, enabling output to be controlled. This latter principle has been developed further for use in apartments and residences by adding ducting, so that a number of rooms may be served from a single central unit.

Gas. The burning of gas within the space to be heated requires a flue, the provision of which restricts the system to buildings in which a rise through the roof or through an outer wall can be arranged. The presence of a flue, however, also assists in ventilation.

Old types of gas fire gave an efficiency of no more than 45 percent because of the loss via the flue. Modern types, arranging convection passages around the heated surfaces, increase efficiency to about 60 percent.

For the heating of factories and large spaces, overhead gas-burning radiant units, which use incandescent refractory plates, have been developed. Dull radiant heaters are also used; in this case the gas heats a refractory plate in front of which a metal surface forms the dull emitter.

Gas convectors are now usually of balanced draft type; that is, they receive combustion air from and discharge products of combustion through the same wall. Provided with a fan on a larger scale, they become forced-draft convectors to cover extensive areas.

Gas warm-air. This is a system developed for multistory apartments or single dwellings. Gas is consumed in a combustion chamber, from which a flue takes the products of combustion to the outside. A fan draws air from the space to be heated and delivers it over the outer surface of the combustion chamber, from which it is conveyed by short ducts to the various rooms of a dwelling. Fast heat-up and response to thermostatic controls render this system economical and acceptable, particularly for a household in which heating is required primarily in the morning and evening.

Oil stoves and oil warm-air. Portable oil stoves burning kerosene present a fire hazard, and represent only temporary means of heating. Oil-fired warm-air heaters have been developed on similar lines to gas, often fed

with oil via a meter from a central source. The burner is automatic as with gas, and also requires a flue.

Solid fuel heating. During the period since World War II, much research has been devoted to improving the open fire, still widely used in some countries. Efficiency has been raised, excess ventilation via the flue reduced, and all-night burning rendered possible. Anti-pollution legislation has restricted the open fire increasingly to smokeless fuel, such as the manufactured briquettes or low-temperature carbonization cokes. In rural districts wood is widely used and in some countries, such as Ireland and Scotland, peat.

The stove has already been mentioned as a customary means of domestic heating in cold countries on the continent of Europe and in North America. Versions using anthracite, burning for long periods without attention, were popular at one time, but declined with the rising cost of such fuel and the growing preference for other labour-saving methods.

District heating. District heating, a logical development of central heating, supplies heat for many buildings in a district or a complete town from mains laid in the street. The first such installation was at Lockport, New York, in 1877. Others followed in New York City, Detroit, and elsewhere, operated by steam. Recent technology favours greater use of hot water. There has been extensive development of district heating in Germany, the Soviet Union, France, and Denmark. In many systems heat has been derived as a by-product of electric power generation, thereby raising the overall thermal efficiency of a station from 25 or 30 percent to 50 or 60 percent. In some cases, notably in Hamburg and in the Pimlico Scheme, London, large, hot-water storage vessels have been included to level out the peaks and valleys of input. In Reykjavik, Iceland, hot water from thermal springs is piped to all buildings, and is run to a drain after use.

Various forms of conduit have been adopted for enclosing the mains of district heating systems, a recent one being the "pipe within a pipe." In this type, the pipe lengths are factory-insulated and enclosed within an outer protected steel or asbestos cement tube, the whole being buried in an excavation without structural duct of any kind.

Primary distribution from heat stations is usually in the form of high-pressure hot water circulated by pumps: it is reduced in temperature by heat exchangers at substations from which smaller areas are served. The apparatus in the buildings or dwellings is then conventional in form. Industrial areas may be served at higher temperatures.

Advantages of district heating include the avoidance of a multiplicity of small heating boilers, the reduction in pollution that results from the use of higher stacks, and the utilization of flue-gas cleaning facilities that would not be economical in small boilers, thus permitting the use of cheap fuel, including town refuse. There is also a saving of the traffic that would be involved in retail fuel deliveries.

The total energy system

A recent concept is the total energy system, which serves all of the heating, cooking, and electrical requirements of a group of buildings from a single source, such as natural gas. Electricity is generated in gas turbines, and the waste heat recovered to heat water. The hot water is used in winter for heating and in summer for cooling by an absorption refrigeration plant. The efficiency and economic advantages of such a system depend on a careful appraisal of relative year-round heat and power demands.

Control equipment. If a specified temperature is to be maintained, the input of heat to a space must be equivalent to the heat losses: thus, some means of control is necessary. Most systems of heating, whether direct or indirect, are governed by some form of thermostatic control, designed to maintain the air temperature set by a thermostat, which regulates the input of heat. In the central system, control is often by water temperature, the adjustment being made according to outside weather. Alternatively, control may be by room thermostat, coupled with a clock switch to change over to a lower setting at night. Electrical night storage systems are controlled by an anticipatory method, regulating the period of night charge according to the trend of external temperature.

Thermostatic control has the object of economizing heat as far as possible while avoiding underheating or overheating. Internal heat gains and solar gains may often contribute significantly to the total daily heat requirement of a building and, superimposed on the normal designed output of a heating system, can lead to uncomfortably high internal temperatures. A quickly responsive system is therefore desirable.

HOT-WATER SUPPLY (HWS)

The use of hot water from hot mineral springs for bathing dates from ancient civilizations. Water in shallow pools heated by the sun was also doubtless used. The Romans were the first to heat water artificially for bathing, using a furnace below a cistern; examples are found at Pompeii, Verulamium (Britain), and in the baths of Caracalla in Rome.

After the Romans no progress was made in heating water until the Industrial Revolution. In factories in which steam was available, it was injected "live" into water to be heated. In wealthy houses water was heated by a boiler forming part of the cooking range. In England in 1846, municipal bodies were first allowed by Act of Parliament to build public baths and washhouses.

Modern methods of providing hot water may, as with heating, be classed as either local or centralized. The local unit is fixed in the bathroom or kitchen adjacent to the draw-off points. The central system delivers hot water from a distance.

Hot water heaters

Local systems. These derive their heat from gas or electricity and are generally of small capacity. The gas heater may be instantaneous (*i.e.*, heating the incoming cold water as it flows through), or the heater may be packaged with a vessel to store a larger quantity of hot water for immediate use. The heater in this case is thermostatically controlled. The electric heater makes use of an immersion element, with thermostat, inserted in a storage cylinder or other vessel. No flue is needed.

Central systems. Installed in a residence, a boiler may form part of a cooking range fired with solid fuel (a back-boiler), or the boiler may be independent and fired with coke, oil, or gas. It is connected by circulating piping to a storage cylinder from which draw-off piping is taken to serve the various taps. The cylinder is supplied with cold water from a tank in the roof, served from the main through a ball valve. There is now a preference for the indirect cylinder in which water circulation from the boiler is on a closed circuit. In many countries, the cold feed is taken directly from the city main and the heating element is formed by a coil in the boiler.

In large installations, such as those for a hospital or hotel, the water is heated in calorifiers by steam or high-pressure hot water. The hot water is stored in cylinders from which secondary flow and return pipes distribute hot water throughout the building. Circulation is by pump. A water temperature of 140° F (60° C) is normal for most purposes but for kitchen use and sterilizing crockery and plates 180° F (80° C) is required. The extra temperature is usually added by a booster.

Solar water heating. Heating is usually necessary in a building only in winter; a hot water supply system usually requires heat in both summer and winter. As solar radiation is at a maximum in summer, it is more feasible for hot water supply than for direct heating. A number of examples exist in warmer climates. The water heated in the collector circulates to a storage vessel, devised to prevent undue mixing of the hot and cold water. Auxiliary means of supplying heat are needed for use on sunless days.

District water heating. If a supply of heat from district mains is available, hot water is heated from the same source. Local heat exchangers are provided, served from the mains.

III. Ventilation

The word "ventilation," from the Latin *ventus* or wind, means either the supply of fresh air or the circulation of air in a room. Life, of course, depends on oxygen. The supply of air to an enclosed space involves the removal of a corresponding volume of expired or vitiated air, possibly laden with smells, heat, noxious gases, or dust resulting from industrial processes. In such a case, air is a vehicle or scavenger of pollution. In a mine, air is a dilutant of inflammable gases. In a dye works, air is a conveyor of steam, removing fog that would otherwise render operations invisible.

Effects of human breathing

The normal adult at rest inhales about 18 cubic feet (500 litres) of air per hour, of which some five percent is absorbed as oxygen in the lungs. The exhaled breath gains from 3 to 4 percent of carbon dioxide (CO_2), equal to about 0.6 cubic feet (17 litres) per hour. If the same air is breathed over and over, a build-up of CO_2 takes place at the expense of oxygen. While CO_2 is non-toxic, such a concentration threatens asphyxiation as a result of insufficient oxygen.

As a criterion of ventilation, however, CO_2 content is inadequate. Of greater importance are temperature, humidity, and odour. The heat from the human body at rest at normal temperatures is sufficient to raise the temperature of 1,000 cubic feet (28,000 litres) of air per hour by 15° F (8° C), and it adds moisture by perspiration and exhalation from the lungs to the extent of about 0.8 grains per cubic foot (1.8 grams per cubic metre). Thus, for a given effective temperature to be maintained, a certain air flow rate can be postulated.

For crowded spaces the problems of smoking and body odour must be considered. Desirable fresh air quantities per person may vary from less than 1,000 to more than 2,000 cubic feet (36 to 60 cubic metres) per hour per occupant in spaces in which human occupancy alone is a factor.

Industrial ventilation rates are dependent on degrees of contamination, quantity of added heat or steam, and amount of solid matter added. Ventilation may be either natural or mechanical.

Natural ventilation results from thermal effects, such as those from a flue, or may be caused by wind, or both. The forces are small and often variable. Their effectiveness is aided by opening or closing windows. In large spaces, such as factories, natural roof ventilators often achieve a sufficient ventilation rate. Some form of inlet is provided at low level.

An interesting method of industrial ventilation, particularly suited to areas of high heat concentration, is the open roof, known as the Colt system. Controllable louvres are opened when conditions permit, giving a clear view of the sky; during rain they are partially closed under control of a rain-sensing device. The same opening acts as a smoke escape in the event of fire. The psychological effect of an open roof is considered highly favourable.

MECHANICAL VENTILATION: APPLICATIONS

Mechanical ventilation depends on the use of fans that either deliver fresh air into the space or exhaust it from the space; in the balanced system, both inlet and exhaust are by fans. Mechanical ventilation, being independent of thermal or wind effects, is positive and readily controllable; hence, if specified air-change rates are obligatory for health or other reasons, it is the only reliable method.

Mechanical inlet with natural exhaust tends to cause a slight pressure within the space, so that air leakage is outward. Thus, if it is installed in an internal office in a factory having a dust- or fume-laden atmosphere, the office will remain essentially contamination-free.

Mechanical exhaust with natural inlet causes a slight negative pressure in the space, so that air leakages are inward. In many cases this is desirable in order to avoid escape of fumes or smells into surrounding areas of the building. Examples occur in laboratories forming part of a college teaching block, in a hotel kitchen adjacent to restaurant areas, and in toilet accommodations generally. In industry, compartments and areas that entail grinding dust, paint spray, fumes, and smoke are similarly treated; these undesirable contaminants are then confined to the spaces in question, leaving surrounding areas free from pollution.

The balanced system is appropriate if the pattern of air distribution is important. In a theatre, cinema, hotel dining room, assembly hall, library, department store, or similar space occupied by large numbers of people, high air-flow rates are necessary, but drafts must be avoided. Air inlets must be situated to avoid direct impingement on the occupants, and vitiated air must be removed to maintain the required flow direction, particularly if smoking is permitted. According to building design, the flow pattern may be upward, downward, or mixed, *i.e.*, part up and part down. The downward system is more appropriate with air-conditioning in which the air is brought in cool and falls by gravity. For "straight" ventilation the upward system is most common: it comprises air inlets in side walls or in other locations above head level; air is removed by the exhaust system through apertures in the roof. In the mixed system some air is exhausted near floor level in order to remove smoke.

Air-flow patterns inside buildings

There are many variations of these methods to suit specific requirements. A few special cases follow.

Hospital operating theatres. A general downward air flow is favoured in order to remove bacteria carried in the air stream away from the body to the floor. Air is introduced at the ceiling at a rate equivalent to some 16 to 20 air changes per hour. Extraction of air is by means of apertures near the floor. Recent research, however, has shown that the downward action does not obtain in practice because of convection currents set up by body heat.

Exhibition halls. An interesting example of both downward and upward ventilation is afforded by the system installed at Earls Court Exhibition, London, which is among the largest covered exhibition areas in Europe. Figure 9 shows a section through half of the Main Hall.

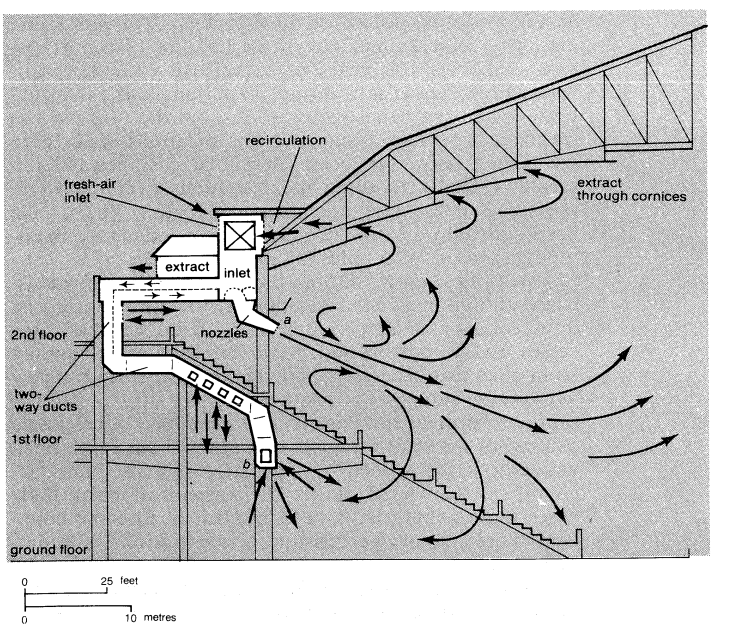


Figure 9: Section of Earls Court Exhibition Building Hall B, showing a two-way ventilating system, (A) for audiences and (B) for exhibitions. Air from *a* is extracted via side ducts under steps; from *b* via cornices in roof.

For exhibition purposes an upward system was required, so that in case of fire smoke could escape through the roof. For sporting events the stepped seating area extends from the balcony to the floor and a downward system was necessary.

The ventilating plants at roof level are arranged to deliver either through the nozzles (downward system) or through the dropping ducts to the spaces below and above the gallery (upward system). When functioning as

a downward system these ducts are reversed and act as extracts, some extraction also taking place through the roof. When functioning as an upward system, all extraction is via the roof.

Vehicular tunnels. Railway tunnels, though once the source of complaint because of the unpleasant smell of smoke and later from diesel fumes, constitute a nuisance rather than a danger because of the infrequency of trains. Natural ventilation caused by wind and the motion of the train has proved sufficient.

Long and busy highway tunnels, through which some 5,000 internal-combustion vehicles per hour may pass, constitute a serious potential hazard because of the exhaust gases, which must be diluted by mechanical ventilation. The principal hazard is carbon monoxide (CO), which may constitute from 3 to 6 percent of the exhaust fumes; unburned hydrocarbons, aldehydes, sulfur dioxide, and oxides of nitrogen may also be present. Ventilation rates of the order of 150 cubic feet of fresh air per minute per foot have been postulated as desirable for a two-lane tunnel and 300 cubic feet (eight cubic metres) for a four-lane tunnel, limiting the concentration of CO to one part per 10,000.

Carbon monoxide and other hazards

Methods of ventilation include the following: the longitudinal system, in which air is drawn in at the two ends of the tunnel and exhausted by a shaft with fans at the centre (*e.g.*, Saint-Cloud Tunnel, Paris); the upward system, in which blowing fans deliver air to a duct below road level, entering at curbside and exhausting through a duct overhead, which is connected to exhaust shafts with fans (*e.g.*, Lincoln and Holland tunnels, New York); the upward transverse system, in which the inlet is also at curbside but the exhaust is through the traffic space to exhaust shafts near the ends (*e.g.*, Mersey Tunnel, Liverpool, England); and the cross-flow system, in which blowing ducts and exhaust ducts occur on opposite sides of the tunnel (*e.g.*, Velsen, The Netherlands).

Underground railway tunnels. The ventilation of an electric railway tunnel is chiefly designed to remove heat. A single tunnel accommodates only one train, which acts as a loose-fitting piston for the distribution of ventilation air. The Victoria Line, part of the London Transport system, embodies 100 years of experience with tube railways. Fresh air is introduced at stations and exhausted by fans and shafts between stations, but the fans may be reversed in winter. Because drafts of almost gale force have been known to occur in public passageways and escalator shafts in some older sections of the system, pressure-relief shafts and cross-connecting ducts have been built into the new line, in order to reduce the maximum air speed to 15 miles (24 kilometres) per hour.

Industrial exhaust. Often, the main function of industrial ventilation is exhaust: that is, to remove airborne dust, solid particulate matter (*e.g.*, from grinding processes and woodworking), toxic fumes, steam, vapours from chemical processes, and smoke (*e.g.*, from foundry work and welding). The primary purpose of such ventilation is to protect the worker and to reduce atmospheric content of harmful products to an innocuous level. Such ventilation is governed by legislation in most industrial countries. Threshold limits of concentration of toxic gases and vapours have been published; in some countries these include permissible concentrations of radioactive materials.

Filtering airborne, solid particles

Airborne solids are usually collected as near to the point of production as possible and are then exhausted by fans through ductwork. The fan discharge to atmosphere is through a cyclone or filter to avoid external pollution.

Smokes, fumes, vapours, and dust are collected in hoods over the area of production and are exhausted to the atmosphere. Wet scrubbers, or chemical absorbent towers, are usually arranged to arrest the discharge of harmful gases. If greasy fumes emanate from the operation, as in a kitchen or food-processing department, filtration may be necessary at the source to avoid the hazard of fire in the ducts.

Extraction of air is often on a considerable scale and means for replacement are necessary. These may be through inlets around the building perimeter near the floor. To prevent drafts, mechanical inlet fans followed by an air heater and filter are used; sometimes fresh air is drawn through the roof by means of units strung overhead.

MECHANICAL VENTILATION: EQUIPMENT

Fans. All systems of mechanical ventilation include a fan, driven by an electric motor. Fans are available in various types according to application. The simple propeller or disc type, with three, four, six, or more blades arranged on a shaft, is suitable for free-air ventilation. It is noisy and inefficient, however, if working against the resistance of heaters, filters, and ducts.

The axial-flow fan, with airfoil blades, is more efficient, but its higher speeds cause noise. This noise can be absorbed in suitably designed silencers.

The centrifugal fan consists of a drum on which is mounted a series of curved blades, coaxial with the spindle or shaft. On rotation, air is drawn along the centre line of the shaft, and thrown out by centrifugal force from the blades in a spiral flow.

The centrifugal fan, which is capable of delivering air against appreciable resistance without undue speed and noise, finds wide use in applications that require quiet running. Another type, the mixed-flow, is a cross between the propeller and the centrifugal. The tangential fan consists of an extended drum that creates a flow in which the air is drawn in on one side of the drum and expelled from the other side. Its use has been confined to small convectors and electric heaters.

Heaters. Most ventilating applications demand some means of air warming during winter. The source of heat is commonly hot water or steam, but electricity is used on a limited and local scale, as is indirect-fired gas. Heaters take a variety of forms, with thermostatic control of heat input regulating outgoing temperature.

Filters. Air cleanliness is of paramount importance in many applications, such as hospitals, office blocks, computer rooms, photographic chemical production, food industries, and so on. Particulate matter in the air of urban localities may contain soots from chimneys, dust of various kinds, pollens, industrial waste, carbon from wear of tires on roads, and oily deposits from diesel fume exhausts. Particles range in size from 0.01 micron to 100 microns (1000 microns = 1 millimetre). Tobacco smoke particles are of the order of 0.1 micron, and atmospheric dusts range from 0.5 to 10 microns. The smallest dust particle visible to the naked eye is about 12 microns, and a human hair is about 100 microns in diameter.

Air filters are classified either on a weight test or on a blackness test. On a weight test, heavier particles account for the greatest mass and a coarse filter may show a high efficiency, but on a blackness test the stain resulting from minute particles may predominate and show such a filter to be very inefficient. For test purposes, therefore, sample dusts of various orders of fineness are used to permit reliable performance data for filters of various types. Filters are available in several types.

Types of air filters

The impact type is usually composed of metal plates, metal coils, or turnings, coated with a viscous fluid to which particles adhere. Impact filters are efficient on a weight test but inefficient on a blackness test.

Fabric types consist of some form of textile material or woven fibre that arrests particles in the interstices until they are clogged. Paper types and others using glass fibre or similar material also work in this way. Of low efficiency when used with fine particles, they improve sharply in blackness tests of particles larger than about 1 micron.

In electrostatic filters, dust-laden air is passed over electrically charged wires, giving a charge to all particulate matter present. The air and particles then enter a series of parallel plates, which are alternately charged and grounded. The charged dust particles are repelled onto the oil-coated grounded plates, to which they adhere. In this way airborne solids of all sizes are arrested for subsequent removal by washing. Efficiency is high throughout the range of particle size.

An absolute filter uses a special form of paper capable of removing the last traces of staining.

Activated charcoal has a high degree of porosity and is very useful in absorbing odours and undesirable gases. Such removal may be desirable if air is recirculated for economy of heating or cooling, or if stray gases, such as sulfur dioxide, are present.

The air washer consists of a series of water sprays causing a mist through which the air from the plant flows. Most of the particulate matter in the air stream in urban localities is, however, not readily wettable, and the air washer as an air cleaner is of low efficiency. It is more useful as a means of humidity control and in air conditioning as a means of cooling.

Cleaning. The cost of cleaning and maintenance of filters rises according to their efficiency. Impact types may be cleaned and re-coated manually or automatically. Fabric types may be thrown away when dirty and replaced, or the filter material may be moved from a clean roll to a dirty roll, as in a camera film. Electrostatic filters are cleaned by periodic washing down with hot water. Other systems rely on sparking between the plates to cause agglomerated material to be discharged and subsequently arrested on a coarse after-filter. Absolute filters are used only as secondary filters; when used, they are simply discarded.

There is a relationship between cost of filtration equipment and its maintenance that determines the degree of filtration efficiency. In a machine shop a coarse degree of filtration may be adequate, while in a hospital or in a factory producing transistors, the highest efficiency obtainable may be required.

Noise control. Air moved by any means is liable to set up sound waves because air itself is the medium by which sounds reach the ear. The impulse given to air on impact with successive blades of a fan is an obvious source of noise that may be conveyed through ducts to areas at a distance.

Degrees of noise toleration vary according to environment. In an environment in which machinery is at work the ear becomes accustomed to a high level of sound, and noise from ventilation systems is lost in the overall din. In a board room with double windows and heavy carpeting, on the other hand, a small noise from ventilating plant would be intolerable.

The punkah fan

Circulation of air. The punkah ceiling fan, once common in restaurants, stores, and offices all over the world, originated in the tropics before the days of air conditioning. Perspiration from the skin was the chief means of body-heat release, and the mere movement of air over the surface gave relief by evaporation, until temperature approached blood temperature and air movement ceased to be effective. For this reason, circulating fans fail to provide real relief in warm climates.

IV. Air conditioning

The term "air conditioning," first used in connection with the ventilation of textile mills, has come to mean a system in an enclosed space in which temperature and humidity are controlled within close limits, regardless of external and internal influences. It also involves control of air cleanliness and avoidance of drafts and noise.

Temperature may be controlled only by adding or by removing heat; *i.e.*, by heating or cooling. Humidity may be controlled only by adding or by removing moisture. An air-conditioning plant is designed to perform both these functions.

The need for air conditioning in temperate climates arises from many causes. Modern building construction, using large areas of glass in light curtain walling, creates conditions in which solar gain is often so great as to make such buildings untenable without means for cooling. Internal heat gains have also increased considerably because of greater intensity of lighting and the presence of electrical machinery. In offices it has been shown that efficiency is higher, and absenteeism through illness lower, in air-conditioned space. In shops and stores, air conditioning improves the atmosphere and attracts customers. Air conditioning for comfort is especially necessary in crowded assembly spaces such as theatres, cinemas, exhibition halls, night clubs, and lecture halls.

Importance of air conditioning in hospitals

In hospitals, air conditioning is now common for operating theatres, special treatment areas, pediatric units, X-ray departments, and spaces using radioactive isotopes and electron microscopes. The completely air-conditioned hospital is becoming more common, particularly in areas of high urban development. Here, considerations of air cleanliness and the need for sealed windows to exclude noise and dirt are determining factors. These considerations also apply to hotels.

Other spaces in which air conditioning has become accepted as a basic amenity include passenger aircraft, passenger ships, railroad passenger cars, motor coaches, and, in many regions, automobiles. Television and radio studios, with their high heat inputs from lighting, would become unbearably hot without air conditioning. The need in subtropical and tropical climates for relief from enervating and oppressive conditions is, of course, especially pronounced.

Reference has been made earlier to the effective temperature scale and the zone of comfort. All air conditioning applications in which human comfort is the design aim must create conditions lying within this zone, though there may be need for variation upwards in very warm climates to avoid undue thermal shock on entering from outside into an air-conditioned space. Representative examples of external dry-bulb and wet-bulb temperatures as a basis of design are given in Table 3.

Table 3: Dry- and Wet-Bulb Temperatures as Basis of Design

	dry bulb		wet bulb	
	°F	°C	°F	°C
Allahabad, India	115	46	—	—
Auckland	79	26	67	19
Berlin	89	32	69	21
Cape Town	93	34	72	22
Chicago	96	36	75	24
Cologne	87	31	69	21
Hong Kong	91	33	82	28
Madrid	96	36	72	22
New York City	94	34	75	24
Paris	90	32	70	21
Rome	96	36	74	23
Seville	107	42	82	28
Sydney	95	35	72	22
Tokyo	92	33	79	26
United Kingdom	85	29	68	20
Washington, D.C.	97	36	77	25

The inside temperature suggested for an arid climate is 76° F (24° C), being reduced to 71° F (22° C) effective temperature as the outside temperature falls below maximum.

In industry, another consideration applies, namely the quality of the product. An atmosphere in which temperature and humidity are controlled is essential in a wide range of manufacture, including food processing, colour printing, and textile printing, and the manufacture of textiles, photographic materials, electrical precision instruments, and gauges. The manufacture of synthetic materials and other modern manufacturing techniques often required stable atmospheric conditions in order to meet traditional product standards, such as those of the leather industry, with its long tradition of the treatment and tanning of hides, for which natural conditions of the environment had sufficed.

The windowless factory, a recent innovation, can only function with air conditioning, which can be justified on economic grounds, partly through saving on window heat losses and gains.

The properties of mixtures of air and water vapour are best understood from a study of the psychrometric chart of which Figure 10 is a simplified version. The horizontal scale is dry-bulb temperature, the vertical scale is absolute moisture content. Oblique lines represent wet-bulb temperatures. The limiting curve or 100 percent saturation line also represents dew-point, at which dry-bulb

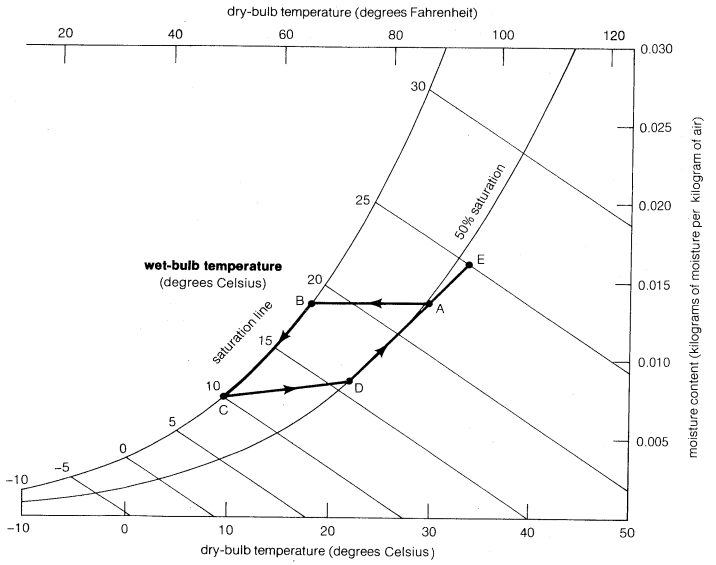


Figure 10: Portion of a simplified psychrometric chart showing the dehumidifying effect of cooling.

and wet-bulb temperatures are the same. Intermediate curves between 100 percent and 0 percent saturation are inserted on the full chart to give percentage saturation, such as the 50 percent drawn. Other data, not shown here, give total heats or enthalpies, vapour pressures, and volume lines.

The saturation curve

If an air–water-vapour mixture from a room is cooled progressively from a point A, it will arrive at point B on the saturation curve. Further cooling is then only accompanied by reduction in moisture content; *i.e.*, by condensation. Having then arrived at a point C, the air may be said to be dehumidified with respect to its initial condition at A. If then delivered to the room, heat gains within the room or through the enclosure cause a rise of temperature and a moisture gain from occupants or from other sources of water vapour, so that the air will arrive at condition D, which may be taken as the design condition for the room.

Fresh air for ventilation may be introduced at condition E; the balance of air in circulation is made up of recirculated air from the room. The mixture is then at the starting point state A, determined by the relative proportions of the two sources of air. From this point the cycle repeats. Some heat might in practice be added at point C to reduce the temperature differential of air entering the room.

The above example illustrates the process of cooling and dehumidification during summer. Another cycle obtains in winter when warming of incoming air is necessary, as well as addition of moisture; *i.e.*, humidification.

Cooling and dehumidification involve reduction of sensible heat (lowering the temperature) of air and vapour, and removal of latent heat of water vapour. The sum of these two heats or enthalpies constitutes the cooling load, generally accomplished by a refrigerating plant. Similarly, warming and humidification involve the addition of sensible heat (raising the temperature) and of latent heat of evaporation, the total being the heating load supplied by a boiler or other source of energy.

TEMPERATURE AND HUMIDITY CONTROL

Cooling. *Evaporation.* One of the earlier methods of cooling was by evaporation of water. In tropical climates of low humidity such as India, a wetted brattis or shutter, fanned by hand, provided some measure of cooling.

The modern equivalent is the air washer in which water is sprayed into the air stream by a pump; evaporation results, causing a reduction of temperature, though at the expense of an increase of humidity. In terms of effective temperature, it adds little to comfort.

Refrigeration. By the application of refrigeration to air conditioning, true control of atmospheric conditions may be achieved. On reduction of temperature, the desired condition of moisture is obtained, subject to prescribed control. Dehumidification in a crowded atmosphere is at least as important as lowering of temperature. The lowered air temperature that is issuing from the dehumidifying plant may, however, be too low for delivery into the occupied space without causing drafts. To overcome this problem of drafts, some reheating may be applied.

Four principal means of cooling by refrigeration

Cooling by refrigeration is applied in four principal ways.

1. By direct expansion of refrigerant gas in cooling coils, usually tubular, with fins over which the air passes. The fins run vertically to allow condensation to fall to the base and then to a drain. The coils constitute the evaporator of the refrigeration plant. Direct expansion is used with refrigerants such as one of the Freon group. This method is used in most small air-conditioning units.
2. By chilling water in a tubular evaporator, with the water circulated by pump through cooling coils as in 1. The system may be extended to serve a number of such cooling coils in various parts of a building, confining the refrigerant gas to the central plant.
3. Chilled water may alternatively be sprayed into the air to be cooled in an air washer and dehumidifier. As explained earlier, the washing function has low efficiency, hence the prime purpose of the sprays is for cooling. This method is often used in large plants to avoid the use of massive cooling coil surfaces. In an alternative form of washer, water is sprayed over cells of fibre-glass through which the air passes.
4. In the sprayed-coil system, chilled water coils are used as in 2 but they are contained in a washer casing as in 3. Water, sprayed over the coils by pump, is constantly recirculating. As condensation occurs the spray water grows in volume and overflows from the base tank to a drain. One merit of this system is compactness.

The refrigeration plant used for air conditioning follows the principles of normal commercial equipment (see REFRIGERATION EQUIPMENT).

Another class of refrigerator is the absorption machine using lithium bromide as the cooling medium, and requiring only heat from gas, hot water, or steam. The relative merits of absorption versus mechanical compression are beyond the scope of the present article.

A combination of centrifugal compressor with turbine drive and an absorption plant is used in a number of installations in the United States. The exhaust steam from the turbine supplies the heat for the absorption machine, effecting a high thermal efficiency.

The condenser, forming part of any refrigerating plant, is, for small direct-expansion plants, usually air cooled. For larger sizes, water circulation is employed in a conventional manner, making use of evaporative coolers in any of the various forms.

Adsorbents. Another method of dehumidification is by the use of an adsorbent such as silica gel, a form of silica in a finely divided state. Water vapour in air passed through a bed of such material is adsorbed, generating heat that is removed by cooling water. Regeneration by heating to a higher temperature drives off the moisture, the bed being cooled by air for re-use. Two plants are needed for continuous use, one dehydrating and one regenerating, with change-over arrangements. Very low humidities are obtainable with this system, rendering it particularly suitable for industries in which very dry air is required. This method is not primarily used for air conditioning.

Liquid adsorbents such as aqueous solutions of lithium chloride and calcium chloride have also been used industrially as a means of dehydration.

Peltier effect. J.-C.-A. Peltier, a French physicist, in 1834 noted that heat was produced at the junction of two dissimilar metals when a current was passed through, and that if the current were reversed, the junction would be cooled. If current were passed between two properly constructed junctions connected in series, one junction would be heated and one cooled; in effect, heat was transferred from one junction to the other.

In the application of the Peltier effect to air conditioning, materials having low thermal conductivity and high

Thermoelectric cooling and heating

electrical conductivity are used. Such materials include semiconductors formed of alloys of bismuth telluride (Bi_2Te_3) and bismuth selenide (Bi_2Se_3). In practical applications, an assembly is made up in the form of fins alternately in the hot air stream and in the cool air stream. Separate fans deliver air over the two sets of fins, that from the hot junction to the atmosphere and that from the cooled air fins to the space to be cooled. By changing over the air streams the same system may be used for heating the space. The hot junction to the atmosphere may alternatively be water cooled. (See also THERMOELECTRIC DEVICES).

Warming and humidifying. For year-round use in countries in which heating is required in winter, the air-conditioning system includes means for warming the air. In small self-contained units this may be accomplished with a heat pump. By changing over the direction of circulation of refrigerant, the evaporator becomes the condenser so that warm air is delivered to the room; meanwhile the condenser becomes the evaporator, accepting outside air and returning it to outside at a lower temperature.

In larger plants the air warming is achieved by heating coils supplied with hot water or steam or, in special cases, heated by gas or electricity.

Humidification is necessary under many winter conditions in which low outside temperatures are accompanied by low moisture content. Vapour may be added to the air stream by direct injection of steam, or water spray from the main, or by evaporation from pans heated electrically. If an air washer is used, the same unit serves the double purpose of dehumidifying in summer, when supplied with chilled water, and humidifying in winter when the spray water is warmed.

Equipment. *Central plant.* An air-conditioning plant serving a single space such as a banquet room is shown diagrammatically in Figure 11. The functions of the various parts are noted in the caption. In this example, direct-expansion cooling coils are used, in addition to an air heater, and water or steam is injected for humidification. The arrangements for air circulation within the space may take many forms. As shown, a small amount of air is exhausted to the atmosphere from the ceiling to remove smoke; the low level extract is recirculated, or, when favourable weather prevails, may be rejected, in which case the plant handles entirely fresh air through the fresh-air intake. This method saves power for operating the refrigerating compressor. Air filtration takes one of the forms referred to above.

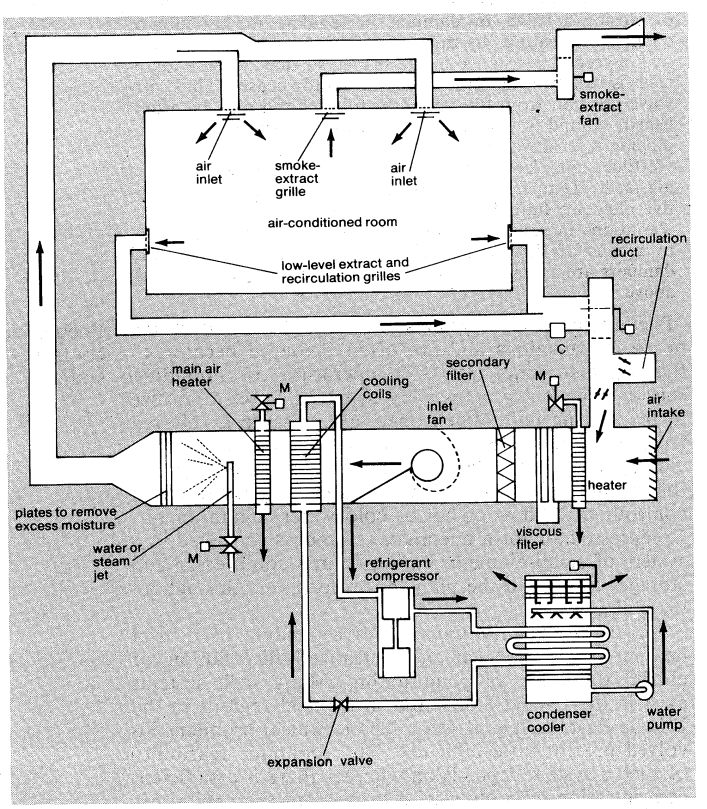


Figure 11: Complete air-conditioning system for a single room, involving heating, cooling, and humidifying as needed. M denotes motorized valves; C denotes temperature and humidity controls.

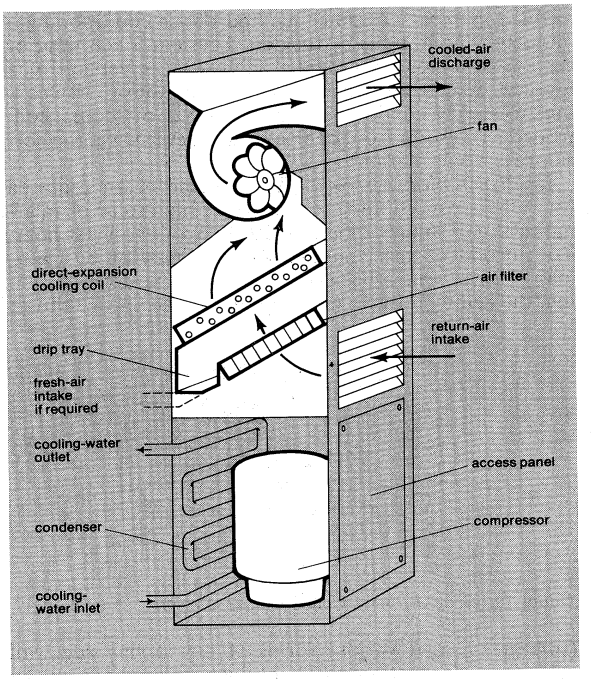


Figure 12: Simplified version of a self-contained air-conditioning unit.

Self-contained units. One form of this type is shown in Figure 12. The condenser here is water cooled, but it may alternatively be air cooled. Provision for warming is not included, but may be added. Some relatively crude forms of air conditioning merely consist of a series of such self-contained units.

Fan-coil units. If a large number of rooms are to be served, one method is to provide each with a unit supplied with conditioned air from a central plant. Another method takes in fresh air from outside. In both cases, a fan also recirculates air from the room. The temperature is regulated by a thermostat controlling water flow to the coil, which receives chilled or warm water according to the time of year.

Induction system. This system has been extensively developed for multistory office blocks and hotels. Conditioned air from a central plant is conveyed at high velocity through small ducts to each unit. This "primary" air is discharged through a series of small jets, so that a flow of "secondary" room air is induced that passes over the coil surface within the unit. Chilled or warm water is supplied to the coil on a change-over system according to outside temperature. Alternatively three-pipe and four-pipe systems have been developed so that some units on the warm side of the building that are subject to solar gain may select chilled water, while those on the cool side may select warm water.

Dual-duct system. This system dispenses entirely with water circulation and makes use of two ducts, one supplying warm air and one cool air. A blender unit serves to control admission of air from the ducts according to the dictates of a controlling thermostat in the room. Greater air quantities are handled by the central plant than in the induction system, but the change-over problem is avoided and the system is more versatile in dealing with internal areas as well as perimeter rooms.

Variable volume system. This mainly applies to internal areas of a building. It achieves control of temperature by variation of the quantity of air admitted from the central plant.

Zoned system. Another method of air conditioning a multistory building is by the zoned system. Each floor or section of a floor is provided with a large version of a fan-coil unit, from which ducts run out above false ceilings to supply conditioned air to a group of rooms, controlled as a group from one point. Air returns via the corridor for recirculation.

Air distribution systems. The volume of air required for full air conditioning of a space of high solar or internal heat gains may amount to 20 changes of air per hour. To introduce this volume without noticeable drafts, many devices have been developed: the perforated ceiling, the ceiling diffuser (see Figure 13), the linear diffuser

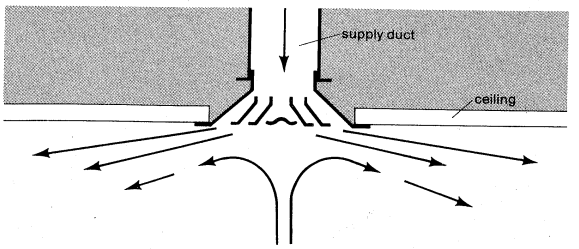


Figure 13: Air flow from a typical ceiling diffuser.

(Figure 14), and various forms of directional side wall grilles. The induction and dual-duct systems rely on an under-window position for perimeter distribution. Diffusion from there is accomplished without noticeable draft.

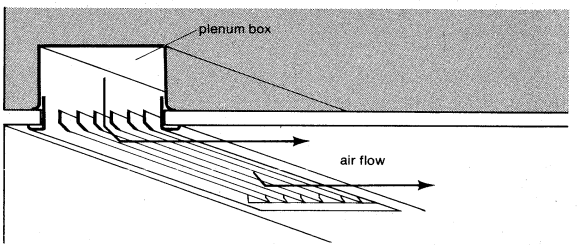


Figure 14: Linear diffuser mounted in ceiling.

Ceiling panels. A fundamentally different approach to cooling is that in which surfaces of the ceiling are cooled by panels maintained at a temperature lower than that of the room. The sole purpose of air distribution is then ventilation. An essential feature of such a system is that the ventilation air is conditioned so that its dew-point remains below the surface temperature of the cooling surface in the room, avoiding condensation.

The same concept has been applied in removing heat from ceiling light fixtures. If the heat from intensive systems of lighting is removed by cool water circulating around the fitting, heat gain to the space is stopped at the source. This avoids large air flows after the heat is radiated to the wall and floor surfaces of the room. Another approach to the same problem is to arrange for extraction of air via the lighting fittings, though this may increase cleaning maintenance.

The heat pump in air conditioning. Self-contained heat pump units may be used for cooling in summer and heating in winter by reversing valves in the refrigerant piping, so that the condenser and evaporator change function.

Another air-conditioning application of the heat pump is in the multistory block of offices in which areas in the core have a perpetual cooling problem because of lighting and occupancy heat gains, while the perimeter spaces are subject to heat losses. The heat removed from the centre core of such a building by means of air extraction is up-graded in the condenser to supply the heating needs of the exterior while the core receives the cool air from the evaporator.

In another system, the heat-recovery system, individual room-conditioning units are connected to a water circulation system to receive water from a constant-temperature circuit and return it either warmed (if cooling) or cooled (if heating) to the common return. The latter is then either heated, through a heat exchanger and boiler, or cooled, by means of a cooling tower on the roof. Recovery of heat from exhaust air may be included in the system.

The heat recovery system

Controls. Perfect functioning of any air-conditioning system relies entirely on the correct operation of the control equipment. The principles of control apparatus are simple enough, requiring only a means of detecting temperature and humidity and a source of power to operate valves or dampers in response to the demands of the controller.

Various systems have been evolved, some of them electrical, some pneumatic, all consisting basically of one element to sense conditions and one to convert signals into practical realities of heat or water vapour. In a large system a series of controls is necessary. In a multistory building, for example, there is the control of the central plant, the refrigerators and the boilers; then the control is required for the circulating water or air sent out, either heated or cooled. Some aspects of the building are subject to solar gain, others not. Some areas may be sparsely occupied, some may contain conference rooms holding several hundred persons. In industry the quality of product may depend on temperature and humidity being controlled within narrow limits.

To deal with the variety of conditions encountered, each control system must be individually designed. The permissible tolerance in variation of conditions must be decided. A plus or minus of 2° F (1° C) may be acceptable in an exhibition hall but would not suit a gauge room of a machine shop.

Electrical control systems are of three main types.

1. On-off thermostatic switching. This suffices for the small unit room-conditioner in which cooling is the main object and humidity control incidental.
2. Potentiometric. In this type the temperature-sensing device may use a volatile liquid and Bourdon tube or bimetallic helix to move a wiper arm (mechanically) over a resistance forming one arm of a bridge circuit. The out-of-balance current in the bridge then actuates a contactor to cause a motor to rotate in the appropriate direction, either opening or closing a valve, or damper, to regulate the amount of cooling or heating. In turn, this motor runs a wiper arm over a second resistance in the bridge so that, when equilibrium is restored, the motion of the motor ceases. Thus, any valve position from fully open to fully closed may be established. Humidity control may likewise be achieved by employing a sensitive material.
3. Electronic. This system dispenses with mechanical moving parts. Temperature is measured by a temperature-sensitive element, humidity by a hygrometric-sensitive element. Variations in resistance are converted into sizeable current by means of thermionic valves or transistors. Valves and dampers are actuated by the current through motors as in 2, above.

Pneumatic control systems

Pneumatic control systems may equally perform the necessary functions and are often favoured because of their greater simplicity. A temperature- or humidity-sensitive element progressively opens or closes an orifice in the sensing unit. When open, air pressure is relieved from above a diaphragm in a pneumatic actuator; when closed, pressure builds up. The actuator is connected to the spindle of a valve or to the lever of a damper, for controlling the flow of hot or cold water, steam, or air. Compressed air for a pneumatic system is conveyed in a system of small-diameter piping from a compressor and storage cylinder in the plant room to points throughout the building.

District air conditioning. Corresponding to district heating is the concept of centralized distribution of chilled water for air conditioning. Many such systems exist in the United States and in tropical regions in the Middle East, such as Kuwait. The technical problems involved present little difficulty. The economic feasibility of such a system depends on whether there is a sufficient concentration of cooling load to warrant the capital expenditure on plant and mains distribution. This situation occurs in high-density urban areas that typically include department stores, hospitals, office blocks, blocks of

apartments, theatres, and exhibition halls. Each building requiring air conditioning is connected to chilled water mains, which are fed to air-cooling batteries of the central plants; alternatively, fan-coil cooler units are served from the system, obviating the need for local refrigerating machinery and simplifying maintenance problems.

Centralization of cooling load brings a bonus in terms of total capacity. Diversity of demand as between one building and another results in a considerable reduction in overall load as compared with the sum of all the loads taken separately. As much as 60 percent reduction in total demand may apply to an extensive system.

Noise and vibration control. Early air-conditioning equipment was likely to be noisy, particularly the self-contained units. Such units have since achieved a much better standard of quietness. In all air conditioning the first problem in respect to noise is the fan. Given the sound power level at each frequency range it is possible to determine in advance the attenuation of fan noise resulting from ducts, changes of direction, and grilles. If too great a sound level remains, a slower speed fan may be selected, or means may be incorporated for sound absorption, such as lining ducts with acoustic absorbing material or inserting a silencer. The latter usually takes the form of an enlarged section of duct in which the air stream is split up by vanes of honeycomb covered with absorbent material. Use is also made of perforated plate backed by such material. The length of a silencer is determined by the need for a certain sound-level reduction at various frequency ranges. High frequencies are more easily absorbed by this means than low, but the ear has a greater tolerance to low-pitch sounds.

Silencing a fan

Ductwork systems were at one time designed on the assumption that low velocities of air movement resulted in quietness. The general use of high velocities coincided with the induction and dual-duct systems. Systems with velocities in the range of 3,000 to 4,000 feet (900 to 1,200 metres) per minute are as quiet as any conventional system using large ducts. This quietness is achieved by careful acoustic design throughout, with special attention to final terminal units.

In addition to airborne noise conveyed through ducts, there is the problem of mechanical vibration entering the structure. A refrigerating compressor mounted in a plant room on the roof may set up vibrations that are conveyed through the columns to a lower floor. Such a plant may not in itself be noisy, but it may set up resonance in the structure. Pumps, fans, motors, and compressors all contribute to the problem, which can be minimized by careful selection of speeds, types of equipment, and mounting arrangements.

ASSOCIATED COMFORT FACTORS

Ionization of air. Gases contain molecules or groups of molecules that become charged positively or negatively by natural ionizing forces near the Earth's surface; these charged molecules, or ions, include radioactive elements in the soil, cosmic rays, frictional electricity generated by rain drops, and lightning. Artificial ionization may be brought about by electrical discharges and short-wave ultraviolet radiation.

Research has shown that occupation of a room reduces the ionic content. Ventilation with fresh air at a rate of 160 cubic feet (five cubic metres) per minute per person is required to maintain the normal content.

It has also been shown that mechanical ventilation may reduce the ion content by 30 percent as a result of absorbence on metal duct walls, negative ions being removed at a faster rate than positive. Heating in a central fan system increases ion count, while cooling reduces it. Washing by sprays removes the small ions but produces many large negative ions.

Observations reported by research workers are summarized briefly in Table 4.

Table 4: Effects of Ions on Humans

	positive ions	negative ions
Mucous membranes	irritation	relief
Respiration	accelerated	decelerated
Hay fever	discomfort	relief
High blood pressure and rapid pulse	rise in pressure and pulse	relaxation

It has been postulated that the effect of ions occurs at the cellular level, but whether the electric charge causes a change of conductivity of the cell is not known.

Tests conducted in 1959 on tissues of rabbits and mice showed that positive ions irritate by causing a decrease in mucous flow, a drying of the surface and other effects. Negative ions have the reverse effect.

All this evidence seems to show that ion content has some physiological impact, but a study by the American Society of Heating, Refrigeration and Air Conditioning Engineers in 1962 produced no statistical evidence that the ionic state had any effect on sensations. A British authority concluded from this and other evidence

The question of ion content

> that the differences which normally occur in the ion content of the air of occupied rooms do not exert any significant influence on subjective sensations . . . . if there is any effect it is too small to be of engineering significance.

Ozone. Ozone is a gas with the chemical symbol O_3 (that is, each molecule has three oxygen atoms as compared with two in normal oxygen, O_2) and is present in the air in small quantities. Ozone is unstable and readily breaks down into normal oxygen plus a detached oxygen atom, O. The detached atom appears to combine rapidly with most available substances, producing a virulent oxidizing effect. This property gives ozone various applications, such as eliminating odours and clarifying water and other substances. It is no longer applied in ventilation and air conditioning owing to the smell caused by oxides of nitrogen that always accompany ozone.

Ultraviolet irradiation. Micro-organisms, especially those in moist droplets, are vulnerable to ultraviolet radiation. Serious, though temporary, inflammation of the eyes is caused by exposure to such rays, so use must be indirect in applications involving human exposure.

Sterile chambers, such as laboratories, may be irradiated with ultraviolet light to produce complete sterility before being opened for work. Indirect radiation of the upper part of rooms is another method, reliance being placed on mixing of air from the lower area by natural or mechanical means. Another method used in some hospitals has been to direct irradiation onto a curtain at the entrance to infants' cubicles.

For use in ventilation ducts, ultraviolet lamps are arranged so that air passing through the duct is sterilized, but the intensities and air flows required in ducts of various sizes are uncertain. The use of ultraviolet irradiation has an application in the biochemical field, though it has been stated that no method of sterilization, whether by irradiation or chemical means, within the limits of what is tolerable, deals with more than 50 to 70 percent of bacteria. With the present state of knowledge it therefore appears that for air conditioning in general, irradiation by ultraviolet light cannot be regarded as a practical means of reducing the spread of infection and disease.

Mixing of air. Brief reference has been made earlier to the problem of air distribution in the large quantities necessary for air conditioning without draft. Air speeds much more than 30 feet (9 metres) per minute at head level are generally regarded as a draft.

The temperature of the incoming air is important. In order to minimize air quantity and hence cost on a cooling cycle, the greater the differential between cool inlet air and room air the better. On the other hand, cold air entering a warm room at low velocity drops rapidly, thus being liable to cause local cold spots. No more than 10° or 12° F (6°–7° C) differential can be allowed.

Use of air entrainment

Air issuing from a jet entrains "secondary" air from the surrounding space in the ratio of three to one or even six to one depending on configuration and velocity. Techniques have been developed to make use of this feature to enable a smaller quantity of primary cooled air to mix with a larger volume of room air at point of entry, reducing an initial temperature differential of 30° F (17° C) to no more than 10° F (6° C) at a distance from the outlet

at which the stream breaks up into nondirectional eddies.

The technique must take into consideration relationship to wall surfaces and proximity of one discharge point to another. Streams of air in opposite directions meeting at ceiling level have a propensity to fall toward the floor or, if striking a side wall, to continue to descend, as in a waterfall.

Another technique, already referred to, is to split the air flow up into a large number of very small streams, each in effect constituting its own jet pattern. Equipment of this type breaks down the entraining air of high-temperature differential by mixing with room air and so allows the conditioned space to benefit from cooling without noticeable local drafts.

Possibly less acute is the problem of the location of extract or return-air grilles. Discomfort caused by these can usually be alleviated by spacing. In many cases extraction may occur at ceiling level, even though inlet air also enters at ceiling.

BIBLIOGRAPHY

General works: C.W. BRABEE, *Heating and Ventilation* (1927); AMERICA SOCIETY OF HEATING, REFRIGERATION, AND AIR-CONDITIONING ENGINEERS, *Guide and Data Book*, 3 vol. (published at intervals), is universally accepted as the best source of authoritative data—much of which is based on the Society's own research and therefore slanted to American practice; INSTITUTION OF VENTILATING AND HEATING ENGINEERS, *Guide*, 3 vol. (published at intervals), equally authoritative (slanted to British practice); OSCAR FABER and J.R. KELL, *Heating and Air Conditioning of Buildings*, 5th ed. (1971), a classic in Britain, giving a general review of heating, ventilation, and air conditioning; JAMES L. THRELKELD, *Thermal Environmental Engineering*, 2nd ed. (1970), a more advanced textbook suitable for the student.

Theory: THOMAS BEDFORD, *Basic Principles of Ventilation and Heating*, 2nd ed. (1964), by an acknowledged authority on physiological reactions to atmospheric conditions; F.C. HOUGHTON and C.P. YAGLOU, *Determination of the Comfort Zone* (1923), another classic study; T.C. ANGUS, *The Control of Indoor Climate* (1968), a condensed version of his research papers in the field of physiological reactions to engineering problems.

Refrigeration and air conditioning: NORMAN R. SPARKS and CHARLES C. DILLIO, *Mechanical Refrigeration*, 2nd ed. (1959); OSCAR FABER, "The Value of Heat with Special Reference to the Heat Pump," *Proc. Instn. Mech. Engrs.*, 154:144–163 (1946); J.C. DAVIS, "Energy Use and Power Demands in All-Electric Houses Equipped with Air-to-Air Heat Pumps," ASHRAE *Jl.*, 4:87–95 (1962), are representative of the many books and papers available on the subject of the heat pump. WILLIS H. CARRIER, *Modern Air-Conditioning, Heating, and Ventilating* (1940), met a need for a comprehensive treatise on air conditioning. It is now published as CARRIER CORPORATION, CARRIER AIR CONDITIONING COMPANY, *Handbook of Air Conditioning System Design* (1966). WILLIAM PETER JONES, *Air Conditioning Engineering* (1967), is a textbook for the engineering designer.

Condensation: The question of condensation in buildings is thoroughly investigated in the HMSO Building Research Station Digests, "Prevention of Condensation," 2nd series, no. 91 (1968) and "Condensation in Dwellings—A Design Guide" (1970).

Insulation: NEVILLE S. BILLINGTON, *Thermal Properties of Buildings* (1952), provides in-depth coverage of building construction and insulation as it affects heating. See also the essays by G.D. NASH, J. COMRIE, and H.F. BROUGHTON in BUILDING RESEARCH STATION, *The Thermal Insulation of Buildings* (1955), another work specializing in the field of insulation.

Air Purity: R. WILLIAMS, "Methods for Maintaining Bacterial Purity of Air," IHVE *Jl.*, 16:404–438 (1949), summarizes much research work in the field of immunization by various methods.

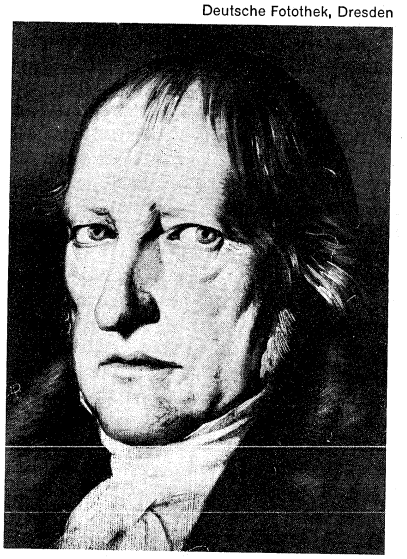
Solar energy: "British Climatological Research Unit Report, 1966" *Architects Jl.*, vol. 13 (1966); J.C. MCVEIGH, "Some Experiments on Heating Swimming Pools by Solar Energy," IHVE *Jl.*, 39:53–55 (1971); G. HASSAN, "A Design Procedure Suitable for Calculating the Size of a Flat Plate Solar Heat Collector Needed to Warm an Outdoor Swimming Pool in Great Britain," IHVE *Jl.*, 39:56–62 (1971); H. HEYWOOD, "Operating Experiences with Solar Water Heating," IHVE *Jl.*, 39: 63–67 (1971), papers dealing with applications of solar heating; A.M. ZAREM and DUANE D. ERWAY, *Introduction to the Utilization of Solar Energy* (1963), a general review of the application and limitations in the use of solar energy; SOLAR ENERGY SYMPOSIUM, *Air Conditioning, Heating, and Ventilating* (1959), papers with greater depth in specialized application.

(J.R.K.)

Hegel, Georg Wilhelm Friedrich

G.W.F. Hegel was the last of the great philosophical-system builders of modern times. His work, following upon that of Kant, Fichte, and Schelling, thus marks the pinnacle of classical German philosophy. As an absolute Idealist inspired by Christian insights and grounded in his mastery of a fantastic fund of concrete knowledge, Hegel found a place for everything—logical, natural, human, and divine—in a dialectical scheme that repeatedly swung from thesis to antithesis and back again to a higher and richer synthesis. His influence has been as fertile in the reactions that he precipitated—in Kierkegaard, the Danish Existentialist; in the Marxists, who turned to social action; in the Vienna Positivists; and in G.E. Moore, a pioneering figure in British Analytic philosophy—as in his positive impact.

Deutsche Fotothek, Dresden



Hegel, oil painting by Jakob von Schlesinger, *c.* 1825. In the Staatliche Museen zu Berlin.

Early life. Hegel was born at Stuttgart on August 27, 1770, the son of a revenue officer. He had already learned the elements of Latin from his mother when he entered the Stuttgart grammar school, where he remained until he was 18. As a schoolboy he made a collection of extracts, alphabetically arranged, comprising annotations on classical authors, passages from newspapers, and treatises on morals and mathematics from the standard works of the period.

University and private studies

In 1788 Hegel went as a student to Tübingen with a view to taking orders, as his parents wished. Here he studied philosophy and classics for two years and graduated in 1790. Though he then took the theological course, he was impatient with the orthodoxy of his teachers; and the certificate given to him when he left in 1793 states that, whereas he had devoted himself vigorously to philosophy, his industry in theology was intermittent. He was also said to be poor in oral exposition, a deficiency that was to dog him throughout his life. Though his fellow students called him "the old man," he liked cheerful company and a "sacrifice to Bacchus" and enjoyed the ladies as well. His chief friends were a pantheistic poet, J.C.F. Hölderlin, his contemporary, and the nature philosopher F.W.J. Schelling, five years his junior. Together they read the Greek tragedians and celebrated the glories of the French Revolution.

On leaving college, Hegel did not enter the ministry; instead, wishing to have leisure for the study of philosophy and Greek literature, he became a private tutor. For the next three years he lived in Berne, with time on his hands and the run of a good library, where he read Edward Gibbon on the fall of Rome and *De l'esprit des loix*, by

Charles Louis, baron de Montesquieu, as well as the Greek and Roman classics. He also studied the critical philosopher Immanuel Kant and was stimulated by his essay on religion to write certain papers that became noteworthy only when, more than a century later, they were published as a part of *Hegels theologische Jugendschriften* (1907). Kant had maintained that, whereas orthodoxy requires a faith in historical facts and in doctrines that reason alone cannot justify and imposes on the faithful a moral system of arbitrary commands alleged to be revealed, Jesus, on the contrary, had originally taught a rational morality, reconcilable with the teaching of Kant's ethical works, and a religion that, unlike Judaism, was adapted to the reason of all men. Hegel accepted this teaching; but, being more of a historian than Kant, he put it to the test of history by writing two essays: first, a life of Jesus in which he attempted to reinterpret the gospel on Kantian lines and, second, an answer to the question of how Christianity had ever become the authoritarian religion that it was, if in fact the teaching of Jesus was not authoritarian but rationalistic.

Move to Frankfurt am Main

Hegel was lonely in Berne and was glad to move, at the end of 1796, to Frankfurt am Main, where Hölderlin had gotten him a tutorship. His hopes of more companionship, however, were unfulfilled: Hölderlin was engrossed in an illicit love affair and shortly lost his reason. Hegel began to suffer from melancholia and, to cure himself, worked harder than ever, especially at Greek philosophy and modern history and politics. He read and made clippings from English newspapers, wrote about the internal affairs of his native Wurtemberg, and studied economics. Hegel was now able to free himself from the domination of Kant's influence and to look with a fresh eye on the problem of Christian origins.

Emancipation from Kantianism. It is impossible to exaggerate the importance that this problem had for Hegel. It is true that his early theological writings contain hard sayings about Christianity and the churches; but the object of his attack was orthodoxy, not theology itself. All that he wrote at this period throbs with a religious conviction of a kind that is totally absent from Kant and Hegel's other 18th-century teachers. Above all, he was inspired by a doctrine of the Holy Spirit. The spirit of man, his reason, is the candle of the Lord, he held, and therefore cannot be subject to the limitations that Kant had imposed upon it. This faith in reason, with its religious basis, henceforth animated the whole of Hegel's work.

His outlook had also become that of a historian—which again distinguishes him from Kant, who was much more influenced by the concepts of physical science. Every one of Hegel's major works was a history; and, indeed, it was among historians and classical scholars rather than among philosophers that his work mainly fructified in the 19th century.

"The Spirit of Christianity"

When in 1798 Hegel turned back to look over the essays that he had written in Berne two or three years earlier, he saw with a historian's eye that, under Kant's influence, he had misrepresented the life and teachings of Jesus and the history of the Christian Church. His new-won insight then found expression in his essay "Der Geist des Christentums und sein Schicksal" ("The Spirit of Christianity and Its Fate"), likewise unpublished until 1907. This is one of Hegel's most remarkable works. Its style is often difficult and the connection of thought not always plain, but it is written with passion, insight, and conviction.

He begins by sketching the essence of Judaism, which he paints in the darkest colours. The Jews were slaves to the Mosaic Law, leading a life unlovely in comparison with that of the ancient Greeks and content with the material satisfaction of a land flowing with milk and honey. Jesus taught something entirely different. Men are not to be the slaves of objective commands: the law is made for man. They are even to rise above the tension in moral experience between inclination and reason's law of duty, for the law is to be "fulfilled" in the love of God, wherein all tension ceases and the believer does God's will wholeheartedly and single-mindedly. A community of such believers is the Kingdom of God.

This is the kingdom that Jesus came to teach. It is founded on a belief in the unity of the divine and the human. The life that flows in them both is one; and it is only because man is spirit that he can grasp and comprehend the Spirit of God. Hegel works out this conception in an exegesis of passages in the Gospel According to John. The kingdom, however, can never be realized in this world: man is not spirit alone but flesh also. "Church and state, worship and life, piety and virtue, spiritual and worldly action can never dissolve into one."

In this essay the leading ideas of Hegel's system of philosophy are rooted. Kant had argued that man can have knowledge only of a finite world of appearances and that, whenever his reason attempts to go beyond this sphere and grapple with the infinite or with ultimate reality, it becomes entangled in insoluble contradictions. Hegel, however, found in love, conceived as a union of opposites, a prefigurement of spirit as the unity in which contradictions, such as infinite and finite, are embraced and synthesized. His choice of the word *Geist* to express this his leading conception was deliberate: the word means "spirit" as well as "mind" and thus has religious overtones. Contradictions in thinking at the scientific level of Kant's "understanding" are indeed inevitable, but thinking as an activity of spirit or "reason" can rise above them to a synthesis in which the contradictions are resolved. All of this, expressed in religious phraseology, is contained in the manuscripts written toward the end of Hegel's stay in Frankfurt. "In religion," he wrote, "finite life rises to infinite life." Kant's philosophy had to stop short of religion. But there is room for another philosophy, based on the concept of spirit, that will distill into conceptual form the insights of religion. This was the philosophy that Hegel now felt himself ready to expound.

Career as lecturer at Jena. Fortunately, his circumstances changed at this moment, and he was at last able to embark on the academic career that had long been his ambition. His father's death in 1799 had left him an inheritance, slender, indeed, but sufficient to enable him to surrender a regular income and take the risk of becoming a *Privatdozent*. In January of 1801 he arrived in Jena, where Schelling had been a professor since 1798. Jena, which had harboured the fantastic mysticism of the Schlegel brothers and their colleagues and the Kantianism and ethical Idealism of J.G. Fichte, had already seen its golden age, for these great scholars had all left. The precocious Schelling, who was but 26 on Hegel's arrival, already had several books to his credit. Apt to "philosophize in public," Schelling had been fighting a lone battle in the university against the rather dull followers of Kant. It was suggested that Hegel had been summoned as a new champion to aid his friend. This impression received some confirmation from the dissertation by which Hegel qualified as a university teacher, which betrays the influence of Schelling's philosophy of nature, as well as from Hegel's first publication, an essay entitled "Differenz des Fichte'schen und Schelling'schen Systems der Philosophie" (1801), in which he gave preference to the latter. Nevertheless, even in this essay and still more in its successors, Hegel's difference from Schelling was clearly marked; they had a common interest in the Greeks, they both wished to carry forward Kant's work, they were both iconoclasts; but Schelling had too many romantic enthusiasms for Hegel's liking; and all that Hegel took from him—and then only for a very short period—was a terminology.

Relationship with Schelling

Hegel's lectures, delivered in the winter of 1801–02, on logic and metaphysics, were attended by about 11 students. Later, in 1804, with a class of about 30, he lectured on his whole system, gradually working it out as he taught. Notice after notice of his lectures promised a textbook of philosophy—which, however, failed to appear. After the departure of Schelling from Jena (1803), Hegel was left to work out his own views untrammelled. Besides philosophical and political studies, he made extracts from books, attended lectures on physiology, and dabbled in other sciences. As a result of representations made by himself at Weimar, he was in February 1805 appointed extraordinary professor at Jena: and in July

1806, on Goethe's intervention, he drew his first stipend —100 thalers. Though some of his hearers became attached to him, Hegel was not yet a popular lecturer.

Hegel, like Goethe, felt no patriotic shudder when Napoleon won his victory at Jena (1806): in Prussia he saw only a corrupt and conceited bureaucracy. Writing to a friend on the day before the battle, he spoke with admiration of the "world soul" and the Emperor and with satisfaction at the probable overthrow of the Prussians.

The Phenomenology of Mind

At this time Hegel published his first great work, the *Phänomenologie des Geistes* (1807; Eng. trans., *The Phenomenology of Mind*, 2nd ed., 1931). This, perhaps the most brilliant and difficult of Hegel's books, describes how the human mind has risen from mere consciousness, through self-consciousness, reason, spirit, and religion, to absolute knowledge. Though man's native attitude toward existence is reliance on the senses, a little reflection is sufficient to show that the reality attributed to the external world is due as much to intellectual conceptions as to the senses and that these conceptions elude a man when he tries to fix them. If consciousness cannot detect a permanent object outside itself, so self-consciousness cannot find a permanent subject in itself. Through aloofness, skepticism, or imperfection, self-consciousness has isolated itself from the world; it has closed its gates against the stream of life. The perception of this is reason. Reason thus abandons its efforts to mold the world and is content to let the aims of individuals work out their results independently.

The stage of *Geist*, however, reveals the consciousness no longer as isolated, critical, and antagonistic but as the indwelling spirit of a community. This is the lowest stage of concrete consciousness, the age of unconscious morality. But, through increasing culture, the mind gradually emancipates itself from conventions, which prepares the way for the rule of conscience. From the moral world the next step is religion. But the idea of Godhead, too, has to pass through nature worship and art before it reaches a full utterance in Christianity. Religion thus approaches the stage of absolute knowledge, of "the spirit knowing itself as spirit." Here, according to Hegel, is the field of philosophy.

Period as Gymnasium rector at Nürnberg. In spite of the *Phänomenologie*, however, Hegel's fortunes were now at their lowest ebb. He was, therefore, glad to become editor of the *Bamberger Zeitung* (1807–08). This, however, was not a suitable vocation, and he gladly accepted the rectorship of the Aegidiengymnasium in Nürnberg, a post he held from December 1808 to August 1816 and one that offered him a small but assured income. There Hegel inspired confidence in his pupils and maintained discipline without pedantic interference in their associations and sports.

In 1811 Hegel married Marie von Tucher (22 years his junior), of Nürnberg. The marriage was entirely happy. His wife bore him two sons: Karl, who became eminent as a historian; and Immanuel, whose interests were theological. The family circle was joined by Ludwig, a natural son of Hegel's from Jena.

At Nürnberg in 1812 appeared *Die objektive Logik*, being the first part of his *Wissenschaft der Logik* ("Science of Logic"), which in 1816 was completed by the second part, *Die subjektive Logik*.

Period as university professor. This work, in which his system was first presented in what was essentially its ultimate shape, earned him the offer of professorships at Erlangen, at Berlin, and at Heidelberg.

Hegel's philosophical system

At Heidelberg. He accepted the chair at Heidelberg. For use at his lectures there, he published his *Encyklopädie der philosophischen Wissenschaften im Grundrisse* (1817; "Encyclopaedia of the Philosophical Sciences in Outline"), an exposition of his system as a whole. Hegel's philosophy is an attempt to comprehend the entire universe as a systematic whole. The system is grounded in faith. In the Christian religion God has been revealed as truth and as spirit. As spirit, man can receive this revelation. In religion the truth is veiled in imagery; but in philosophy the veil is torn aside, so that man can know the infinite and see all things in God. Hegel's system is thus a spiritual monism but a monism in which differentiation is essential. Only through an experience of difference can the identity of thought and the object of thought be achieved—an identity in which thinking attains the through-and-through intelligibility that is its goal. Thus, truth is known only because error has been experienced and truth has triumphed; and God is infinite only because he has assumed the limitations of finitude and triumphed over them. Similarly, man's Fall was necessary if he was to attain moral goodness. Spirit, including the Infinite Spirit, knows itself as spirit only by contrast with nature. Hegel's system is monistic in having a single theme: what makes the universe intelligible is to see it as the eternal cyclical process whereby Absolute Spirit comes to knowledge of itself as spirit (1) through its own thinking; (2) through nature; and (3) through finite spirits and their self-expression in history and their self-discovery, in art, in religion, and in philosophy, as one with Absolute Spirit itself.

"Encyclopaedia of the Philosophical Sciences"

The compendium of Hegel's system, the "Encyclopaedia of the Philosophical Sciences," is in three parts: "Logic," "Nature," and "Mind." Hegel's method of exposition is dialectical. It often happens that in a discussion two people who at first present diametrically opposed points of view ultimately agree to reject their own partial views and to accept a new and broader view that does justice to the substance of each. Hegel believed that thinking always proceeds according to this pattern: it begins by laying down a positive thesis that is at once negated by its antithesis; then further thought produces the synthesis. But this in turn generates an antithesis, and the same process continues once more. The process, however, is circular: ultimately, thinking reaches a synthesis that is identical with its starting point, except that all that was implicit there has now been made explicit. Thus, thinking itself, as a process, has negativity as one of its constituent moments, and the finite is, as God's self-manifestation, part and parcel of the infinite itself.

This is the sort of dialectical process of which Hegel's system provides an account in three phases.

"*Logic.*" The system begins with an account of God's thinking "before the creation of nature and finite spirit"; *i.e.*, with the categories or pure forms of thought, which are the structure of all physical and intellectual life. Throughout, Hegel is dealing with pure essentialities, with spirit thinking its own essence; and these are linked together in a dialectical process that advances from abstract to concrete. If a man tries to think the notion of pure Being (the most abstract category of all), he finds that it is simply emptiness; *i.e.*, Nothing. Yet Nothing *is*. The notion of pure Being and the notion of Nothing are opposites; and yet each, as one tries to think it, passes over into the other. But the way out of the contradiction is at once to reject both notions separately and to affirm them both together; *i.e.*, to assert the notion of becoming, since what becomes both is and is not at once. The dialectical process advances through categories of increasing complexity and culminates with the absolute idea, or with the spirit as objective to itself.

"*Nature.*" Nature is the opposite of spirit. The categories studied in "Logic" were all internally related to one another; they grew out of one another. Nature, on the other hand, is a sphere of external relations. Parts of space and moments of time exclude one another; and everything in nature is in space and time and is thus finite. But nature is created by spirit and bears the mark of its creator. Categories appear in it as its essential structure, and it is the task of the philosophy of nature to detect that structure and its dialectic; but nature, as the realm of externality, cannot be rational through and through, though the rationality prefigured in it becomes gradually explicit when man appears. In man nature rises to self-consciousness.

"*Mind.*" Here Hegel follows the development of the human mind through the subconscious, consciousness, and the rational will; then through human institutions and human history as the embodiment or objectification of that will; and finally to art, religion, and philosophy, in which finally man knows himself as spirit, as one with

God and possessed of absolute truth. Thus, it is now open to him to think his own essence; *i.e.*, the thoughts expounded in "Logic." He has finally returned to the starting point of the system, but en route he has made explicit all that was implicit in it and has discovered that "nothing but spirit is, and spirit is pure activity."

Hegel's system depends throughout on the results of scientific, historical, theological, and philosophical inquiry. No reader can fail to be impressed by the penetration and breadth of his mind nor by the immense range of knowledge that, in his view, had to precede the work of philosophizing. A civilization must be mature and, indeed, in its death throes before, in the philosophic thinking that has implicitly been its substance, it becomes conscious of itself and of its own significance. Thus, when philosophy comes on the scene, some form of the world has grown old.

At Berlin. In 1818 Hegel accepted the renewed offer of the chair of philosophy at Berlin, which had been vacant since Fichte's death. There his influence over his pupils was immense, and there he published his *Naturrecht und Staatswissenschaft im Grundrisse*, alternatively entitled *Grundlinien der Philosophie des Rechts* (1821; Eng. trans., *The Philosophy of Right*, 1942). In Hegel's works on politics and history, the human mind objectifies itself in its endeavour to find an object identical with itself. *The Philosophy of Right* (or *of Law*) falls into three main divisions. The first is concerned with law and rights as such: persons (*i.e.*, men as men, quite independently of their individual characters) are the subject of rights, and what is required of them is mere obedience, no matter what the motives of obedience may be. Right is thus an abstract universal and therefore does justice only to the universal element in the human will. The individual, however, cannot be satisfied unless the act that he does accords not merely with law but also with his own conscientious convictions. Thus, the problem in the modern world is to construct a social and political order that satisfies the claims of both. And thus no political order can satisfy the demands of reason unless it is organized so as to avoid, on the one hand, a centralization that would make men slaves or ignore conscience and, on the other hand, an antinomianism that would allow freedom of conviction to any individual and so produce a licentiousness that would make social and political order impossible. The state that achieves this synthesis rests on the family and on the guild. It is unlike any state existing in Hegel's day; it is a form of limited monarchy, with parliamentary government, trial by jury, and toleration for Jews and dissenters.

The Philosophy of Right

After his publication of *The Philosophy of Right*, Hegel seems to have devoted himself almost entirely to his lectures. Between 1823 and 1827 his activity reached its maximum. His notes were subjected to perpetual revisions and additions. It is possible to form an idea of them from the shape in which they appear in his published writings. Those on *Aesthetics*, on the *Philosophy of Religion*, on the *Philosophy of History*, and on the *History of Philosophy* have been published by his editors, mainly from the notes of his students, whereas those on logic, psychology, and the philosophy of nature have been appended in the form of illustrative and explanatory notes to the corresponding sections of his *Encyklopädie*. During these years hundreds of hearers from all parts of Germany and beyond came under his influence; and his fame was carried abroad by eager or intelligent disciples.

Three courses of lectures are especially the product of his Berlin period: those on aesthetics, on the philosophy of religion, and on the philosophy of history. In the years preceding the revolution of 1830, public interest, excluded from political life, turned to theatres, concert rooms, and picture galleries. At these Hegel became a frequent and appreciative visitor, and he made extracts from the art notes in the newspapers. During his holiday excursions, his interest in the fine arts more than once took him out of his way to see some old painting. This familiarity with the facts of art, though neither deep nor historical, gave a freshness to his lectures on aesthetics, which, as put together from the notes taken in different years from 1820 to 1829, are among his most successful efforts.

Philosophies of aesthetics; of religion; of history

The lectures on the philosophy of religion are another application of his method, and shortly before his death he had prepared for the press a course of lectures on the proofs for the existence of God. On the one hand, he turned his weapons against the Rationalistic school, which reduced religion to the modicum compatible with an ordinary worldly mind. On the other hand, he criticized the school of Schleiermacher, who elevated feeling to a place in religion above systematic theology. In his middle way, Hegel attempted to show that the dogmatic creed is the rational development of what was implicit in religious feeling. To do so, of course, philosophy must be made the interpreter and the superior discipline.

In his philosophy of history, Hegel presupposed that the whole of human history is a process through which mankind has been making spiritual and moral progress and advancing to self-knowledge. History has a plot, and the philosopher's task is to discern it. Some historians have found its key in the operation of natural laws of various kinds. Hegel's attitude, however, rested on the faith that history is the enactment of God's purpose and that man had now advanced far enough to descry what that purpose is: it is the gradual realization of human freedom.

The first step was to make the transition from a natural life of savagery to a state of order and law. States had to be founded by force and violence; there is no other way to make men law-abiding before they have advanced far enough mentally to accept the rationality of an ordered life. There will be a stage at which some men have accepted the law and become free, while others remain slaves. In the modern world man has come to appreciate that all men, as minds, are free in essence, and his task is thus to frame institutions under which they will be free in fact.

Hegel did not believe, despite the charge of some critics, that history had ended in his lifetime. In particular, he maintained against Kant that to eliminate war is impossible. Each nation-state is an individual; and, as Hobbes had said of relations between individuals in the state of nature, pacts without the sword are but words. Clearly, Hegel's reverence for fact prevented him from accepting Kant's Idealism.

The lectures on the history of philosophy are especially remarkable for their treatment of Greek philosophy. Working without modern indexes and annotated editions, Hegel's grasp of Plato and Aristotle is astounding, and it is only just to recognize that it was from Hegel that the scholarship lavished on Greek philosophy in the century after his death received its original impetus.

Recognition and honours

At this time a Hegelian school began to gather. The flock included intelligent pupils, empty-headed imitators, and romantics who turned philosophy into lyric measures. Opposition and criticism only served to define more precisely the adherents of the new doctrine. Though he had soon resigned all direct official connection with the schools of Brandenburg, Hegel's real influence in Prussia was considerable. In 1830 he was rector of the university. In 1831 he received a decoration from Frederick William III. One of his last literary undertakings was the establishment of the Berlin *Jahrbücher für wissenschaftliche Kritik* ("Yearbook for Philosophical Criticism").

The revolution of 1830 was a great blow to Hegel, and the prospect of mob rule almost made him ill. His last literary work, the first part of which appeared in the *Preussische Staatszeitung* while the rest was censored, was an essay on the English Reform Bill of 1832, considering its probable effects on the character of the new members of Parliament and the measures that they might introduce. In the latter connection he enlarged on several points in which England had done less than many continental states for the abolition of monopolies and abuses.

In 1831 cholera entered Germany. Hegel and his family retired for the summer to the suburbs, and there he finished the revision of the first part of his *Science of Logic*. Home again for the winter session, on November 14, after one day's illness, he died of cholera and was buried, as

he had wished, between Fichte and Karl Solger, author of an ironic dialectic.

Personage and influence. In his classroom Hegel was more impressive than fascinating. His students saw a plain, old-fashioned face, without life or lustre—a figure that had never looked young and was now prematurely aged. Sitting with his snuffbox before him and his head bent down, he looked ill at ease and kept turning the folios of his notes. His utterance was interrupted by frequent coughing; every sentence came out with a struggle. The style was no less irregular: sometimes in plain narrative the lecturer would be specially awkward, while in abstruse passages he seemed especially at home, rose into a natural eloquence, and carried away the hearer by the grandeur of his diction.

The early theological writings and the *Phenomenology of Mind* are packed with brilliant metaphors. In his later works, produced as textbooks for his lectures, the "Encyclopaedia of the Philosophical Sciences" and the *Philosophy of Right*, he compresses his material into relatively short, numbered paragraphs. It is only necessary to translate them to appreciate their conciseness and precision. The common idea that Hegel's is a philosophy of exceptional difficulty is quite mistaken. Once his terminology is understood and his main principles grasped, he presents far less difficulty than Kant, for example. One reason for this is a certain air of dogmatism: Kant's statements are often hedged around with qualifications; but Hegel had, as it were, seen a vision of absolute truth, and he expounds it with confidence.

Emergence of Hegelianism

Hegel's system is avowedly an attempt to unify opposites—spirit and nature, universal and particular, ideal and real—and to be a synthesis in which all the partial and contradictory philosophies of his predecessors are alike contained and transcended. It is thus both Idealism and Realism at once; hence, it is not surprising that his successors, emphasizing now one and now another strain in his thought, have interpreted him variously. Conservatives and revolutionaries, believers and atheists alike have professed to draw inspiration from him. In one form or another his teaching dominated German universities for some years after his death and spread to France and to Italy. The vicissitudes of Hegelian thought to the present day are detailed in the article HEGELIANISM. In the mid-20th century, interest in the early theological writings and in the *Phänomenologie* was increased by the spread of Existentialism. At the same time, the growing importance of Communism encouraged political thinkers to study Hegel's political works, as well as his "Logic," because of their influence on Karl Marx. And, by the time of his bicentennial in 1970, a Hegelian renascence seemed to be in the making, at least in the United States and in Germany.

BIBLIOGRAPHY. A collected edition of Hegel's published works, together with a great deal of material culled from his lectures, was published by his pupils within a few years of his death in 1831. This edition, with some rearrangement, was reissued by HERMANN GLOCKNER in 26 volumes, including a comprehensive index (1927–40). In 1905 the Philosophische Bibliothek (Leipzig, later Hamburg) began publication of a new edition with a carefully revised text edited by GEORG LASSON and later by JOHANNES HOFFMEISTER; volumes appeared for over 50 years, but it was never completed. It has now been enhanced by a comprehensive edition sponsored by the Deutsche Forschungsgemeinschaft; it is to contain about 50 volumes. The first volume appeared in 1968.

English translations of most of Hegel's works were published in the later 19th and early 20th century, but, apart from those by W. WALLACE (*Logic* and *Mind*—*i.e.*, the first and third parts of the *Encyclopaedia*), they are not always very satisfactory and they have no notes. With a view to remedying this deficiency, new English translations have appeared of the *Philosophy of Right* (1942, often reprinted), the *Science of Logic* (1969), and the *Philosophy of Art* (1973), as well as translations of writings not translated previously, namely, the *Early Theological Writings* (1948; rev. ed., 1971), some *Political Writings* (1969), and, twice, the *Philosophy of Nature*—the second part of the *Encyclopaedia* (1970), the latter in three volumes. With the exception of the *Science of Logic* and the Oxford translation of the *Philosophy of Nature*, all these translations are annotated.

The best short account of Hegel's philosophy in English is still EDWARD CAIRD, *Hegel* (1883, reprinted 1968), but it has been brought up to date in certain respects by G.R.G. MURE, *The Philosophy of Hegel* (1965); and in more detail by W.A. KAUFMANN, *Hegel* (1965). An attempt to interest modern philosophers in Hegel is contained in J.N. FINDLAY, *Hegel* (1958), but this important and lively work is for consideration only by those already acquainted with Hegel. As an introduction to Hegel, G.R.G. MURE, *An Introduction to Hegel* (1940), is more reliable but it is not an exposition. The standard long exposition of Hegel's mature system is still KUNO FISCHER, *Hegels Leben, Werke und Lehre*, esp. as edited by H. FALKENHEIM (1911); while in English there is W.T. STACE, *The Philosophy of Hegel* (1924). This concentration on Hegel's mature system alone, as if it were self-explanatory, has long been outdated. In 1900 Wilhelm Dilthey maintained that Hegel could only be understood if there were a study of his early manuscripts and, on the basis of these, a history of his development written. He wrote such a history in *Die Jugendgeschichte Hegels* (1906). This seminal work, hardly noticed at all by writers in English before 1965, gave rise to an immense literature in Germany, France, and Italy. The most important and most brilliant study of the young Hegel may be that of GEORG LUKACS, *Der junge Hegel* (1948), although he writes from a Marxist point of view. But the most exhaustive is that of THEODOR HAERING, *Hegel: Sein Wollen und sein Werk*, 2 vol. (1929–38). Reliable and more readable than this are the two volumes of HERMANN GLOCKNER issued (1929 and 1940) as an appendix to his edition of the collected works. Even these, however, have been outdated by the flood of new material collected by the Hegel-Archiv at Bochum in West Germany and published in a series of volumes of *Hegel-Studien* (1961 and subsequent years). This material has not yet been condensed into monographs. The same is true of the four volumes of Hegel's letters, edited by JOHANNES HOFFMEISTER (1952–60). These could be used for the preparation of a new biography of Hegel to supersede the admirable, but now obsolete, one by K. ROSENKRANZ (1824).

The following are the most recent commentaries on individual works, or short guides to them: (*Phenomenology of Spirit*): JEAN HYPPOLITE, *Genèse et structure de la Phénoménologie de l'Espirit de Hegel* (1946). (*Logic*): G.R.G. MURE, *A Study of Hegel's Logic* (1950). (*Nature*): There are few, even short, studies in any language, but the apparatus in the English translation by M.J. PETRY, 3 vol. (1970), provides a full and learned commentary. (*Law, Morality, and the State*): WILHELM SEEBERGER, *Hegel; oder, die Entwicklung des Geistes zur Freiheit* (1961), which, incidentally, is a good introduction to Hegel's thought as a whole; but HUGH A. REYBURN, *The Ethical Theory of Hegel* (1921, reprinted 1967); and FRANZ ROSENZWEIG, *Hegel und der Staat*, 2 vol. (1920), have not been superseded. Both are excellent summaries, but the latter is a commentary as well. (*Art*): JACK KAMINSKY, *Hegel on Art* (1962), is a fair summary of Hegel's lectures. (*Religion*): T.M. KNOX, *A Layman's Quest* (1969), deals in chapters 5 and 6 not only with Hegel's lectures on the philosophy of religion but with all his other writings elsewhere on religion.

(T.M.K.)

Hegelianism

Hegelianism is the name given to a diversified philosophical movement that developed out of a monumental system of thought constructed during the first third of the 19th century by the German Idealist G.W.F. Hegel, generally considered to have been one of the most broadly and profoundly learned philosophers of modern times. The term is here so construed as to exclude Hegel himself and to include, therefore, only the ensuing Hegelian movements. As such, its thought is focussed upon history and logic, a history in which it sees, in various perspectives, that "the rational is the real" and a logic in which it sees that "the truth is the Whole."

GENERAL CONSIDERATIONS

Problems of the Hegelian heritage. The Hegelian system, in which German Idealism reached its fulfillment, claimed to provide a unitary solution to all of the problems of philosophy. It held that the speculative point of view, which transcends all particular and separate perspectives, must grasp the *one* truth, bringing back to its proper centre all of the problems of logic, of metaphysics (or the nature of Being), and of the philosophies of nature, law, history, and culture (artistic, religious, and philosophical). According to Hegel, this attitude is more than a formal method that remains extraneous to its own content; rather, it represents the actual development of

Dialectical development of the Absolute

the Absolute—of the all-embracing totality of reality—considered "as Subject and not merely as Substance" (*i.e.*, as a conscious agent or Spirit and not merely as a real being). This Absolute, Hegel held, first puts forth (or posits) itself in the immediacy of its own inner consciousness and then negates this positing—expressing itself now in the particularity and determinateness of the factual elements of life and culture—and finally regains itself, through the negation of the former negation that had constituted the finite world.

Such a dialectical scheme (immediateness–alienation–negation of the negation) accomplished the self-resolution of the aforementioned problem areas—of logic, of metaphysics, and so on. This panoramic system thus had the merit of engaging philosophy in the consideration of all of the problems of history and culture, none of which could any longer be deemed foreign to its competence. At the same time, however, the system deprived all of the implicated elements and problems of their autonomy and particular authenticity, reducing them to symbolic manifestations of the one process, that of the Absolute Spirit's quest for and conquest of its own self. Moreover, such a speculative mediation between opposites, when directed to the more impending problems of the time, such as those of religion and politics, led ultimately to the evasion of the most urgent and imperious ideological demands and was hardly able to escape the charge of ambiguity and opportunism.

Stages in the history of the interpretation of Hegel. The explanation of the success of Hegelianism—marked by the formation of a school that, for more than 30 years, brought together the best energies of German philosophy—lies in the fact that no other system could compete with it in the richness of its content or the rigour of its formulation or challenge its claim to express the total spirit of the culture of its time. Moreover, as Hegelianism diffused outward, it was destined to provoke increasingly lively and gripping reactions and to take on various articulations as, in its historical development, it intermingled with contrasting positions.

Four stages can be distinguished within the development of Hegelianism. The first of these was that of the immediate crisis of the Hegelian school in Germany during the period from 1827 through 1850. Always involved in polemics against its adversaries, the school soon divided into three currents: (1) The right, in which the direct disciples of Hegel participated, defended his philosophy from the accusation that it was liberal and pantheistic (defining God as the All). These "old Hegelians" sought to uphold the compatibility of Hegelianism with evangelical orthodoxy and with the conservative political policies of the Restoration (the new order in Europe that followed the defeat of Napoleon). (2) The left—formed of the "young Hegelians," for the most part indirect disciples of Hegel—considered the dialectic as a "principle of movement" and viewed Hegel's identification of the rational with the real as a command to modify the cultural and political reality that reactionism was merely justifying and to make it rational. Thus the young Hegelians interpreted Hegelianism in a revolutionary sense—*i.e.*, as pantheistic and then, consecutively, as atheistic in religion and as liberal democratic in politics. (3) The centre, which preferred to fall back upon interpretations of the Hegelian system in its genesis and significance, with special interest in logical problems.

Hegelian right, centre, and left

In the second phase (1850–1904), in which Hegelianism diffused into other countries, the works of the centre played a preponderant role; thus in this phase of the history of the interpretation of Hegel, usually called Neo-Hegelian, the primary interest was in logic and a reform of the dialectic.

In the first decade of the 20th century, on the other hand, there arose still in Germany a different movement, after Wilhelm Dilthey, originator of a critical approach to history and humanistic studies, discovered unpublished papers from the period of Hegel's youth. This third phase, that of the Hegel renaissance, was characterized by an interest in philology, by the publication of texts, and by historical studies; and it stressed the reconstruction of the genesis of Hegel's thought, considering especially its cultural matrices—both Enlightenment and Romanticist—and the extent to which it might present irrationalistic and so-called pre-Existentialist attitudes.

In the fourth stage, after World War II, the revival of Marxist studies in Europe finally thrust into the foreground the interest in Hegel–Marx relationships and in the value of the Hegelian heritage for Marxism, with particular regard to political and social problems. This fourth phase of the history of Hegelianism thus appropriated many of the polemical themes of the earlier years of the school.

CRISES IN THE EARLIER HEGELIAN SCHOOL

The earlier development of Hegelianism can be divided, according to predominant concerns, into three periods: (1) polemics during the life of Hegel (1816–31), (2) controversies in the religious field (1831–39), and (3) political debates (1840–44), though discussions on all of the problems continued through all three periods.

Polemics during the life of Hegel: 1816–31. While Hegel was still living, discussion was dominated by the master. It was not a matter of polemics within the school but only one of objections against the system from various quarters: from speculative theists; from Johann Herbart, a prominent student of the philosophy of mind, and his followers; and from disciples of Friedrich Schelling, an objective and aesthetic Idealist, and of Friedrich Schleiermacher, a seminal thinker of modern theology.

The substantive history of the school stems from Hegel's later teaching at Berlin and from the publication of his *Naturrecht und Staatswissenschaft im Grundrisse* (1821; Eng. trans., *The Philosophy of Right*, 1942). This book was reviewed by Herbart, who reprimanded Hegel for mixing the monism of the Rationalist Spinoza with the transcendentalism of Kant, which had explored the conditions of the possibility of knowledge in general. There were also certain critics who directed the liberal press against Hegel for attacking Jakob Fries, a psychologizing Neo-Kantian, in the introduction of *The Philosophy of Right*. Some of the polemical writings of Hegel made a notable impact—*e.g.*, a preface that he wrote for a book by one of his earliest disciples, Hermann Hinrichs, on the relation of faith to reason (1822). In this preface, Hegel saw the two things as the same in content but different in form—which for faith is the representation and for reason is the concept.

Articles in *Jahrbücher*

Particularly significant were eight articles in the *Jahrbücher für wissenschaftliche Kritik* (founded 1827; "Yearbooks for Scientific Critique"), a journal of the Hegelian Right. Important among these were a review by Hegel that was unexpectedly eulogistic about the thesis that philosophy and evangelical orthodoxy are compatible and another review in which Hegel responded indirectly to arguments of Herbart. Among Hegel's critics can be distinguished speculative theists such as Christian Weisse of Leipzig and Immanuel Fichte, the son of the more famous Johann Fichte, who reproached him for his panlogism and proposed to unify thought and experience in the concept of a free God, the Creator. Among the most loyal disciples of Hegel were Hermann Hinrichs, his collaborator and Karl Rosenkranz, who defended the Hegelian solution of the faith–reason problem (which had asserted the identity of content and difference of form), thus aptly defending the free rationality of religion.

Period of controversies chiefly in religion: 1831–39. The tone of these early polemics became animated and embittered after the death of Hegel. But, inasmuch as conditions in Germany, during the Restoration, inhibited the liberalization of political discussions, the milieu of controversy shifted to the religious realm and became related to problems of immortality, Christology, and general theology.

Feuerbach and Strauss

Shortly before Hegel's death, the youthful Ludwig Feuerbach, who later became a pioneer of naturalistic humanism, had published his *Gedanken über Tod und Unsterblichkeit* (1830; "Thoughts on Death and Immortality"), in which he contended that, from the Hegelian point of view, death must be necessary in order for man

to be transformed from the finite to the infinite and it is thus a privilege for man preferable to empirical personal survival. This work was held to confirm the charge of pantheism that orthodox adversaries had directed at Hegel's system. On this point, at the appearance of two volumes by Johann Friedrich Richter, a pantheist and critic of religion, Hegel's disciples intervened, in an argument employing not a few dialectical artifices, to conciliate Hegelian statements with the traditional doctrine of immortality.

The polarization of historical positions that the debate on immortality could not adequately express soon came into the open with *Das Leben Jesu kritisch bearbeitet* (1835–36; Eng. trans., *The Life of Jesus Critically Examined*, 1846), of David Friedrich Strauss, a biographer and radical theologian. This work brought the problem of the nature of Christ up to date from the point of view that had been reached by biblical criticism; *i.e.*, Christology was no longer an issue of denominational dogma but, rather, a problem of the interpretation and evaluation of the Gospel sources and of their meaning in the historical development of civilization. In this approach, the narrowly philological outlook was overcome by a reconstruction in terms of a philosophy of history strangely suggestive of the young Hegel. The thesis of the book was that the Gospel account is interwoven with myths that are not the works of individuals but of the collective poetic activity of the first Christian community, myths that resulted in part from messianic expectations, in part from the memory of the historical figure of Jesus, and in part from a transfiguration of the real elements. The aim of the myths was to demonstrate that philosophy and religion are the same in content and to offer, in an imaginative guise (as in parables), the meaning of the one truth that Substance is unification of the divine nature and of the human, which Christ symbolized and which is realized in the spirit of all humanity.

Strauss's work provoked a lively reaction, to which he replied in his *Streitschriften* (1837–38; "Controversial Writings"), proposing the image of a Hegelian school split, like the French Parliament, into a right (Göschel, and several others), a centre (Rosenkranz), and a left (Strauss himself). There were responses from the right and centre and from Bruno Bauer, a philosopher, historian, and biblical critic. From the anti-Hegelian side there was, above all, *Die evangelische Geschichte* (1838; "The History of the Gospels"), by Weisse, who, conceding to Strauss the necessity to rationalize the Gospel story, propounded a speculative interpretation of the Christ figure as an incarnation of the Logos (Word), in contrast to the mystic and pantheistic views.

Weisse, Ruge, Schelling

Meanwhile, Bauer shifted toward the left in a polemic against the orthodox Ernst Hengstenberg, a vehement accuser of the Hegelians, and in his *Kritik der Geschichte der Offenbarung* (1838; "Critique of the History of Revelation"). In 1838 was founded the earliest journal of the left, the *Hallische Jahrbücher für deutsche Wissenschaft und Kunst* ("Halle Yearbooks for German Science and Art"), coedited by the activist philosopher Arnold Ruge and T. Echtermeyer. At first, the journal maintained a moderate tone, and Hegelians of the centre and right also contributed articles. In June, however, it veered to the democratic-liberal side as Ruge struck out against an accuser of the young Hegelians and as Feuerbach attacked earlier Hegelians. Hegelianism, which marks the culmination of speculative philosophy, Feuerbach charged, does not demonstrate its own truth, because its contrast between sensory reality and intellectual concept comprises an irresoluble contradiction. Thus, its dialectic turns out to be a "monologue with itself," bereft of authentic mediation with the world. Hegelian philosophy, he held, is a "rational mystique," and what is needed is a return to nature, which, as objective reason, ought to become a principle of philosophy and of art. Thus an extensive examination of contemporary culture was conducted by the journal's editors in an article that depicted Romanticism as a movement degraded to a reactionary stance and extolled the spirit of reform and of liberal (yet royalist) Prussianism.

As for issues in the fields of logic and metaphysics, after several polemical exchanges the interest of philosophers was attracted to the publicist reawakening that came to Schelling, who reactivated certain anti-Hegelian criticisms. These criticisms dealt with the impossibility of building a valid philosophy upon the pure concept assumed as a point of departure and endowed with autonomous movement. Such a philosophy would be vitiated by presuppositions of what ought to be demonstrated and by hypostatizations (*i.e.*, the making of an idea into an entity). Schelling proposed, on the other hand, that the real itself be taken as the subject of development, to be grasped with a "lively intuition"; and that, while accepting a "negative philosophy" (such as that of Rationalism and Hegel) pointing to the conditions *without which* one cannot think, one must also add a "positive philosophy" delineating the conditions *by means of which* thought and reality can exist, premised on the existence of a free creative God.

Period of atheistic and political radicalism: 1840–44. The ensuing years marked one of the most intense periods in the cultural life of modern Europe.

Anti-Hegelian criticism. Advancing from Aristotelian presuppositions, an important critique against the Hegelian logic was presented by the classical philosopher and philologist Friedrich Adolf Trendelenburg in his *Logische Untersuchungen* (1840; "Logical Investigations"). In Hegel's view, the passage from Being to Nothing and to Becoming can be posited as a pure beginning "without presuppositions" of logic. In Trendelenburg's view, however, this passage is vitiated by its spurious dependence upon the surreptitious presupposition of the Empirical movement, without which support neither the passage from Being to Nothing (and vice versa) nor the recognition of Becoming as the "truth" of this primal opposition of concepts can be justified. Secondly, he charged that Hegel confused (1) the logical opposition or contradiction of A against non-A with (2) the real contradiction or contrariety of A against B. Contradiction (1) consists in the mere repetition of the first term with a negative sign; and from it no concrete movement can proceed. In contrariety (2), however, the opposition of the second term to the first is concrete—thus the second term cannot be deduced from the first and, instead, should be derived on its own account from empirical experience. Thus Hegel constructed his entire system, Trendelenburg charged, on an arbitrary dialectic of elements intrinsically real (contraries), which he mistakenly treated as though they were abstract opposites (contradictories) and were such by logical necessity.

Meanwhile, Schelling continued to teach his "positive philosophy"—of mythology and of revelation (of a personal God). Hence the philosophy of the later Schelling became the target of all of the criticisms from the left and likewise exerted a notable influence on the speculative theists. Meanwhile, the centre, on account of the critique of Trendelenburg, oriented itself toward the future reforms of Hegelianism.

Kierkegaard and Fischer

Among those who attended Schelling's lectures was Søren Kierkegaard, the man who was destined to become one of the founding fathers of Existentialism and whose religious individualism represents the earliest major result of the diffusion of Hegelianism outside of Germany. In all of his works—but above all in his *Philosophiske Smuler* (1844; Eng. trans., *Philosophical Fragments*, 1936) and his *Afsluttende uvidenskabelig Efterskrift* (1846; Eng. trans., *Concluding Unscientific Postscript*, 1941)—Kierkegaard waged a continuous polemic against the philosophy of Hegel. He regarded Hegel as motivated by the spirit of the harmonious dialectical conciliation of every opposition and as committed to imposing universal and panlogistic resolutions upon the authentic antinomies of life. Kierkegaard saw these antinomies as emerging from the condition of the individual, as a single person, who, finding himself always stretching to attain ascendance over his existential limitations in his absorption in God and at the same time always thrust back upon himself by the incommensurability of this relationship, cannot find his salvation except through the paradoxical in-

version of the rational values of speculative philosophy and through the "leap of faith" in the crucified Christ. Kierkegaard's claim that the nexus of problems characterizing man's condition as an existing being is irreducible to any other terms lay at the very roots of Existentialism. It was destined to condition the critical relationship of this current of thought to Hegelianism throughout its subsequent history. Moreover, Kierkegaard's thought—still more than that of Strauss—seemed reminiscent of those problem areas explored in the young Hegel's religious thought—issues that were destined to appear only later when Hegel research would gain precise knowledge of the writings of Hegel's youth.

At this time the attitude of the centre was oriented toward reforms of the Hegelian system in the field of logic and historiography, as reflected especially in the emergence of Kuno Fischer, one of the foremost historians of philosophy. In the fundamental triad of the dialectic, as Fischer saw it, Being and Nothing are not equally static and neutralizing. The real movement does not interpose itself into their relationship because Being is here to be understood as the Being of thought, which, to the degree that it is a thinking of Nothing, possesses that dynamic surplus that becomes manifest in the moment of Becoming. It was in making responses to this view that the forthcoming Neo-Hegelian movement in Europe found some of its motivations.

Theological radicalism. In 1840 political conditions in Germany changed with the succession of the young Frederick William IV, whose minister began to repress the liberal press and summoned to Berlin in an anti-Hegelian capacity both Schelling and the conservative jurist F.J. Stahl, a stubborn critic of Hegel. Far from weakening the movement, however, these actions radicalized its revolutionary manifestations. Strauss, in *Die christliche Glaubenslehre* (1840–41; "The Christian Doctrine of Faith"), reaffirmed the opposition of philosophical pantheism to religious theism as a means of reunifying the finite and the infinite; and Feuerbach established a philosophical anthropology in his major work *Das Wesen des Christentums* (1841; Eng. trans., *The Essence of Christianity*, new ed. 1957), in which man reappropriates his essence, which he had alienated from himself by hypostatizing it in the idea of God. The essence of man is reason, will, and love; and these three faculties comprise the consciousness of the human species as a knowledge of the infinity that man must regain. Man must thus reverse the theological propositions that express the spurious objectification of his universality in God; for this objectification had been effected through the individual consciousness in its effort to surmount its limitations. Thus Feuerbach interpreted the Christian mysteries as symbols of the alienation of human properties absolutized as divine attributes, and he criticized the contradictions of theology that are found in such concepts as God, the Trinity, the sacraments, and faith. Man's reappropriation of his essence from such religious alienation is consummated in the "new religion" of humanity, of which the supreme principle is that "Man is God to man."

Feuerbach and Bauer

To this period belong also the major critiques of Bruno Bauer on the Johannine (1840) and Synoptic (1841–42) Gospels. Differentiating his position from the pantheistic and mysticizing Substance of Strauss, Bauer held that the Gospels were not the unconscious product of the original community but a product of the self-consciousness of the Spirit in a given stage of its development. There followed two works specifically concerning Hegel, in which, feigning an orthodoxy from which he charged Hegel with atheism and radicalism, Bauer maintained, in the form of a parody, the revolutionary interpretation of Hegel that became customary in the current of the Hegelian left.

Sociopolitical radicalism. In the years 1841–43, the repressive measures of the government reached ever more decisive extremes: Bauer was debarred from teaching; Feuerbach did not even attempt to teach; and Ruge was enjoined to publish the *Hallische* in Prussia instead of Leipzig. (Actually, he transferred it to Dresden and changed its name to the *Deutsche Jahrbücher.*) Here also appeared one of Ruge's major writings, "Die Hegelsche Rechtsphilosophie und die Politik unserer Zeit" (1842; "The Hegelian Philosophy of Right and the Politics of our Time"), in which Ruge denounced Hegel's political conservatism, charging that his contemplative reason was reduced to the acceptance of existing conditions, to the exclusion of every effort to modify reality, and to the absolutizing of the Prussian state as the model of an ideal state. Ruge's journal was suppressed early in 1843, but in March he published in Switzerland his *Anekdota zur neuesten deutschen Philosophie und Publicistik* ("Anecdotes for the Latest German Philosophy and Political Journalism"), containing articles by Bauer, Ruge, Marx, Feuerbach, and others.

Feuerbach's article developed the claim that the method of speculative philosophy, which is the ultimate form of theology, is to invert the subject and predicate—*i.e.*, to substantialize the abstract and to treat concrete determinations as attributes or "logical accidents" of hypostatized abstractions. The inversion of speculative propositions, he held, leads to the philosophical reappropriation of man's essence; the philosophy of the future will achieve mastery through the negation of the Hegelian philosophy —and this is exactly what he entitled his forthcoming book: *Grundsätze der Philosophie der Zukunft* (1843; "Basic Principles of the Philosophy of the Future"). In place of the immediate Absolute of Hegel, he argued, there must be substituted the immediate individual existent—corporeal, sensible, and rational. Man's reappropriation of himself will be possible whenever his need to transcend his own limitations finds fulfillment in another person and in the totality of the human species: "thus man is the measure of reason."

Meanwhile, a schism had been ripening in the left wing: (1) On the one hand, there were the "Free Berliners" (initially the young Friedrich Engels, later to become Marx's theoretician, the radical anarchist Max Stirner, and the Bauer brothers), who, deeming themselves faithful to Hegel, developed a philosophy of self-consciousness (understood in a subjective and superindividualistic sense) directed toward treating social and historical problems with aristocratic intellectual detachment. (2) On the other hand, there was the group that included Ruge, the publicist Moses Hess, the scholarly poet Heinrich Heine, and Karl Marx. Influenced in their theories by Feuerbach, this group directed radicalism toward an experience deepened by the classical Enlightenment and embraced the rising Socialism. They thus involved Hegel in their critique of the political, cultural, and philosophical conditions of the time. The most widely known result of the first trend was Stirner's book *Der Einzige und sein Eigentum* (1845; "The Individual and His Property"), in which the fundamental thesis of individualistic anarchism can be discerned. The unique entity, in Stirner's view, is the individual, who must rebel against the attempt made by every authority and social organization to impose upon him a cause not his own and must be regarded as a focus of absolutely free initiative—a goal to be reached by emancipating himself from every idea-value imposed by tradition.

The work of Marx

The years between 1840 and 1844, however, saw the emergence of a figure incomparably more representative of the crisis of German Hegelianism than any already cited, that of Karl Marx, who was destined to guide the experience of this crisis toward a revolution of world-historical scope. Marx's study of Hegel dates from his university years in Berlin, the earliest result of which was his doctoral dissertation with the exceedingly important preparatory notes, in which he ventured an original application of Hegelian method to the problem of the great crises in the history of philosophy. At first a friend of Bauer, Marx clung closely, however, to the democratic wing of the left. In 1843 he completed an important critical study of Hegel's *Philosophy of Right*, in which he reproached Hegel for having absolutized into an ideal state the Prussian state of the time. Such absolutizing, he charged, lent itself to generalizations of broad critical scope with respect to the idealistic procedure of hypostatizing the Idea and brought about (as allegorical deriv-

atives from it) certain concrete political and social determinations, such as family, classes, and the state powers. Not yet a Communist, Marx nonetheless completed, in his *Kritik der hegelschen Staatsrechts* (written in the summer of 1843, published 1929; "Critique of Hegel's Constitutional Law"), a criticism of the erroneous relationship initiated in Hegel between society and the state, which was destined to lead Marx from the criticism of the modern state to that of modern society and its alienation.

It will be recalled that Hegel had likewise proposed the concept of alienation, describing the dialectic as a movement of the Absolute that was determined by its alienating and then regaining itself (thus overcoming the self-negation). Already in the *Economic and Philosophic Manuscripts of 1844* (German ed., 1932; Eng. trans., 1959), Marx had enunciated a general critique of the Hegelian dialectic that revealed its a priori nature, which, in Marx's view, was mystifying and alienated inasmuch as Hegel did nothing but sanction, by a method inverted with respect to real relationships, the alienation of all the concrete historical and human determinations.

Marx then directed himself against his former colleagues on the left—against Bauer in his *Die heilige Familie* (1845; Eng. trans., *The Holy Family*, 1956) and against Stirner in his *Die deutsche Ideologie* (1845–46; Eng. trans., *The German Ideology*, 1938), criticizing their "ideologism" (*i.e.*, the illusion that Idealism can be carried into the revolutionary camp since it is ideas that make history). The historical Materialism that Marx counterposed against Idealism expressed the conviction that the basis comprising the relations of production, both economic and social, conditions the superstructure of political, juridical, and cultural institutions and that the interchange among these spheres of production within the totality of an historical epoch must be designed to overcome their contradictions. This Materialism, though not belonging any more to Hegelianism, was destined nonetheless to remain linked to it by continuing polemical relationships and overlapping problem areas throughout the subsequent history of the movement.

Along with Marx must, of course, be mentioned his colleague Friedrich Engels, who was more tied, however, to the Hegelian conception of the dialectic—particularly regarding the dialectic of nature—than Marx was.

HEGELIANISM THROUGH THE MID-20TH CENTURY

Development and diffusion of Hegelianism in the later 19th century. In Germany, the second half of the 19th century witnessed a decline in the fortunes of Hegelianism, beginning with the *Hegel und seine Zeit* (1857; "Hegel and His Age"), by Rudolph Haym, a historian of the modern German spirit. The decline was urged on by Neo-Kantianism and Positivism as well as by the political realism of Bismarck. Hegelian influences still appeared in the first representatives of historicism (which urged that all things be viewed in the perspective of historical change). The surviving Hegelians, however, such as Kuno Fischer and Johann Erdmann, devoted themselves to the history of philosophy. Strauss and the Bauer brothers were won over to conservatism, and even Ruge, returning from exile in England, became a conservative.

Decline in Germany

Political and cultural problems: East Europe and the U.S. The diffusion of Hegelianism outside of Germany was oriented in two directions. With respect to its political and cultural problems, the Hegelian experience developed in east European philosophers and critics such as the Polish count Augustus Cieszkowski, a religious thinker whose philosophy of action was initially influenced by the left; and the theistic metaphysician Bronisław Trentowski. Among the Russians can be cited the literary critic Vissarion Belinsky, the democratic revolutionary writers Aleksandr Herzen and Nikolay Chernyshevsky, and certain anarchists such as the Russian exile and revolutionist Mikhail Bakunin. And among the French there were Hegelian Socialists such as Pierre-Joseph Proudhon.

In the United States, the interest in Hegelianism was stimulated by its political aspects and its philosophy of history. Its two centres, the St. Louis and Cincinnati schools, seemed to duplicate the German schism between a conservative and a revolutionary tendency. The former was represented by the Hegelians of the St. Louis school: the German Henry Brokmeyer and the New Englander William Harris, a pedagogue and politician, and the circle that they founded called the St. Louis Philosophical Society, which published an influential organ, *The Journal of Speculative Philosophy*. Their legitimism, or support for legitimate sovereignty, was expressed in the quest for a foundation, dialectical as well as speculative, for American democracy and in a dialectical interpretation of the history of the United States. The Cincinnati group, on the other hand, gathered around August Willich, a former Prussian officer, and John Bernard Stallo, an organizer of the Republican Party. Willich had participated in the Revolution of 1848 as a democratic partisan in south Germany, and, as an exile, had been in lively intercourse with Marx. He founded the *Cincinnati Republikaner,* in which he reviewed Marx's *Zur Kritik der politischen Ökonomie* (1859) and endeavoured to base the principles of social democracy upon the humanistic foundations of Feuerbach. Stallo, on the other hand, tried to interpret the political philosophy of Hegel in republican terms. The democratic community became, for him, the realization of the dialectic rationality of the Spirit with a rigorous separation of church and state.

St. Louis and Cincinnati schools

Logic and Metaphysics: Italy, England. The second trend in non-German Hegelianism was directed, in Italy and in England, to problems of logic and metaphysics. A vigorously speculative rethinking of the foundations of Hegel's *Wissenschaft der Logik* was engaged in by the major liberal Italian philosopher Bertrando Spaventa and his associates. Spaventa's *Studi sull' etica di Hegel* (1869) consisted of a direct liberal translation of Hegel's *Philosophy of Right.* Seeking to rediscover the connection between the thinking of the Italians of the 16th century and that of the German Idealists, Spaventa encountered the system of problems involved in the relationship between Kant and Hegel. He adopted from Kuno Fischer the solutions by which Fichte, Schelling, and Hegel had rendered Kant's transcendental ego consummatively veritable. He thus proposed an epistemological (idealistic theory of knowledge) interpretation of the Hegelian logic, according to which one premise of the logic is the dialectic of consciousness described in Hegel's *Phenomenology of Mind*, and the problems of the genesis of logic are resolved in the sense that Being is, from first to last, Becoming; *i.e.*, it is thought in action, which negates the objective residue of thought-out Being and, for that reason, is confirmed as a creative process. From Spaventa, whose intention was to vindicate the freedom and autonomy of thought against denominational dogmatism, was derived the foundation for the subjectivistic formalization of Hegelianism soon undertaken by Giovanni Gentile, an early-20th-century Idealist.

As in Italy, so also in England, interest in Hegel arose from the philosopher's need to round out his experience of classical German thought by tracing its vicissitudes since the time of Kant; and this interest was directed toward the fields of epistemology and logic and in this instance was applied to problems of religion and not of politics. The pioneer in English Hegelianism was James Hutchison Stirling, through his work *The Secret of Hegel* (1865). Stirling reaffirmed the lineage of thought that Fischer had traced "from Kant to Hegel," endeavouring to penetrate the dialectic–speculative relationship of unity in multiplicity as the central point of the dialectic. Toward Hegelianism as a unifying experience the ethics scholar Thomas Hill Green, the foremost representative of Hegelianism at Oxford, applied himself, though with more original attitudes; and the brothers John and Edward Caird dedicated themselves to right-wing interpretations of religious subjects—Edward in a well-known monograph entitled *Hegel* (1883).

Hegelianism in the first half of the 20th century. At this point, the development of Hegelianism branched out in two directions: one of which, in England and Italy,

pursued the tendencies of the Neo-Hegelians of the preceding decades, while the other, in Germany and France, accomplished the philological interpretative renewal known as the Hegel renaissance.

Neo-Hegelianism in England and Italy. With respect to the first tendency, there appeared in England at the turn of the century various outstanding works on Hegel's logic by authors who were partly Hegelian in spirit. These scholars, toiling through the system of problems that they shared—which focussed on establishing a criterion for the unification of the multiplicity of experience—ended up in diverse positions: those of Bernard Bosanquet and John Ellis MacTaggart, for example, who were translators and commentators of Hegelian works; but above all that of the foremost spiritualistic philosopher then in England, F.H. Bradley, author of the renowned *Appearance and Reality* (1893), whose development led him to positions more and more at odds with the absolute panlogism of Hegel. His affirmation of the dualism of appearance and reality was the result of a critique of the category of relations, which, by introducing contradictions between the qualities of the thing, utterly shattered the unity of experience in which it might seem that true reality could be reached—a reality that in Bradley's view it is not given to thought to attain.

Bradley, Royce, Croce

The echoes of this Idealistic system were not long in being felt in the United States by one of its most profound philosophers, an absolute Idealist, Josiah Royce, who, in *The World and the Individual* (1900–01), discussed the skeptical Idealism of Bradley in order to overthrow its consequences in favour of a conception of the infinite as a self-representative system and of the world (or the All) as an individualized realization of the intentional aims of the Idea copresent in a superior eternal consciousness. In Anglo-Saxon Neo-Hegelianism, the Hegelian experience has always been merely an episode—which fact serves to refine, by contrast, the methods of experimentalism that are more congenial to the Empirical tradition in England.

In Italy, on the other hand, the Neo-Hegelianism of the 20th century took the form of a spiritualistic reaction to the spread of Positivism that had followed upon the unification of Italy. This reaction developed in two directions: that of the historicism of Benedetto Croce and that of the actualism of Giovanni Gentile, two scholars who divided the realm of philosophy between themselves and occupied it—rather heavy-handedly—for four decades. The Crocean reform of Hegelianism dates from his volume *Ciò che è vivo e ciò che è morto della filosofia di Hegel* (1907; "What Is Living and What Is Dead in the Philosophy of Hegel") and from the systematic works of his so-called "philosophy of the spirit." Croce accepted the dialectic from Hegel as a requirement for the unification of opposites; but he rejected its system, in which Hegel would put in opposition and treat dialectically certain intellectual forms that are not really opposite but only distinct—such as the beautiful, the true, the useful, and the good, each of which has its dialectical opposite over against itself that it has to overcome within the purview of each grade. Consequently, renouncing the possibility of a philosophy of nature or of history, Croce formulated a development of so-called "distinct grades" according to the spiritual forms of art, of philosophy, of economics, and of ethics and contended that the comprehensive meaning of the development of the Spirit is given by history "as thought and as action" and a realization of freedom.

Gentile, on the other hand, accentuated the opposition of subject and object by considering every objective factuality as surpassed by the living dialectical development of the act—*i.e.*, the becoming of the Spirit in its own self-making, proceeding from an originating self-establishment, or *autoktisis*, of the Spirit itself. From this position he derived an absolute subjectivism that exploited all the possibilities for dialectically transforming every fixed position into its opposite, a downright sophistry of disengagement. Gentile's pro-Fascist stance, however, condemned his actualism to collapse.

Hegelian renaissance in Germany and France. Already from the beginnings of the century, however, there had been in Germany a change in Hegelian interpretation instigated by Wilhelm Dilthey's re-examination, in 1905, of the youthful manuscripts of Hegel and by the publication by one of Dilthey's principal disciples, Herman Nohl, of *Hegels theologische Jugendschriften* (1907; "The Theological Writings of Hegel's Youth"). Inasmuch as there had been heretofore only fragmentary notices on these unpublished literary remains, the effect of this rereading of the texts was to place them in contrast with the works of his maturity; they thus emerged as dealing, for the most part, with various problem areas in ethics, religion, and history; as lacking systematic preoccupations; and as rich discourse, tending to the mystic, which invited their comparison with the severe technical uniformity of his major works. Hermeneutical interest, however, centred especially on the problem of the beginnings of the philosophy and dialectic of Hegel, of which the first formulations were investigated in order to collate their meanings with those of the major works and of the *Phenomenology*, which was a key work of the Hegelian evolution inasmuch as it participated both in the romanticized colouring of the youthful writings and in the systematic demands of the *Encyklopädie der philosophischen Wissenschaften im Grundrisse*.

Publication of the *Jungendschriften*

Scholars were soon led to investigate the historical matrices of Hegel's intellectual culture—the late Enlightenment and dawning Romanticism—a direction of inquiry that yielded imposing contributions rich in discussions that continue to this day. These studies began with Dilthey's monograph, which pointed out the irrationalistic and vitalistic aspects of Hegel's youthful writings. In addition, a basic work by Franz Rosenzweig, *Hegel und der Staat* (1920), genetically reconstructed the political thought of the young Hegel in relation to its historical sources and concluded that the influence of Rousseau prevented Hegel from becoming the genuine "national philosopher of Germany." Jean Wahl, a French metaphysician and historian of philosophy, wrote on the "wretched conscience," interpreting Hegel existentially. Further, the German philosopher Richard Kroner studied the development from Kant to Hegel integrating it with the contributions of early Romanticism. And Hermann Glockner, a Bavarian aesthetic intuitionist, saw following one another in the development of Hegel a so-called "pantragistic" phase up to the *Phenomenology* and, subsequently, an opposing "panlogistic" phase that betrayed the most lively and concrete instances of the preceding phase—a work that approached the efforts at interpreting Hegel that were made by the Nazis.

HEGELIAN STUDIES TODAY

Today one has to speak not of the presence of Hegelianism as an operating philosophical current but only of studies on Hegel and of an experience of the Hegelian philosophy, to which, however, almost none of the present-day orientations in philosophy is foreign. The repeated encounter of Western culture with Marxist thought after World War II has brought to the fore the political, ethical, and religious implications of Hegelianism; and a marshalling into opposing camps analogous to that of the earlier crisis of the school is taking shape. Today there are no orthodox Hegelians, but there are denominational critics of Hegelianism, especially Catholic, whose cognizance of Hegel's painful development invokes, despite their differences, a certain fellow feeling with him.

In the centre are found scholars of a liberal and radical frame of mind but with varying orientations with respect to historical interpretations. Karl Löwith, a German philosopher of history and culture, sees Hegel as the initiator of the "historicist" crisis in modern thought, culminating in Marx and in Kierkegaard; and to this he contrasts the metahistorical perspective reflected in the Nietzschean motif of the "eternal return," based on the ideal of a Goethean serenity. In France, Alexandra Kojève, noteworthy for his effort to harmonize Hegel with Martin Heidegger, proposes a reinterpretation of the *Phänomenologie* as a manifesto of the emancipation of "man the servant" from all alienations. Jean Hyppolite, author

of an outstanding commentary on the *Phänomenologie*, usually presents a restrained humanistic interpretation of the Hegel of Jena. This renaissance of the study of Hegel has conditioned the thought of some of the major thinkers of France. Particularly notable, however, is the Hegelian conditioning of German philosopher-sociologists such as Theodor Wiesengrund Adorno and Herbert Marcuse. The former is sometimes regarded as the most Hegelian thinker of the mid-20th century because he sought to bring again to the fore Hegel's dialectic, understood in a new anti-intellectualistic sense, as a method for the solution of present-day social problems. Marcuse, a partisan of a Diltheian interpretation, approaches the position of the first Hegelian left, ending up in what critics see as a neoromantic anarchism. The major merit of both of these thinkers lies in their incisive analyses of aspects of modern consumer societies, especially American—though their proposed remedies remain uncertain.

Three currents of Marxist interpretation

The major interest, however, in the contemporary interpretation of Hegel is displayed by the Marxist camp. Marxist interpretation of Hegel had permeated the entire history of Hegelianism (notwithstanding the fact that the critical activity of young Marx against Hegel had been vehemently conducted and had led to various effects). This interpretation had settled upon the distinction made by Engels between the method and the system of Hegel's philosophy—*i.e.*, between the dialectic considered as a revolutionary "principle of movement" that achieves fulfillment in human culture and regarded, on the other hand, as reactionary because idealistic and conservative. With varying emphases on critical issues, this interpretation was continued in subsequent Marxist thinkers—from the Russians Georgy Plekhanov and Lenin to Mao Tse-tung and Stalin—the latter of whom affirmed the complementariness of historical and dialectical Materialism.

Today many Marxist scholars, especially in the countries of eastern Europe, remain favourable to the traditional line of Engels; and above all György Lukács, a Hungarian philosopher and literary critic and author of a volume on the young Hegel, does so. With the intention of revealing the romantic and irrationalistic presuppositions of Naziism, Lukács re-evaluates, in German culture, the tendency of the Enlightenment and of democracy, which he recognizes in the young Goethe, in Schiller, in Hölderlin, and in the young Hegel—in whom he sees, however, a reactionary involution.

A secondary tendency, which is drawing attention in France, with the work of Louis Althusser, draws Marx close to Structuralism, a recent school that seeks, through a "human science," to probe the systematic structures evinced in cultural life. In this school Marx's humanism is viewed as a temporary, Feuerbachian phase, surpassed by commitment to the scientific observation of the structure of bourgeois society. Such Structuralistic interpretation of Marxism thus runs the risk of departing from a due emphasis on the *historical* substance of Marxian Materialism.

The latter motive is, on the other hand, the essential aim of a third Marxist current, in Italy, initiated by Galvano della Volpe, a critical aesthetician who discusses the relationship between bourgeois and Socialist democracy and champions, in aesthetics, a critical and antiromantic Aristotelianism. This current has been continued by Mario Rossi, who asks one to read again in full the texts of Hegel and Marx, to reconstruct the related movements, and to compare the Materialistic conception of history with more recent philosophical currents such as Structuralism, present-day sociology, and the logic of the sciences.

A conclusion of a theoretical–systematic nature concerning Hegelianism has today become not only impossible but also inopportune, because its possible interest has been effectively replaced by that of the sheer history of the movement. The latter has shown how the substantial ambiguity of the philosophy and dialectic of Hegel can be resolved only when its claim to be able to solve all problems on a theoretical level and to achieve a "circular" decisiveness in its arguments—which violates the conditioning specificity of historical facts—is refuted. It is then the scholar's task to explore the limits of Hegel's thought as well as its conditioned inadequacies—but also its merits, which are above all those of having expressed and documented the major part of the cultural problems of modern civilization.

BIBLIOGRAPHY

Critical works: Works presenting a critical consideration of Hegelianism viewed as a whole are very few. See, however: STEPHAN D. CRITES, "Hegelianism," in the *Encyclopedia of Philosophy*, vol. 3, pp. 451–459 (1967); MARIO ROSSI, *Introduzione alla storia delle interpretazioni di Hegel* (1953), and *Da Hegel a Marx*, 2 vol. (1970); and RENE SERREAU, *Hegel et l'hégélianisme*, 3rd ed. (1968).

Historical works: J.E. ERDMANN, *Darstellung der Deutschen Philosophie seit Hegels Tode* (1963); WILLY MOOG, *Hegel und die Hegelsche Schule* (1930); KARL LOWITH, *Von Hegel zu Nietzsche*, 3rd ed. (1953; Eng. trans., *From Hegel to Nietzsche: The Revolution in Nineteenth-Century Thought*, 1964); the two anthologies edited by KARL LOWITH, *Die Hegelsche Linke* (1962), and HERMANN LUBBE, *Die Hegelsche Rechte* (1962), on the Left and Right, respectively.

In various countries: (*Germany*): HEINRICH LEVY, *Die Hegel-Renaissance in der deutschen Philosophie* (1927). (*Italy*): MARIO ROSSI (ed.), *Sviluppi dello Hegelismo in Italia* (1957); BENEDETTO CROCE, *Saggio sullo Hegel*, 5th ed. (1967). (*Slavic countries*): Contributions of authors from Russia, Poland, the Balkans, and Czechoslovakia are presented in *Hegel bei den Slaven*, 2nd ed., ed. by DMITRIJ TSCHIZEWSKIJ (1961); see also BORIS JAKOWENKO, *Ein Beitrag zur Geschichte des Hegelianismus in Russland* (1934). (*England*): HIRA-LAL HALDAR, *Neo-Hegelianism* (1927). (*United States*): LOYD D. EASTON, "Hegelianism in Nineteenth-Century Ohio," *Journal of the History of Ideas*, 23:355–378 (1962), for the Cincinnati school; HENRY A. POCHMANN, *German Culture in America: Philosophical and Literary Influences 1600–1900*, pp. 257–294 (1957), for the St. Louis school.

Other works: A. CORNU, *Karl Marx et Friedrich Engels*, 2 vol. (1955–58), is very rich in materials and citations from the *Hallische* and *Deutsche Jahrbücher*. See also HERBERT MARCUSE, *Reason and Revolution: Hegel and the Rise of Social Theory*, 2nd ed. (1954); SIDNEY HOOK, *From Hegel to Marx* (1950); and *The Monist*, vol. 48, no. 1 (1964), of which the entire issue is on the topic "Hegel Today."

(M.R.)

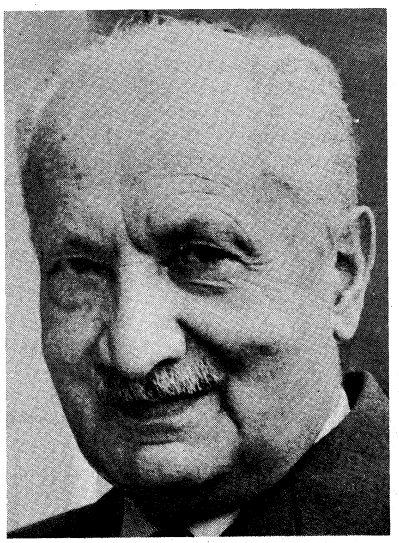
Heidegger, Martin

Martin Heidegger has been called the most original thinker in the field of contemporary German philosophy. As a leading exponent of Existentialism in the broad sense of that term, he exerted a profound influence on a younger generation of continental European cultural personalities. A critic of technological society and of the role of science, Heidegger has no place for God, whose absence nevertheless plays an important role in his thinking. He does not exalt human goals but sees human existence as a cult of Being—a notion not very unlike certain notions of God. Heidegger invented a strange and difficult vocabulary in order to obviate the trivializing effect of traditional philosophic jargon; but he believed, on the other hand, that some of the most common words, like "to dwell" and "to see," can reveal basic truths about man's existence. Heidegger has made the study of human existence subservient to more profound purposes—to the quest for the meaning of Being conceived as the fundamental principle of reality. This concern is also revealed in the central issue explored in his *Einführung in die Metaphysik* (1953; Eng. trans., *An Introduction to Metaphysics*, 1959); namely, the question "Why is there anything at all and not rather nothing?" Heidegger is thus the most influential ontologist, or student of "Being" as Being, among contemporary philosophers.

Background and youth. Heidegger was born September 26, 1889, in the small town of Messkirch, in Baden, southwest Germany. The son of a Catholic sexton, Heidegger showed an early interest in religion and, upon finishing high school, joined the Jesuits as a novice.

His education

At the University of Freiburg, he studied Catholic theology and medieval Christian philosophy. In fact, his interest in philosophy had already begun when, at secondary school, he started an intensive study of the late 19th-



Heidegger.
Camera Press—Pix

century Catholic philosopher Franz Brentano, author of a "descriptive" psychology, as presented in Brentano's *Von der mannigfachen Bedeutung des Seienden nach Aristoteles* (1862; "On the Manifold Meaning of Being According to Aristotle").

For the rest of his life Heidegger was to contemplate the possibility that there is a basic sense of the verb "to be" that lies behind its variety of usages. From his early study of Brentano also stems his enthusiasm for the Greeks, especially the pre-Socratics, whose thought marks the dawning of the penetrating reflection that transpired before the cleavage of thinking into poetry, philosophy, and science occurred.

The philosophy of Heidegger is obviously dependent upon the philosophers prior to Socrates, upon Plato and Aristotle, and upon the Gnostics, who claimed secret knowledge. He was particularly influenced, however—positively or negatively—by several 19th- and early 20th-century philosophers: by the Danish theological thinker Søren Kierkegaard and the Dionysian vitalist Friedrich Nietzsche, founders of Existentialism; by the historical vitalist Wilhelm Dilthey, noted for directing the attention of philosophers to the human and historical sciences; and by the founder of Phenomenology, Edmund Husserl.

When still in his 20s, Heidegger studied at Freiburg with Heinrich Rickert, later of the southwest school of axiological Kantianism, and with Husserl, who was then already famous; and for five years he was schooled in the Phenomenological method, in which the philosopher closely examines immediate experience and its objects in their essential aspects. Husserl's Phenomenology, and especially his struggle against the intrusion of psychology into essential studies of man—which he felt should, instead, be conducted on the philosophical level—determined the background of the young Heidegger's doctoral dissertation, completed in 1914. Consequently, what Heidegger later said and wrote about anxiety, thinking, forgetfulness, curiosity, distress, care, or awe was not meant as psychology; and what he said about man, publicness, and other-directedness was not intended to be sociology, anthropology, or political science. His utterances were meant, instead, to disclose ways of Being.

His magnum opus: "Being and Time." Heidegger started teaching at the University of Freiburg during the winter semester of 1915 and earned his habilitation through a study of the 13th-century British Franciscan philosopher Duns Scotus. In this position, now as a colleague of Husserl, Heidegger was expected to carry the Phenomenological movement further along within the spirit of his former master. As a religiously inclined young man, however, he went his own way instead and in 1927 astonished the German philosophical world with an almost unreadable work, *Sein und Zeit* (Eng. trans., *Being and Time*, 1962)—a work that, however, was immediately felt to be of prime importance, whatever its relation to Husserl might be. In spite of, or perhaps partly because of, its intriguingly difficult style, this book, which was Heidegger's magnum opus, was acclaimed as a deep and important work not only in German-speaking countries but also in Latin countries, where Phenomenology was already well known. It strongly influenced Sartre in France and other Existentialists; and despite Heidegger's protestations, he was classed, on the strength of this book, as the leading atheistic Existentialist. Its reception in the English-speaking world, however, was rather chilly, and its influence negligible until the 1960s. Since then, more and more universities have offered courses on Heidegger, and careful translations and commentaries have begun to appear. The influence of this book up to 1970 was still profound, but hardly stirring.

Heidegger's notion of the meaning of Being

In *Being and Time*, Heidegger's declared purpose is to bring to light what it means for a man to be, or, more accurately, *how* it is to be. This leads to a more fundamental question: what it means to ask "What is the meaning of Being?" These questions lie behind the obviousness of everyday life and, therefore, also behind the empirical questions of natural science. They are usually overlooked because they are too near to be grasped in everyday life. One might say that the whole prophetic mission of Heidegger amounts to making each man ask that question with maximum involvement. Whether he will ever arrive at any definite answer or not is, in the present crisis of mankind, of secondary importance.

This crisis, according to Heidegger, stems from a deep fall (*Verfall*) that Western thought has undergone, owing to a one-sided technical development, a development that results in alienation (*Entfremdung*), or, as expressed in terms more central to Heidegger's thought, in a "highly inauthentic way of being." Fallenness, or inauthenticity, belongs to the inescapable way of human existence; *i.e.*, it is an existential, an essential, potentiality (*Möglichkeit*), but epochs and individuals may be coloured by it in different degrees. This somewhat stern outlook has been mitigated, however, in Heidegger's later writings, in which he suggests that there are possibilities of redemption by "thinking of Being" and, thus, again coming closer to Being—a process in which, he believes, continental European rather than Eastern or Western countries are to lead the way.

The wealth of ideas contained in *Being and Time* is best discussed, however, in conjunction with those developed in another, short work, *Was ist Metaphysik?* (1929; Eng. trans., *What is Metaphysics?*, 1949). At the time of publishing *Being and Time*, Heidegger had been a professor ordinarius at Marburg for several years (since 1923). He resigned that post and, in 1928, returned to Freiburg, this time as Husserl's successor. *What Is Metaphysics?* was Heidegger's inaugural lecture; it elaborates one of his favourite themes, *das Nichts* ("nothing"); *i.e.*, the no-thing (see below).

Phenomenological and etymological methods

As Heidegger learned from Husserl, it is the phenomenological and not the scientific method that unveils man's ways of Being. Thus, in pursuing this method, Heidegger comes into conflict with the dichotomy of the subject–object relation, which has traditionally implied that man, as knower, is something (some-thing) within an environment that stands over against him. This relation, however, must be transcended. The deepest knowing, on the contrary, is a matter of *phainesthai* (Greek: "to show itself" or "to be in the light"), the word from which phenomenology, as a method, is derived. Something is just "there" in the light. Thus, the distinction between subject and object is not immediate but comes only later through conceptualization, as in the sciences.

As an aid in the effort to get back to "Thinking of Being" and its redemptive effects, Heidegger employs linguistic or hermeneutical techniques. He develops his own German, his own Greek, and his own kind of etymologies. He coins, for example, about 100 new complex words ending with "-being." In reading his works one must, thus, translate many of its key terms back into Greek words and then consider his free, often very spe-

cial (but never uninteresting), interpretations and etymologies.

Man stands out (*ex-sists*, not merely *ex-ists*) from things, says Heidegger in *Being and Time*, never being completely absorbed by them, but nevertheless being nothing (no-thing) apart from them. Man dwells in a world that he has been and continues to be thrown into until death. Being thrown into things, being-there (*Da-sein*), he falls away (*Verfall*) and is on the point of being submerged into things. He is continually a pro-ject (*Ent-wurf*); but periodically, or even normally, he may be submerged in things to such a degree that he is temporarily absorbed (*Aufgehen in*). He is then nobody in particular; and a structure that Heidegger calls *das Man* ("the they") is revealed, which recalls certain Anglo-American sociological criticisms of modern industrial society that stress man's "other-directedness," his tendency to measure himself in terms of his peers. But Heidegger's phenomenological metaphors avoid social science terms as much as possible in favour of ontological terms. Characteristic of *das Man* are idle talk (*Gerede*) and curiosity (*Neugier*). In *Gerede*, talker and listener do not stand in any genuine personal relation or in any intimate relation to what is talked about; hence, it leads to shallowness. Curiosity is a form of distraction, a need for the "new," a need for something "different," without real interest or capability of wonder.

Anxiety and freedom

But there is a mood, anxiety or dread (*Angst*), that functions to disclose (dis-close) authentic being, freedom (*Frei-sein*), as a potentiality. It manifests the freedom of man to choose himself and take hold of himself. The relevance of time, of the finiteness of human existence, is then experienced as a freedom to meet his own death (*das Freisein für den Tod*), a preparedness for and continuous relatedness to his own death (*Sein zum Tode*). In anxiety, all entities (*Seiendes*) sink away into a "nothing and nowhere," man hovers in himself as ex-sisting, being nowhere at home (*Un-heimlichkeit, Un-zu-hause*). He faces no-thing-ness (*das Nichts*); and all average, obvious everydayness disappears—and this is good, since he now faces the potentiality of authentic being.

Thus, the "sober" (*nüchtern*) anxiety and the implied confrontation with death are for Heidegger primarily of methodological importance: fundamentals are revealed. Among the structures revealed are potentialities for being joyfully active (". . . knowing joy [*die wissende Heiterkeit*] is a door to the eternal"). Anxiety opens man up to Being. But this does not imply that Being partakes in the dark aspect of dread; Being is associated with "light" and with "the joyful" (*das Heitere*). Being "calls the tune"; "to think Being" is to arrive at one's (true) home. Though Heideggerian students are often baffled by just what Being and thinking stand for, it is clear that Heidegger opposes a mere cult of mankind and wishes to call attention to something greater.

Later life. In the early 1930s, there occurred an event in the thought of Heidegger that scholars call his *Kehre* ("turning around"), which is said by some specialists to involve a turning away from the problem of *Being and Time*. This was denied by Heidegger himself, who insisted that he had been asking the same basic question since his youth. But in his later years he clearly became more reluctant to offer any answer. He did not even indicate a way in which to reach an answer to the basic problem of *Being and Time*.

At about the time of the *Kehre*, there also occurred Heidegger's short but eloquent pro-Nazi participation in the cultural politics of the Third Reich, which became a matter of considerable controversy. Even before Hitler assumed power in November 1933, German universities were exposed to heavy pressures. They were supposed to support the "national revolution" and eliminate Jewish scholars and doctrines (such as relativity). The anti-Nazi scientist who had been the rector at Freiburg resigned in protest, and the teaching staff unanimously elected Heidegger as his successor.

Heidegger's inauguration speech ("The German University's Self-Affirmation") was widely declared to be an affirmation of Nazism. To be sure, he divided student tasks into work service, military service, and scientific service; but this fell within the area of the authoritarian educational policy of Plato; and the speech ended not with a "*Heil*, Hitler!" but with a quotation from Plato's *Republic:* "All great things stand in peril." The speech turned against scientific specialization; it urged the asking of the question "What is it to be?"; and it warned against losing oneself in "things" (*Seiendes;* opposite *das Sein*). On other occasions, however, Heidegger gave solidly pro-Hitler speeches. "The Führer himself," he said, "and he alone is the German reality, present and future, and its law." In short, Heidegger succumbed to Hitlerism, but not to Nazi cultural policy or philosophy.

Under some pressure, Heidegger joined the Nazi Party and did not try to leave it. His relations to the party, however, and to the whole Nazi environment rapidly deteriorated. He resigned as rector as early as the beginning of 1934. After World War II, Heidegger characterized Hitlerism as the historical explosion of a structural sickness in mankind as a whole and expressed concern that it would take time to get rid of the poison.

In November 1944 Heidegger terminated his university lectures, and in 1945 the occupying powers forbade him to take up official lecturing again. He was "investigated"; but his support of Hitler in 1933–34 was not found to be of the serious, "active" kind, and he did not lose his professional rights. His status remained a matter of controversy, however, until he reached the age of retirement in 1959. Nevertheless, he gave influential regular lectures in the years 1951–58, and his attitude in 1933–34 did not affect his strong position within the international Phenomenological movement.

Evaluation

Perhaps the specific pretension of Heidegger's phenomenological method rests on a grandiose illusion, and perhaps the search for thinking Being is merely a disguised quest for a kind of belief in God; perhaps his abstruse terminology is only a mask covering and mystifying a more traditional approach. Such irreverent evaluations would scarcely be unsympathetic to Heidegger, if joined with the intent to verify or falsify it by genuinely following his own path through his writings. After all, he asks, or rather, provokes, us *to question*, not to listen to any answers. It is, therefore, misleading to present Heidegger's philosophy as a set of clearly understandable results. His metaphors must remain, rather than be translated into a usual philosophical terminology that he rejected.

Heidegger preferred to live in his retreat in the Black Forest and in his last years had few followers in his effort to think Being. He died in the town where he was born on May 26, 1976.

MAJOR WORKS

"Das Realitätsproblem in der modernen Philosophie," *Philosophisches Jahrbuch der Görresgesellschaft*, 25 (1912); *Die Lehre vom Urteil im Psychologismus: Ein kritisch-positiver Beitrag zur Logik* (1914); *Die Kategorien- und Bedeutungslehre des Duns Scotus* (1916); "Der Zeitbegriff in der Geschichtswissenschaft," *Zeitschrift für Philosophie und philosophische Kritik*, 161 (1916); *Sein und Zeit: Erste Hälfte*, first as a contribution to the *Jahrbuch für Philosophie und phänomenologische Forschung*, 8 (1927), then as a separate book (1927; 11th ed., 1967; *Being and Time*, 1962); *Kant und das Problem der Metaphysik* (1929; Eng. trans. by James Churchill, *Kant and the Problem of Metaphysics*, 1962); *Vom Wesen des Grundes* (1929; *The Essence of Reasons*, 1969); *Was ist Metaphysik?* (1929; 10th ed., 1969; "What Is Metaphysics?" in the selective volume *Existence and Being*, ed. by W. Brock, 1949); *Die Selbstbehauptung der deutschen Universität* (1933); *Hölderlin und das Wesen der Dichtung* (lecture of 1936, printed 1937; Eng. trans. in *Existence and Being*); *Platons Lehre von der Wahrheit*, first as a contribution to *Geistige Überlieferung* (1942), then in book form (1947), Eng. trans. in William Barrett and H.D. Aiken (eds.), *Philosophy in the Twentieth Century*, vol. 2 (1962); *Vom Wesen der Wahrheit* (1943, from lectures given 1930–32; 4th ed., 1961; Eng. trans. in *Existence and Being*); *Erläuterungen zu Hölderlins Dichtung* (1944; 3rd ed., 1963); *Brief über den "Humanismus,"* first with *Platons Lehre . . .* (1947), then separately (1949), Eng. trans. in *Philosophy in the Twentieth Century; Holzwege* (1950; 4th ed., 1963); *Einführung in die Metaphysik* (1953; 3rd ed., 1967; *An Introduction to Metaphysics*, 1959); *Der Feldweg* (1953; 4th ed.,

1969); *Vorräge und Aufsätze* (1954; 3rd ed., 1967); *Was heisst Denken?* (1954; *What is Called Thinking?*, 1968); *Aus der Erfahrung des Denkens* (1954); *Was ist das—die Philosophie?* (1956; *What Is Philosophy?*, 1958); *Zur Seinsfrage* (1956; *The Question of Being*, 1958); *Der Satz vom Grund* (1957); *Identität und Differenz* (1957; *Essays in Metaphysics: Identity and Difference*; 1960); *Unterwegs zur Sprache* (1959; *On the Way to Language*, 1971); *Der Ursprung des Kunstwerkes*, (1960; Eng. trans. in A. Hofstadter and R. Kuhns [eds.], *Philosophies of Art and Beauty*, 1964); *Nietzsche*, 2 vol. (1961); *Die Frage nach dem Ding: Zu Kants Lehre von den transzendentalen Grundsätzen* (1962; *What Is a Thing?*, 1967); *Kants These über das Sein* (1962).

BIBLIOGRAPHY. H. ALBERT, *Traktat über kritische Vernunft* (1968), a very competent critique of Heidegger's conception of cognition as revelation and of the general tradition to which it belongs; A. DE WAELHENS, *La Philosophie de Martin Heidegger*, 5th ed. (1967); H. FEICK, *Index zu Heideggers Sein und Zeit* (1961), a useful collection of definitions and a survey of occurrences of key terms, not severely restricted to *Sein und Zeit;* M. GRENE, "Heidegger, Martin," in P. EDWARDS (ed.), *Encyclopedia of Philosophy*, vol. 3 (1967), a courageous attempt to furnish an understandable translation and a concentrated survey of Heideggerian conceptions (perhaps too far removed, however, from the German tradition); H. LUBBE, "Bibliographie der Heidegger-Litteratur 1917–1955," *Zeitschrift für Philosophische Forschung*, vol. 11 (1957), excellent; G. LUKACS, "Heidegger redivivus," in *Sinn und Form*, 1:37–62 (1949), an influential philosophical and Marxist evaluation of the work and influence of Heidegger by a brilliant theoretician; J. MACQUARRIE, *An Existentialist Theology: A Comparison of Heidegger and Bultmann* (1955), an interpretation of Heidegger's system as a partial rediscovery of fallenness, care, death, and guilt—of the biblical understanding of man; A. NAESS, "Heidegger," in *Four Modern Philosophers: Carnap, Wittgenstein, Heidegger, Sartre* (1968), an introduction along the lines of the present article; O. POGGELER, *Der Denkweg Martin Heideggers* (1963), perhaps the most Heideggerian initiation to Heidegger's thinking; W.J. RICHARDSON, *Heidegger: Through Phenomenology to Thought*, with preface by Heidegger (1963), a useful reference, indispensable for advanced study, though not more difficult than others within the phenomenological tradition; G.J. SEIDEL, *Martin Heidegger and the Pre-Socratics: An Introduction to His Thought* (1964), a clear account of an obscure theme with penetrating critical conclusions and bibliography.

(A.D.N.)

Heilungkiang

Heilungkiang (Hei-long-jiang in Pin-yin romanization), the northernmost province of China's Northeast Region (formerly known as Manchuria), is bounded on the north and east by the Union of Soviet Socialist Republics along the Hei-lung Chiang (Hei-lung River)—known in the Soviet Union as the Amur River—and Ussuri River (Wu-su-li Chiang), on the west by the Inner Mongolian Autonomous Region of China, and on the south by the Chinese province of Kirin. The name of the province is derived from Hei-lung Chiang (Black Dragon River). The province has an area of about 272,300 square miles (705,300 square kilometres), including the administrative area of Hu-lun-pei-erh-meng (Hulunbuir League), which was transferred to it from the Inner Mongolian Autonomous Region in 1969. Late in the 1950s the population was over 14,000,000; in 1970 it was estimated at 25,000,000. The capital is Harbin.

Heilungkiang occupies about three-fifths of the area of the three Northeast provinces and has about one-third of this region's population. All but about 8 percent of its population are Chinese; the rest are Manchus, Mongolians, Koreans, Hui (Chinese Muslims), and other small groups. The long international border with the Soviet Union has been the scene of Sino-Russian clashes from the 17th century into the 1960s.

HISTORY

Early peoples of the province

The prehistoric population of the region appears to have consisted of people who bred pigs and horses; known as Tungids, they occupied much of northeastern Asia. Stone Age fishermen, the Sibirids and Ainoids, lived along the rivers and coast. Northern Manchuria remained until the 19th century an undeveloped steppe and forest region occupied by a few primitive nomadic tribes. In 1650 the Russians built a fort on the Amur River, which the Chinese regarded as their territory. Consequently, they destroyed the fort in 1685. In 1689 the Treaty of Nerchinsk was signed between Russia and China, recognizing Chinese dominion over both banks of the Amur, extending to its mouth and Sakhalin Island. Russian expansion was thus checked for almost two centuries.

In 1858 Russia annexed the region north of the Amur River to its mouth and two years later the region east of the Ussuri River to the Sea of Japan, including the important seaport of Vladivostok (Hai-sheng-wei) and the Ussuri–Amur river port of Khabarovsk (Pai-li). The Russians occupied Heilungkiang from 1900 to 1905 and maintained their domination—despite their defeat by Japan in the Far East in 1904–05—because of Russian control of the strategic Chinese Eastern Railway, running through the region from west to east. After the 1917 Russian Revolution, the Bolsheviks renounced special privileges in northern Manchuria as a friendly gesture toward China. The province remained under Chinese control until Japan invaded Manchuria on September 18, 1931. Heilungkiang became a part of the Japanese puppet state of Manchoukuo from 1932 to 1945. On August 15, 1945, just before Japan's unconditional surrender, Soviet troops entered Manchuria, but they evacuated it later to make way for the Chinese People's Liberation Army. After the Sino-Soviet rift in 1960, border clashes occurred repeatedly along the international border. (For a detailed description of the provincial capital, see HARBIN.)

THE NATURAL ENVIRONMENT

Physical features. The province of Heilungkiang occupies more than half of the huge Manchurian Plain, surrounded on three sides by old mountain ranges of medium elevation. Its central part is the Sungari–Nen river plain, delimited by the Greater Khingan Range (Ta-hsing-an-ling Shan-mo) on the west, the Lesser Khingan Range (Hsian-hsing-an-ling Shan-mo) on the north, and the Chang-kuang-ts'ai Ling (Chang-Kuang-t'sai Mountains) and the Lao-yeh Ling (Lao-yeh Mountains) on the east. The general elevation of the southern part of the Greater Khingan Range is around 3,000 feet above sea level; the elevation gradually drops toward the northeast to about 1,600 feet at its junction with the Lesser Khingan Range.

Mountain ranges

The Greater Khingan Range is composed mainly of igneous rocks resistant to erosion and weathering. Glaciers toward the end of the Tertiary Period (10,000,000 to 2,500,000 years ago) dug U-shaped valleys through which the tributaries of the Amur run. The structure of the Lesser Khingan Range is more complex. Its northern part is composed of granite, volcanic basalt, and other metamorphic rocks. The average elevation is about 2,300 feet; the granite peaks near I-ch'un rise to about 3,770 feet. The western slope facing the Nen Chiang (Nen River) is gentle, while the eastern slope is steep. The southern end of the Lesser Khingan is composed of arch-like, folded, stratified rock. The highest peaks may reach 3,280 feet or more, but the hills are generally lower. The valleys of the foreland are often broad and smooth, dotted with swamps. The rolling Sungari–Nen Chiang plain, at an elevation of 490 to 600 feet, has many bogs and swamps. In contrast, sand dunes occur in the drier western part of the plain.

The Hei-lung Chiang (Amur) is the longest stream in the province. Its upper and middle sections serve as the international boundary for a distance of 1,180 miles. Its average frozen period is 178 days at Hei-ho and 183 days at Mo-ho. Its chief tributary, the 1,159-mile-long Sungari River, is the main waterway of the province, however. It has two sources: the second Sungari River, originating in Kirin Province, and the Nen Chiang, originating on the border of Inner Mongolia. The two streams join at San-ch'a Ho (Three-Prong River). The Sungari drainage system covers an area of 202,154 square miles, most of which lies within the province.

The Ussuri River forms the Sino-Soviet boundary on

the east. It flows along a longitudinal valley between mountains. It is a broad, slow-moving river and has a tributary linking it with Lake Khanka (Hsing-k'ai Hu), the largest freshwater lake in East Asia (1,691 square miles). Only a quarter of the lake, which is on the Sino-Soviet border, belongs to China.

Climate and soils. The province has severe winters, lasting five to eight months. Summer is short but coincides with the rainy season, making it possible to raise temperate crops in most areas. There are considerable regional differences in climate. The northwest has a cold, wet, temperate climate with very cold winters; summer thaw is only superficial. Hu-ma, at the confluence of the Amur River and its tributary, the Hu-ma-erh Ho, has a mean temperature of −17.7° F (−27.6° C) in January. The July mean temperature is 74.5° F (23.6° C). There are only four months with mean temperatures over 50° F (10° C).

A temperate, wet climate prevails in the eastern section, in the drainage basin of the Ussuri River and the lower Sungari River. In the central core of the province the climate is temperate, with a deficiency of water and very severe winters. Nen Chiang, in the northern Tung-pei plain, 728 feet above sea level, has a mean temperature of −15.9° F (−26.6° C) in January, and 70° F (21.1° C) in July. The mean annual precipitation is 20 inches, most of which falls during the active growing season from June to September. Spring wheat and soybeans are grown in the plain, with its fertile black soil, put under cultivation after 1950 as a part of the opening up of the "great northern wilderness" (*pei-ta-huan*).

The southern part of the province is also very cold in winter but enjoys a warmer summer and a longer growing period. Harbin has mean temperatures of −2° F (−19° C) in January and 72° F (22° C) in July. Its mean annual precipitation is 21.6 inches. The natural vegetation is steppe grass.

Soil patterns

The soils in the province are complex. In both the Greater and Lesser Khingan mountains soils differ with altitude. Black earths are prevalent in the foothills and mountain gray forest soils higher up. Still higher, the cold, wet soils are podsilized; *i.e.*, the soluble salts (carbonates and iron and aluminum oxides) along with soluble organic matter are leached out of the topsoil and deposited in an underlying subsoil. Such soils are of low fertility, and their cultivation causes erosion. The famous chernozem, or black soil, that covers one-fourth of the province is found in the Sungari–Nen river plain (the northern section of the Great Manchurian Plain). Its eastern part has the best soils, two feet deep or more, with good texture and high humus content, yielding crops for years without fertilization. The chernozem lands form the main agricultural region of the province, growing wheat, soybeans, kaoliang, and millet. The soil is suitable for sugar beets and flax.

Vegetable and animal life. The original vegetation of the province was forest–prairie, but it has been largely destroyed by cultivation; the remaining trees are predominantly poplars. There are many species of herbaceous plants, pasture grasses, and sorghums. The central part of the plain was originally prairie–steppe; the western part of the plain is a drier steppe.

The province has two distinct fauna districts. The first is the Manchurian Plain, which constitutes the larger part of the province. It has a predominance of temperate mixed forest fauna, with a significant admixture of elements of the Eurasian taiga. Among the district's representative animals are the Manchurian hare, the eastern field vole, the rat hamster, the Far Eastern finches, the Buteo hawk, the needle-footed owl, and some species of flycatchers. Insects include the duckling beetle, the ground beetle, and the bumblebee. The region's fauna yield valuable fur and pelts, including the sable, panther, fox, chipmunk, Manchurian hare, and light-coloured polecat.

The second district, the Greater Khingan taiga, has a fauna more akin to that of the boreal forests of Europe and Siberia. The more common wildlife includes the brown bear, squirrels, chipmunks, some forest voles, the Kolinsky, or Asiatic, mink, the wood hen, the crossbill, and the Siberian frog. Among the insects may be mentioned long-horned beetles, the ground beetle, and the Siberian silkworm. During the long, cold winter the birds migrate to warmer regions as far south as the Malay Peninsula.

THE PEOPLE AND THE ECONOMY

Population. The national census of June 30, 1953, recorded 11,897,309 inhabitants in the province. The registered population at the end of 1957 was 14,860,000. The government did not release detailed demographic data in the 1950s and 1960s. After 1957 even the annual population registration statistics of the provinces were not published. The government's published estimate of 1967–68 was 21,000,000, and a broadcast by Radio Harbin in 1970 cited a figure of 25,000,000, which included the population of Hu-lun-pei-erh-meng (Hulunbuir League). If the latter figures are approximately correct, the population more than doubled in 17 years' time. The rapid increase may reflect the success of the national program of resettling peasants from North China. It also reflects the high rate of natural increase throughout China after World War II, resulting largely from improvements in sanitation, medical care, and nutrition.

Industrial development has drawn a great number of peasants from the countryside to the cities, especially to do unskilled construction work. In 1957 the total number of workers in the province was 1,200,000, an increase of 149 percent since 1952. The increase of rural population during the same period was only 12 percent.

The density of population was 83 per square mile in 1957; by 1970 it had risen to 92 per square mile. Population density varies considerably within the province, which is larger than Sweden. Rural districts in the vicinity of Harbin have densities ranging from 750 to over 1,000, while the forested Greater Khingan mountains in the northernmost part of the province have one or two persons per square mile. In a great part of the Sungari Plain, population densities range from 250 to more than 1,000 per square mile. In the Nen Chiang plain, a region of recent settlement, as well as in the Lower Sungari Plain, the rural population around the large cities has reached 250–500 per square mile, while farther away it is less than 130. The other major areas of the province are sparsely populated.

Ethnic groups

Although the population is predominantly Chinese, the remaining 8 percent contains some significant ethnic groups: Manchus, Koreans, Muslims, Mongolians, Ta-hu-erh, O-lun-chun (Oronchon), O-wen-k'o, Ho-che, and Chi-erh-chi-szu. Minor groups include Tibetans, Russians, and Yakuts. After the establishment of the Communist government, a number of autonomous *hsien* (territorial divisions) and autonomous villages were created in areas inhabited by ethnic minorities. The Manchus form the largest minority group, numbering about 630,000, distributed in the southern part of the province. They have been culturally assimilated by the Chinese majority. Most of them farm; their way of life is similar to the Chinese, and intermarriage is common, especially among the former nobility and the educated.

Korean immigration started during the reign of Hsien-feng (1851–62). After the Japanese annexation of their country (1910–45), a large number of Koreans emigrated to Heilungkiang and Kirin provinces, where they have converted large areas of swampy wasteland into rice paddies. They number about 232,000, living mostly in southeastern Heilungkiang, where many autonomous Korean villages have been established. Next numerically are the Muslims (42,000), who live and work mostly in the bigger cities as merchants, handicraftsmen, and proprietors of beef and mutton restaurants. Those in An-ta and Chao-tung raise goats and dairy cattle. There are about 38,000 Mongolians living in the drier western part of the province, where they engage in farming and animal husbandry. A third of them live in Tu-erh-po-t'e Mongol Autonomous Hsien.

The Ta-hu-erh number about 23,000 (out of 40,000 in China), living mostly in the upper Nen Valley, on the

eastern foreland of the Greater Khingan Range. They are believed to have come from the north side of the Amur River during the 15th and 16th centuries. Hunters originally, they became the earliest farmers of Heilungkiang. Probably the O-lun-ch'un, now numbering only about 1,300, also came from north of the Amur River, later to settle down in he Khingan ranges as farmers and hunters. They had domesticated the deer and were once known as the "deer riders." The O-lun-ch'un were among the earliest inhabitants of the Upper and Middle Amur. The O-wen-k'o tribesmen moved into the province in the 1st century AD. They are believed to be descendants of the Su-shen tribes (Tungusic) of the Chou dynasty. They now live in the Amur River Valley, near Ai-hui and Hei-ho, numbering about 400. Originally hunters, they have learned to farm since 1949.

Russians entered the province at the end of the 19th century and during the early 20th. A great number of emigrés arrived after the Bolshevik Revolution. Those who became Chinese citizens have remained, many of them women who married Chinese and live mostly in Harbin and in the towns along the Amur River.

Urbanization and industrialization. In 1957, city dwellers formed 36 percent of the total population; the urban population in the province had more than doubled since 1949. New industrial centres were established, and older ones were greatly expanded.

The growth of the city of Chia-mu-ssu

The city of Chia-mu-ssu, on the right bank of the Sungari River, was a town of less than 10,000 inhabitants in 1925. During the Japanese occupation (1933–45) it was built up as a military and air base. Four strategic railways were completed linking the city with T'u-men, on the North Korean border; Sui-hua, to the west; Ho-kang, to the north; and Shuang-ya-shan, to the east. By 1941 Chia-mu-ssu had 113,411 people.

Rapid industrial development began under the Chinese Communists. The city began to produce threshing machines, pressing machines, grain sowers, combines, mining machinery, and electrical and telecommunication equipment. The Chia-mu-ssu paper mill is one of the largest in China, producing both for domestic needs and for export. There is also a food-processing industry: sugar refining, meat processing, edible oils and tallow, brewing, and confectioneries. By 1962 Chia-mu-ssu had over 300,000 inhabitants.

Another burgeoning industrial city is Shuang-ya-shan (Twin-Ducks Hill), about 43 miles east of Chia-mu-ssu. Its development began after World War II and was based on the Shuang-ya-shan coalfield, which extends over 117 square miles, with 45 seams of good coking coal. The city also has a number of large plants engaged in metal processing, food processing, and the production of lumber and construction materials. It grew from 26,000 inhabitants in 1949 to almost 110,000 by 1958.

Ch'i-ch'i-ha-erh (Tsitsihar), the second-largest city and former capital of the province, also grew phenomenally in the 1950s. Over a dozen large machinery-manufacturing plants were built, making it a key centre of the industrialization program. It also has a large food-products industry. The population was 704,000 in 1958.

Harbin

Harbin, the largest city and capital of the province, with a population of 2,000,000 (1970 estimate), only grew up in 1898 as a construction base for the Chinese Eastern Railway across northern Manchuria. It soon became the major transportation hub and communications centre of northern Manchuria, with direct rail links to the Russian railroad network and to the Sea of Japan; through the South Manchurian Railway, it is linked with the Chinese and Korean rail networks and the Pacific. In 1932 the city had a population of 380,000. There were numerous handicraft industries and small oil-pressing and flour mills. By the 1950s the Harbin area had become one of China's primary industrial-development centres, with an emphasis on heavy industry. It produced a variety of industrial machines, machine tools, and agricultural machinery. Chemical and fertilizer industries were added. The city is also a food-processing centre, as well as a producer of textiles, lumber, and construction materials.

Harbin is also an important educational centre, especially in engineering and applied science. The Harbin Polytechnical University was founded in 1920 to train technical personnel for the Chinese Eastern Railway. By 1960 it offered 40 specialized programs in ten departments of engineering and technology, as well as a graduate school.

Transportation and administration. Heilungkiang had about 2,500 miles of railroads in 1957, or 12.5 percent of the national total. Little was added in the 1960s. The five trunk lines are the Harbin–Dairen (Lü-ta), Harbin–Manchou-li, Harbin–Chia-mu-ssu, Harbin–Sui-fen-ho, and Ch'i-ch'i-ha-erh–Ssu-p'ing lines. There are also ten secondary lines.

Inland waterways are not important, handling less than 2 percent of the freight volume in 1957. About 47 percent of the freight was carried by railroads, 48 percent by animal- and man-drawn carts, and less than 3 percent by motor trucks.

Administrative units of the province

Heilungkiang Province is one of 29 primary administrative divisions of the People's Republic of China. The province is divided into 6 special districts (*chuan-ch'ü*) and 8 municipalities under the direct control of the province (*shih*). The latter category includes Harbin, Ch'i-ch'i-ha-erh, Chia-mu-ssu, Ho-kang, Shuang-ya-shan, Mu-tan-chiang, Chi-hsi, and I-ch'un. The subprovince-level units are further subdivided into 64 counties (*hsien*), one autonomous county (*tzu-chih-hsien*), and one county level municipality (*shih*). The *hsien* has been the basic territorial division of China since ancient times.

Under the *hsien* is the *hsiang* (rural district, or civil township). In the late 1950s the *hsiang* was supplanted by the people's communes. During the Great Proletarian Cultural Revolution of the late 1960s many local administrative organizations were replaced with revolutionary committees. The Heilungkiang Communist Party organization was in disarray until it was reconstituted on August 26, 1971.

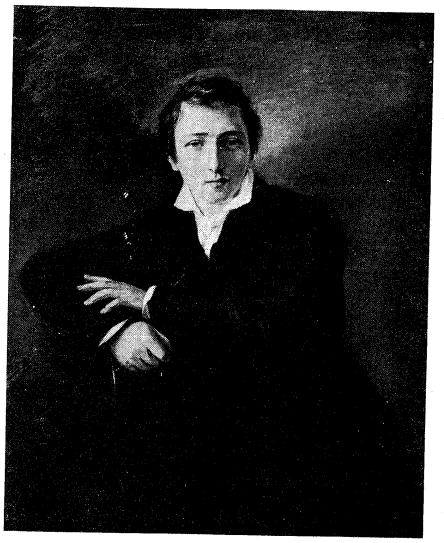
BIBLIOGRAPHY. The most comprehensive contemporary regional monographs of the province are the two volumes edited by SUN CHING-CHIH: *The Economic Geography of the Amur and Ussuri Regions of Heilungkiang Province* (Eng. trans. 1957); and *Economic Geography of Northeast China* (Eng. trans. 1962). In the 1960s the Chinese People's Republic stopped publishing basic economic statistics, but data up to 1958 and more recent estimates may be found in THEODORE SHABAD, *China's Changing Map*, 2nd ed. (1972). The best atlas is the CIA publication: *Communist China: Administrative Atlas* (1969).

(F.Hu.)

Heine, Heinrich

Heinrich Heine's international fame rests to a large extent on his *Buch der Lieder* (1827; *The Book of Songs*), the poems of which have been carried, in Heine's own phrase, "on wings of song" throughout the world by the more than 2,000 lieder settings of Robert Schumann, Franz Schubert, and many other composers. But he was more than the most famous love poet of the 19th century and, perhaps, of all European literature since Petrarch. He wrote many kinds of poetry and prose; he was a satirist and wit of exceptional skill; and his efforts as a writer of political and social commitment and of opposition to the oppression of his time today form the base for a new evaluation of his significance.

Early life. Heine was born of Jewish parents in Düsseldorf on December 13, 1797. His father was a handsome and kindly but somewhat ineffectual merchant; his mother was fairly well educated for her time and sharply ambitious for her son. Much of Heine's early life, however, was influenced by the financial power of his uncle Salomon Heine, a millionaire Hamburg banker who endeavoured to trade generosity for obedience and with whom Heine remained on an awkward and shifting footing for many years. After he had been educated in the Düsseldorf Lyceum, an unsuccessful attempt was undertaken to make a businessman of him, first in banking, then in retailing. Eventually, his uncle was prevailed upon to finance a university education, and Heine attended the universities of Bonn, Göttingen, Berlin, and Göttingen



Heine, oil painting by Moritz Oppenheim, 1831. In the Hamburger Kunsthalle.
By courtesy of the Hamburger Kunsthalle

again, where he finally took a degree in law with absolutely minimal achievement in 1825. In that same year, in order to open up the possibility of a civil service career, closed to Jews at that time, he converted to Protestantism with little enthusiasm and some resentment. He never practiced law, however, nor held a position in government service; and his student years had been primarily devoted not to the studies for which his uncle had been paying but to poetry, literature, and history.

Heine's pre-university years are rather obscure, but during this period he apparently conceived an infatuation for one, and possibly both, of his uncle's daughters, neither of whom had the slightest notion of mortgaging her future to a dreamy and incompetent cousin. Out of the emotional desolation of this experience arose, over a period of years, the poems eventually collected in *The Book of Songs*. The sound of Romantic poetry was firmly lodged in Heine's ear; but the Romantic faith, the hope for a poeticization of life and the world to overcome the revolution, alienation, and anxiety of modern times, was not in his heart. Thus, he became the major representative of the post-Romantic crisis in Germany, a time overshadowed by the stunning achievements of Goethe, Schiller, and the Romantics but increasingly aware of the inadequacy of this tradition to the new stresses and upheavals of the modern age. The most consistent characteristic of Heine's thought and writing throughout his career is a taut and ambiguous tension between "poesy," as he called the artistic sensibility, and reality. His love poems, though they employ Romantic materials, are at the same time suspicious of them and of the feelings they purportedly represent. They are bittersweet and self-ironic, displaying at the same time poetic virtuosity and a skepticism about poetic truth; their music is now liquid, now discordant, and the collection as a whole moves in the direction of desentimentalization and a new integration of the poet's self-regard in the awareness of his artistic genius.

Early poetry

The steady growth of Heine's fame in the 1820s was accelerated by a series of experiments in prose. In the fall of 1824, in order to relax from his hated studies in Göttingen, he took a walking tour through the Harz Mountains and wrote a little book about it, fictionalizing his modest adventure and weaving into it elements both of his poetic imagination and of sharp-eyed social comment. *Die Harzreise* ("*The Harz Journey*") became the first piece of what were to be four volumes of *Reisebilder* (1826–31; *Pictures of Travel*); the whimsical amalgam of its fact and fiction, autobiography, social criticism, and literary polemic was widely imitated by other writers in subsequent years. Some of the pieces were drawn from a journey to England Heine made in 1827 and a trip to Italy in 1828, but the finest of them, "Ideen. Das Buch Le Grand" (1827; "Ideas. The Book Le Grand"), is a journey into the self, a wittily woven fabric of childhood memory, enthusiasm for Napoleon, ironic sorrow at unhappy love, and political allusion.

Later life and works. When the July Revolution of 1830 occurred in France, Heine did not, like many of his liberal and radical contemporaries, race to Paris at once but continued his more or less serious efforts to find some sort of paying position in Germany. In the spring of 1831 he finally went to Paris, where he was to live for the rest of his life. He had originally been attracted by the new Saint-Simonian religion (a socialistic ideology according to which the state should own all property and the worker should be entitled to share according to the quality and amount of his work); it inspired in him hopes for a modern doctrine that would overcome the repressive ideologies of the past and put what he variously called spiritualism and sensualism, or Nazarenism (adherence to Judeo-Christian ideals) and Hellenism (adherence to ancient Greek ideals), into a new balance for a happier human society. His critical concern with political and social matters deepened as he watched the development of limited democracy and a capitalist order in the France of the citizen-king, Louis-Philippe. He wrote a series of penetrating newspaper articles about the new order in France, which he collected in book form as *Französische Zustände* (*French Affairs*, 1906) and followed with two studies of German culture, *Die Romantische Schule* (1833–35; *The Romantic School*) and "Zur Geschichte der Religion und Philosophie in Deutschland" (1834–35; "On the History of Religion and Philosophy in Germany"), in which he mounted a criticism of Germany's present and recent past and argued the long-range revolutionary potential of the German heritage of the Reformation, the Enlightenment, and modern critical philosophy. The books were conceived with a French audience in mind and were originally published in French. In 1840–43 he wrote another series of newspaper articles about French life, culture, and politics, which he re-edited and published as *Lutezia*, the ancient Roman name for Paris, in 1854.

Move to Paris

During these years, then, Heine's attention turned from "poesy" to writing of contemporary relevance. His second volume of poems, *Neue Gedichte* (1844; *New Poems*), illustrates the change. The first group, "Neuer Frühling" ("New Spring," written mostly in 1830/31), is a more mannered reprise of the love poems of *Buch der Lieder*, and the volume also contains some ballad poetry, a genre in which Heine worked all his life. But the second group, "Verschiedene" ("Varia"), is made up of short cycles of sour poems about inconstant relationships with the blithe girls of Paris; the disillusioning tone of the poems was widely misunderstood and held against him. Another section is called "Zeitgedichte" ("Contemporary Poems"), a group of harsh verses of political satire. Several of these were written for Karl Marx's newspaper *Vorwärts* ("Forward"). Heine had become acquainted with the young Marx at the end of 1843, and it was at this time that he produced, after a visit to his family in Germany, a long verse satire, *Deutschland. Ein Wintermärchen* (1844; *Germany, a Winter's Tale*), a stinging attack on reactionary conditions in Germany. Though Heine remained on good, if not intimate, terms with Marx in later years, he never was much taken with Communism, which did not fit his ideal of a revolution of joy and sensuality. About the time that he met Marx, he also wrote another long poem, *Atta Troll. Ein Sommernachtstraum* (*Atta Troll, a Midsummer Night's Dream*, 1843–45), a comic spoof of radical pomposity and the clumsiness of contemporary political verse.

Heine's early years in Paris were the happiest of his life. From an outcast in the society of his own rich uncle, he was transformed into a leading literary personality, and he became acquainted with many of the famous and prominent people of his time. In 1834 he found in an uneducated shopgirl, Crescence Eugénie Mirat, whom for some reason he called "Mathilde," a loyal if obstreperous mistress. He married her in 1841. But troubles were soon hard upon him. His critical and satirical writings brought him into grave difficulties with the German censorship,

and, at the end of 1835, the Federal German Diet tried to enforce a nationwide ban on all his works. He was surrounded by police spies, and his voluntary exile became an imposed one. In 1840 Heine wrote a witty but ill-advised book on the late Ludwig Börne (1786–1837), the leader of the German radicals in Paris, in which Heine attempted to defend his own more subtle stand against what he thought of as the shallowness of political activism; but the arrogance and ruthlessness of the book alienated the public in all camps.

Though never destitute, Heine was always out of money; and when his uncle died in 1844, all but disinheriting him, he began, under the eyes of all Europe, a violent struggle for the inheritance, which was settled with the grant of a right of censorship over his writings to his uncle's family; in this way, apparently, the bulk of Heine's memoirs was lost to posterity. The information, revealed after the French Revolution of 1848, that he had been receiving a secret pension from the French government, further embarrassed him.

Last poetry

The worst of his sufferings, however, were caused by his deteriorating health. An apparently venereal disease began to attack one part of his nervous system after another, and from the spring of 1848 he was confined to his "mattress-grave," paralyzed, tortured with spinal cramps, and partially blind. Heine returned again to "poesy." With sardonic evasiveness he abjured his faith in the divinity of man and acknowledged a personal God in order to squabble with him about the unjust governance on the world. His third volume of poems, *Romanzero* (1851), is full of heartrending laments and bleak glosses on the human condition; many of these poems are now regarded as among his finest. A final collection, *Gedichte 1853 und 1854* (*Poems 1853 and 1854*), is of the same order. After nearly eight years of torment, Heine died on February 17, 1856, and was buried in the Montmartre Cemetery.

Heine's power to annoy was as great as his power to charm and move, and rarely has a great poet been so controversial in his own country. His aggressive satires, radical postures, and insouciance about his methods made him appear to many as an unpatriotic and subversive scoundrel, and the growth of anti-Semitism contributed to the case against him. Efforts in the late 19th and early 20th centuries to erect monuments to him in various German cities touched off riots and shook governments. In view of the popularity of many of his songs, the Nazis were obliged to include them in anthologies but marked them "author unknown." For many decades his literary reputation was stronger abroad, especially in France, England, and America, where his wit and ambivalence were better appreciated, than at home. Today the evaluation of Heine's political role and its relationship to Marxism supplies a bone of contention between East and West. Deplorable as much of this history of his reputation has been, it is testimony to the enduring impact of a genuinely European poet and writer.

MAJOR WORKS

Gedichte (1821), first collection of poems; *Tragödien nebst einem lyrischen Intermezzo* (1823), the tragedies were *Almansor* and *William Ratcliff; Reisebilder*, vol. 1: "Die Heimkehr," "Die Harzreise," "Nordsee I," first of the "North Sea" cycles (1826); vol. 2: "Nordsee II," "Nordsee III," "Ideen. Das Buch Le Grand," "Briefe aus Berlin" (1827); vol. 3 included "Reise von München nach Genoa" and "Die Bäder von Lucca" (1829); vol. 4 included "Die Stadt Lucca" and "Englische Fragmente" (1831); *Buch der Lieder* (1827); *Französische Zustände* (1832), a collection of political essays; *Zur Geschichte der neueren schönen Literatur in Deutschland* (1833), on contemporary German literature; reprinted and expanded in *Die Romantische Schule* (1835); *Der Salon*, 4 vol. (1834–40), vol. 1 included a description of French painting, the "Salon" of 1831, a collection of poems, and "Aus den Memoiren des Herren von Schnabelewopski"; vol. 2 (1835) consisted largely of a synopsis of the history of religion and philosophy in Germany, "Zur Geschichte der Religion und Philosophie in Deutschland"; vol. 3 (1837) contained two prose works, "Elementargeister" and "Florentinische Nächte"; vol. 4 (1840) included "Vertraute Briefe über die französische Bühne" and the opening chapters of the novel *Der Rabbi von Bacherach; Ludwig Börne, eine Denkschrift* (1840); *Deutschland. Ein Wintermärchen* (1844), satirical verse; *Neue Gedichte* (1844), second collection of shorter poems; *Atta Troll. Ein Sommernachtstraum* (1847), mock-epic poem; *Der Doktor Faust. Ein Tanzpoem* (1851); *Romanzero* (1851), included the "Historien," "Lamentationen," and "Hebräische Melodien"; *Gedichte 1853 und 1854* (1854); *Lutezia* (1854), political essays written 1840–43; *Vermischte Schriften* (1854), included "Geständnisse" and "Götter im Exil."

TRANSLATIONS: *Heinrich Heine: The Poems*, trans. by Louis Untermeyer (1937), is the most comprehensive English translation of Heine's poetry; various of Heine's works have been translated into English many times and into some 50 other languages. Other notable English translations are: *The Prose Writings of Heinrich Heine*, ed. by Havelock Ellis (1887); *The Works of Heinrich Heine*, trans. by C.G. Leland *et al.*, 12 vol. (1892–1905); *The Poems of Heine*, trans. by E.A. Bowring (1859); and *Heinrich Heine: The North Sea*, trans. by Vernon Watkins (1951).

BIBLIOGRAPHY. GOTTFRIED WILHELM and EBERHARD GALLEY, *Heine-Bibliographie (1817–1953)*, 2 vol. (1960); SIEGFRIED SEIFERT, *Heine-Bibliographie 1954–1964* (1968).

Manuscripts, papers, and memorabilia: The major collections of materials are in the Heine Archive of the Landes- und Stadtbibliothek, Düsseldorf; in the Bibliothèque Nationale in Paris; and in the Nationale Forschungs- und Gedenkstätten der klassischen deutschen Literatur in Weimar. Manuscript materials are also at Harvard and Yale. Since 1962 the Düsseldorf archive has published a *Heine-Jahrbuch* with a running annual bibliography.

Editions: ERNST ELSTER (ed.), *Sämtliche Werke*, 7 vol. (1887–90); OSKAR WALZEL *et al.* (eds.), *Sämtliche Werke*, 10 vol. (1910–15; index volume, 1920); HANS KAUFMANN (ed.), *Werke und Briefe*, 10 vol. (1961–64); FRIEDRICH HIRTH (ed.), *Briefe*, 6 vol. (1950–51), letters. New critical editions are being prepared in Düsseldorf and Weimar.

Biographies: ADOLF STRODTMANN, *Heinrich Heines Leben und Werke*, 2nd ed., 2 vol. (1873–74), the first and, in some ways, still the best comprehensive biography; LUDWIG MARCUSE, *Heinrich Heine: Ein Leben zwischen Gestern und Morgen* (1932; Eng. trans., *Heine: A Life Between Love and Hate*, 1933); LOUIS UNTERMEYER, *Heinrich Heine: Paradox and Poet*, vol. 1, *The Life* (1937), a companion volume to the translation of poems; E.M. BUTLER, *Heinrich Heine: A Biography* (1956); H.H. HOUBEN, *Gespräche mit Heine* (1926), a compendium of contemporaries' recollections and a valuable biographical sourcebook; FRITZ MENDE, *Heinrich Heine, Chronik seines Lebens und Werkes* (1970), a day-by-day account of all known events and activities in Heine's life.

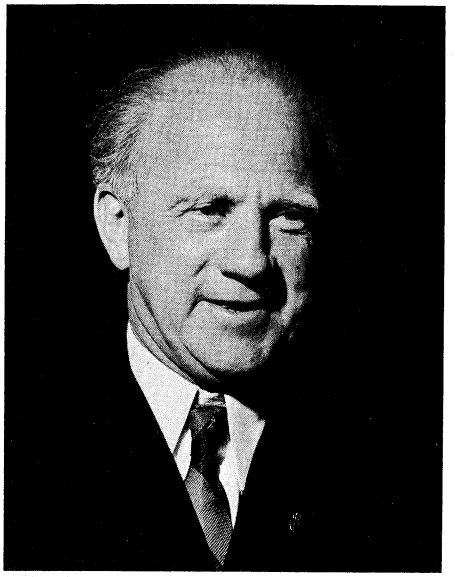
Critical studies: BARKER FAIRLEY, *Heinrich Heine: An Interpretation* (1954); WILLIAM ROSE, *Heinrich Heine: Two Studies of His Thought and Feeling* (1956); S.S. PRAWER, *Heine: The Tragic Satirist* (1961); LAURA HOFRICHTER, *Heinrich Heine* (1963); JEFFREY L. SAMMONS, *Heinrich Heine: The Elusive Poet* (1969).

(J.L.Sa.)

Heisenberg, Werner Karl

Werner Heisenberg, physicist, philosopher, and public figure, helped to establish the modern science of quantum mechanics, out of which came the famous "indeterminacy principle." He contributed significant refinements to a conceptual understanding of the atomic nucleus, ferromagnetism, cosmic rays, and elementary particles. In his philosophical writings, he insisted that the scientist is as much an "actor" as a "spectator" in scientific inquiry. He conceived of a Platonic central order, consisting of a set of universal symmetries that are exhibited in all natural phenomena. In his view, these symmetries constitute the rationale for a mathematical equation that may be applied to all systems of particulate matter. As a public figure, he actively promoted the peaceful use of atomic power after World War II and, in 1957, led other German scientists in opposing a move to equip the West German Army with nuclear weapons.

Born December 5, 1901, in Würzburg, Germany, Heisenberg studied physics at the University of Munich, where he wrote his doctoral dissertation in 1923 on turbulence in fluid streams. Interested in Niels Bohr's account of the planetary atom, Heisenberg went to the University of Göttingen to study under Max Born and then, in the fall of 1924, to the Universitets Institut for Teoretisk Fysik in Copenhagen to study under Bohr. In June 1925, while recuperating from an attack of hay fever on



Heisenberg.
Edo Keonig—Black Star

Helgoland, an island in the North Sea, he solved a major physical problem—how to account for the stationary (discrete) energy states of an anharmonic oscillator—and in so doing laid the program for the development of quantum mechanics (the science that accounts for discrete energy states and other forms of quantized energy, as in the light of atomic spectra and in the phenomenon of stability exhibited by macroscopic pieces of matter). A few months later, he published the article "Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen" ("About the Quantum-Theoretical Reinterpretation of Kinetic and Mechanical Relationships"), in which he proposed a reinterpretation of the basic concepts of mechanics. Physical variables were to be represented by matrices, or arrays, of numbers and would deal only with "observable," or measurable, quantities. Heisenberg and other eminent physicists then used the new quantum mechanics to interpret many atomic and molecular spectra, ferromagnetic phenomena, and electromagnetic behaviour. Alternative forms of the new quantum theory were proposed in 1926 by Erwin Schrödinger and P.A.M. Dirac.

Indeterminacy principle

In 1927 Heisenberg published his indeterminacy principle, which stated the theoretical limitations imposed by quantum mechanics upon certain pairs of variables that constantly affect each other, such as position and momentum. He held that in their new designations as conjugate observables (an interrelated pair of measurable quantities), indeterminacy ruled that no quantum mechanical system could simultaneously possess an exact position and exact momentum. Indeterminacy affects all phenomena, great and small, but its significance is usually confined to the microphysical domain.

Bohr and Heisenberg elaborated a philosophy of complementarity to take into account the new physical variables, each of which would be relative to an appropriate measurement process on which it depends. The new conception of the measurement process in physics emphasizes the active role of the scientist, who, in the act of making measurements, interacts with the observed object and thus causes it to be revealed not as it is in itself but as a function of measurement. Many physicists, including Albert Einstein, Schrödinger, and Louis de Broglie, refused to accept the philosophy of complementarity.

From 1927 to 1941, Heisenberg was professor at the University of Leipzig. For the following four years, he was director of the Kaiser Wilhelm Institute for Physics in Berlin. During World War II, he worked with Otto Hahn, one of the discoverers of nuclear fission, on the development of a nuclear reactor. Although he did not publicly oppose the Nazi regime, he was hostile to its policies and acted so as to prevent Germany from developing effective nuclear weapons. After the war, he organized and became director of the Max Planck Institute for Physics and Astrophysics at Göttingen, later moving with the institute, in 1958, to Munich.

Universal nonlinear spinor equation

In the postwar period, Heisenberg began working on a universal nonlinear spinor equation (a nonlinear differential equation for complex vector-like entities representing all possible states of matter) for particulate systems that would exhibit the basic set of universal symmetries (relative to possible observer viewpoints) in nature and yet be capable of explaining the variety of elementary particles generated in high-energy collisions.

In his writings on the philosophical implications of quantum mechanics, Heisenberg vigorously opposed the Logical Positivism developed by philosophers in Vienna. In the theory of knowledge of quantum physics, he stressed the active role of the observer and replaced the absolute objects of classical science with relativized observational situations. Quantum mechanics implies that on the subatomic level the traditional idea of scientific causality needs broadening, since the behaviour of particles can be predicted only on the basis of probability; and that Newton's laws concerning the motion of bodies in space and time do not fit the basic processes within the atom. But classical physics remains valid in his view for a wide domain of macroscopic phenomena. The indeterminacy principle is of significance therefore primarily for subatomic particles.

Heisenberg married Elizabeth Schumacher (1937); they had seven children. Although known primarily as a physicist, he continued in the classical tradition of philosophy, seeking a new solution to the ancient problem of the One and the Many. Widely acknowledged as one of the seminal thinkers of the 20th century, Heisenberg was awarded the Nobel Prize for Physics in 1932 and honoured with the Max Planck Medal, the Matteucci Medal, and the Barnard College Medal of Columbia University. He died at Munich on February 1, 1976.

BIBLIOGRAPHY. P.A. HEELAN, *Quantum Mechanics and Objectivity* (1965), is the only critical study of Heisenberg's philosophy of science in English. M. JAMMER, *The Conceptual Development of Quantum Mechanics* (1966), is the most complete historical study of Heisenberg's contribution to quantum mechanics. Books by Heisenberg include *Die physikalischen Prinzipien der Quantentheorie* (1930; Eng. trans., *The Physical Principles of the Quantum Theory*, 1930), his most important, containing themes of his early papers amplified as a treatise; *Wandlungen in den Grundlagen der Naturwissenschaft* (1948; Eng. trans., *Philosophic Problems of Nuclear Science*, 1952), a collection of his early essays; *Physics and Philosophy: The Revolution in Modern Science* (1958), his Gifford lectures; and *Der Teil und das Ganze* (1969; Eng. trans., *Physics and Beyond*, 1971), an autobiographical memoir of his early life.

(P.A.H.)

Hellenistic Law

Hellenistic law is that body of law, of essentially Greek origin, that applied to Greek and Hellenized inhabitants of Egypt, Palestine, Syria, Asia Minor, and other countries of the ancient Near East during the period beginning with the death of Alexander the Great in 323 BC and the founding of the Hellenistic monarchies by his Macedonian successors, and ending with the absorption of these monarchies into the Roman Empire. One generally considers the year 30 BC to be the end of this period, when Egypt came under Roman rule; but as early as the mid-2nd century BC, Rome had conquered certain areas of Asia Minor, while Doura-Europus (on the Euphrates) remained independent until about AD 164. With the advent of Rome the influx of Roman law began. But it came slowly, and as a result, Hellenistic law continued to develop for some time. In fact, the development of Hellenistic law continued, be it with Roman influences, at least until AD 212, when Roman citizenship was conferred upon all the inhabitants of the Roman Empire.

Since classical authors give practically no information about Hellenistic law, modern knowledge is derived from such other sources as papyruses, parchments, and inscriptions on which are found some contracts, petitions,

and records of lawsuits. Any investigation into Hellenistic law is further hampered by two factors: first, no legal code nor contemporary legal treatises exist, making it necessary to reconstruct the law from single cases; second, the documents that do survive come mainly from Egypt (though there are records, for instance, from Delphi in Greece that offer some information about slavery, property law, and the status of women), and as a result, knowledge of Hellenistic law is restricted primarily to Egypt.

The character of Hellenistic law. *The origins of Hellenistic law.* In classical Greece every autonomous city-state had its own body of law. Because these legal systems had some general principles in common, some scholars have considered Greek law to be a unified body of law. Yet, because there appear to be many differences in the actual regulations of the various city-states, it can be contended that Greek law was instead a plurality of laws. The legal capacity of women, for example, was very restricted in Athens, much more than it was in Sparta or Gortyn on Crete; further, in Thessaly, women were equal to men in legal matters. There must have been considerable differences in the laws concerning other matters, because by the 4th century BC it was necessary to institute a special tribunal in Athens to deal with cases involving traders from other city-states. The existence of this tribunal implies that some "international" law must have regulated the affairs of those who travelled from one city-state to another (see also GREEK LAW).

As Alexander the Great swept through the Near East, citizens of the Greek city-states began to settle in the conquered areas. The oldest Greek legal document from Hellenistic Egypt shows something of a heterogeneous population living in Elephantine in 311 BC. The document records a marriage, and since the bride came from one city-state and the groom from another, the question arises as to the law of which city-state applied. It appears that neither one did, but rather that a "common law" was used among all Greeks who lived in Hellenistic countries. This Hellenistic law evolved from various Greek laws of the preceding period in much the same way as the common Greek language, *koinē*, evolved from the various Greek dialects spoken by the people. Although the exact origin of this common language and common law is unknown, there is speculation that it lies in the preceding period, when Greeks from various parts of the empire began to come in contact with each other.

The unity of Hellenistic law. Although there were some influences from local law, private Hellenistic law does appear, in many respects, to be universal—in the same way as, in the linguistic field, the *koinē* was largely uninfluenced by other local languages and remained the uniform Greek language of the whole Hellenistic world. Originally, scholars had thought that, at least in Egypt, a special variety of Hellenistic law had existed, but recent finds outside Egypt show clearly that institutions thought to be specific to Egypt also occurred elsewhere in the Hellenistic world: joint wills have been found to have existed in Asia Minor, not only in Egypt; marriages between siblings appear to have occurred in many parts of the Hellenistic world; and finally, the sale of land has been found to have been, in many respects, the same in Doura-Europus, near the border of Mesopotamia, as it was in Egypt.

In the area of public law there were more local variances. This was due to the fact that the first Hellenistic monarchs adopted, to a large extent, the administrative and fiscal regimes of the previous governments—Syrian, Egyptian, and so forth. Other local occurrences, such as the regular flooding of the Nile valley, served to influence legal practice as they did daily life.

Certain unique attributes of Hellenistic law. One of the most striking features of Hellenistic law is the absence of fixed legal terminology such as that found in Roman law. Although, for example, the *kyrios* usually means the guardian of a woman and *epitropos* the guardian of a minor or a ward, scholars acknowledge that either word might be substituted for the other.

Another striking feature is the flexibility of the legal system and the absence of fixed and clearly defined rules of contract. Sometimes, for example, it is difficult to distinguish a loan from a sale on credit or a sale with deferred delivery. Land lease might be combined with loan of seed corn. Manumission of slaves might be combined with the obligation to stay with someone for a certain period of time. Land lease might involve payment of a debt. Yet, at the same time, in penal law and torts (involving civil wrongs and injuries) different kinds of offenses were distinguished and subject to specific actions.

Importance given to written documents

A final striking feature is the importance given to written documents. A written document was the favourite method of proving a right or the extinction of an obligation. This feature is common to all the legal systems of the Hellenistic world, but it was apparently unknown in ancient Greece.

The relation to local law. The first Hellenistic monarchs respected the local law that applied to the indigenous inhabitants. Local law remained in force during the entire Hellenistic Age, so that when the Roman period began Hebraic law was still in Palestine, Babylonian law in Mesopotamia, and Egyptian law in Egypt. Special tribunals had made possible this survival: in Hellenistic Egypt Greek tribunals administered justice according to Greek law, Egyptian tribunals according to Egyptian law, and Jewish tribunals according to Hebraic law. Further, the king received petitions on matters of Jewish law or Egyptian law, and decisions apparently were made according to their respective laws.

In the beginning of the Hellenistic Age everyone was bound by his own law in conformity with the principle of personality; later on, this principle was abandoned so that everyone became free to choose the system of law he wished to be governed by. From that point on, the language used by the parties to a contract determined which system of law applied. The important thing here is that freedom of choice allowed the parties to avail themselves of procedures not found in their own law; for example, Jews entered into Greek loans with interest (forbidden by Jewish law) and divorced by mere consent (unknown to Jewish law).

Freedom of choice

Such freedom of choice does not mean that the law of the Jews or Egyptians or other peoples disappeared. On the contrary, the recent discoveries near the Dead Sea and the evidence from Egypt show that Jews used their own as well as Hellenistic law in the same way as they also used both the Hebrew and Greek languages. The Egyptians also appear to have preserved their language and, in some respects, their law, although both were submerged for several centuries until the Coptic period.

The principal subjects of Hellenistic law. *Law of persons.* Slaves were not very numerous in Hellenistic Egypt, perhaps because the economic condition of many a free person was such that his actual position was not very different from that of a slave. A slave was capable of owning property and of engaging in legal transactions and thus had more rights than a slave in ancient Greece.

Among free persons, women held a special place. Though their capacity was much less restricted than it had been under the law of Athens, the assistance of a guardian was still often required, even if only as a formality. In local Egyptian law there were no restrictions on the legal capacity of women.

Family law

The *oikos* or closed family, characteristic of ancient Greece, disappeared with the broadening of horizons in the Hellenistic Age, and the process of transferring the bride from her father's to her husband's *oikos* disappeared with it. Marriage was constituted simply by actual cohabitation with a view toward a common life as spouses; and divorce was effectuated by ending this cohabitation; in neither case was a legal act involved. Both spouses were free to divorce. No written document was required for constituting a valid marriage or divorce; the main purpose of such documents was to arrange property settlements. Polygamy and marriages between brothers and sisters or other close relatives occurred, though not very often.

Unlike the Roman paterfamilias (head of family), who had absolute power over his children, a Hellenistic fa-

ther had no more than a simple paternal authority. A mother also had a maternal authority over her children, especially when they were fatherless. Adoption existed, and the persons adopted were usually free persons and not foundlings, who were considered as slaves.

Law of property. In line with the character of Hellenistic law, there is no clear terminology that distinguishes such legal notions as "ownership" and "possession," "movables" and "immovables," although in practice these categories of goods existed, as did different qualities of ownership. It is true that because the king was said to own all the land, the question arises as to whether private ownership could exist. Yet private persons did freely buy and sell land, leased it and paid taxes on it, so that it appears that the king's sovereign rights to the land were meant merely to assert his right to levy taxes.

There were several categories of land, although their exact nature often is not known. "Royal land" in Egypt and Seleucia seems to have been the private property of the king, and "holy" or "temple land" that of the temples. On the other hand, so-called temple domains existed; private persons could own land situated there, but they were obliged to pay a tax to the temple in question. Another kind of land was called *klēros*, originally a fee temporarily held either by individual soldiers (in Egypt) or by entire military colonies (in Seleucia and pre-Hellenistic Egypt), granted to them in return for military and other services. Gradually such lands became private property and lost their military aspect, so that by the 1st century BC they even could be owned by women.

Inheritance. The Egyptian law of inheritance was fundamentally different from the Greek, although the differences tended to disappear in Hellenistic law. In Egyptian law, when someone died, a member of his family assumed his position within the family; the heir took over the entire property, including the debts, for which he became responsible. He could not decline the inheritance and was responsible to his prospective heirs to preserve the estate. In contrast, in Greek law, the heir inherited the assets only; he was not responsible for the debts and could decline the inheritance. According to Hellenistic law, the property tended to be divided among the heirs—the primary heirs being the natural or adopted children (male and female) of the deceased. The inability of Athenian women to make a will survived in Alexandria but disappeared in the rest of the Hellenistic world.

Obligations. One of the major problems in the history of law is the question of whether or not debtors could be held responsible for their debts or how, in the proceedings, they became liable. With regard to Hellenistic law, scholars have questioned the existence of contracts based merely on consent and, though the sources appear to show that such contracts did exist, many scholars have denied it and concluded that, for example, a debtor became liable, not by promising something but by submitting himself to punishment when he could not perform what he had promised to do; or that he became liable by adding a surety to his contract rather than simply his consent.

Liability of a debtor

Sale. The evidence on sale is limited almost exclusively to immovables and valuable goods, since the sale of movables usually occurred without a written document. Sale was usually a cash transaction, with the goods sold being delivered at the same time that payment was made. It might happen, however, that the price was paid in advance or that the goods were delivered on credit, in which case there might be a question as to when transfer of ownership took place. It appears that this caused no difficulty in Hellenistic law, although the time of actual transfer of ownership is not known.

Loan and deposit. Credit was common and was found in a variety of forms; interest could be paid in money, in kind, in the performance of services, and so forth. A maximum rate of interest on loans of money was introduced in Egypt around 250 BC, at about 24 percent per year; this was reduced in Roman times to about 12 percent per year. Loans in kind, however, required the borrower to pay 50 percent, regardless of the duration of the loan.

Sometimes the amount named in the contract was the amount to be repaid, either because interest was not required or because it had already been calculated in the amount. This type of contract, very common in Roman Egypt, is generally thought to have influenced the development of a similar Roman loan, because the regular Roman loan did not allow for interest.

Torts and criminal law. There was no distinction in Hellenistic law between public and private law, and therefore between offenses against the state and against individuals. Consequently, the actions and procedures dealing with such offenses are quite mixed. Distinctions were made, however, between premeditated offenses and unpremeditated offenses. Offenses committed at night, in a temple, by armed men, or by a band were considered as aggravated offenses and, therefore, more serious.

The Athenian democratic principle that every citizen was entitled to bring an action when a public offense had taken place disappeared in Egypt except in the case of fiscal offenses against the state. In this instance, the motivating concern was not one of democratic principle but the interest of the state.

There were many different actions for offenses. They were directed usually toward compensation for damages and at civil monetary penalties (many of these cases could be settled by agreement in order to avoid an action), but public penalties such as confiscation of the offender's property and the death penalty occurred, especially in cases of fiscal offenses against the state. In general, corporal punishment seems to have been inflicted only on slaves.

Procedure. It is a striking feature that in Egypt jurisdiction was exercised not only by courts made up of a group of judges or by single magistrates but also, to a certain degree, by influential private persons.

Several Greek, Jewish, and Egyptian courts administered justice in Egypt. It is not known how the Jewish courts were organized, but a royal official was attached to the Greek and Egyptian courts in order to prepare the cases and to carry out the sentences. The king could, in the person of this official, supervise the courts, especially the Egyptian ones.

Execution was accomplished by an executive officer, against the property as well as the person of the debtor. A debtor could be placed in detention, but it is not known if he could be sold as a slave. Hellenistic monarchs tried to abolish punishment against an individual's person for other than fiscal debts, but they did not succeed.

BIBLIOGRAPHY. R. TAUBENSCHLAG, *The Law of Greco-Roman Egypt in the Light of the Papyri, 332 B.C.–640 A.D.*, 2nd ed. (1955); and E. SEIDL, *Ptolemäische Rechtsgeschichte*, 2nd ed. (1962), are the best recent handbooks but are mainly concerned with Hellenistic law in Egypt. No general account of Hellenistic law as such exists. Many important articles are collected in the selected works of R. TAUBENSCHLAG, *Opera Minora*, 2 vol. (1959). Of a more general interest are: J.G. WINTER, *Life and Letters in the Papyri* (1933); E.G. TURNER, *Greek Papyri: An Introduction* (1968); and A. DEISSMANN, *Licht vom Osten: Das Neue Testament und die neuentdeckten Texte der hellenistisch-römischen Welt*, 4th ed. (1923; Eng. trans., *Light from the Ancient East: The New Testament Illustrated by Recently Discovered Texts of the Graeco-Roman World*, new ed. 1927, reprinted 1965).

An extensive bibliography of Hellenistic law may be found in J. MODRZEJEWSKI, "Monde Hellénistique," vol. A/8 of *Introduction bibliographique à l'histoire du droit et à l'ethnologie juridique*, ed. by J. GILISSEN (1965); more recent bibliographies are regularly published in the *Revue Historique de Droit Français et Étranger*, and in *Studia et Documenta Historiae et Iuris*.

Some important collections of Greek papyruses from Egypt are: L. MITTEIS and U. WILCKEN, *Grundzüge und Chrestomathie der Papyruskunde*, 4 vol. (1912), with a thorough commentary; A.S. HUNT and C.C. EDGAR, *Select Papyri*, vol. 1, *Private Documents*, 3rd ed. (1959), vol. 2, *Public Documents*, 2nd ed. (1956), with English translations of the documents; and M. DAVID and B.A. VAN GRONINGEN, *Papyrological Primer*, 4th ed. (1965), with explanatory and bibliographical annotations.

(P.W.P.)

Hellenistic Religions

The Hellenistic Age (*c.* 300 BC–AD 300), when taken as a whole, constitutes one of the most creative periods in the history of religions. It was a time of spiritual revolution in the Greek and Roman empires, when old cults died or were fundamentally transformed and when new religious movements came into being.

Nature and significance. The Hellenistic Age may be broadly defined as the period from the Greco-Macedonian conqueror Alexander the Great (356–323 BC) to Constantine, the first Christian Roman emperor (d. AD 337), and its religions may be confined to those that were active within the Mediterranean world. The empire of Alexander and his successors created a great world community which, whether in Macedonian, Greco-Roman, or its later Christian form, established a cultural unity that was destined to be broken only 1,000 years later with the advent of Muslim imperialism (beginning in 7th century AD). This empire was so vast as truly to stagger the imagination. Extending from the Strait of Gibraltar to the Indus River, from the forests of Germany and the steppes of Russia to the Sahara Desert and the Indian Ocean, it took in an area of 1,500,000 square miles (most of Europe, the Mediterranean, the Near and Middle East, Africa, Persia, and the borderlands of India) and had a total population of over 54,000,000.

The study of Hellenistic religions is a study of the dynamics of religious persistence and change in this vast and culturally varied area. Almost every religion in this period occurred in both its homeland and in diasporic centres—the foreign cities in which its adherents lived as minority groups. For example, Isis (Egypt), Baal (Syria), the Great Mother (Phrygia), Yahweh (Palestine), and Mithra (Kurdistan) were worshipped in their native lands as well as in Rome and other cosmopolitan centres. With few exceptions, each of these religions, originally tied to a specific geographical area and people, had traditions extending back centuries before the Hellenistic period. In their homeland they were inextricably tied to local loyalties and ambitions. Each persisted in its native land with little perceptible change save for its becoming linked to nationalistic or messianic movements (centring on a deliverer figure) seeking to overthrow Greco-Roman political and cultural domination. Indeed, many of these native religions underwent a conscious archaism during this period, attempting to recover earlier forms and practices. Old texts in native languages (especially those related to relevant themes such as kingship) were recopied, national temples were restored, and old, mythic traditions were revived. From Palestine to Persia one may trace the rise of Wisdom literature (the teachings of a sage concerning the hidden purposes of the deity) and Apocalyptic traditions (referring to a belief in the dramatic intervention of a god in human and natural events) that represent these central concerns—*i.e.*, national destiny, the importance of traditional lore, the saving power of kingship, and the revival of mythic images. Each of these native traditions likewise underwent hellenization (modifications based on Greek cultural ideas), but in a manner frequently different from their diasporic counterparts.

The dispersion of local and area cults and the resultant changes

Each of these native religions also had diasporic centres that exhibited marked change during the Hellenistic period. There was a noticeable lessening of concern on the part of the members of the dispersed religious group for the density and fortunes of the native land and also a relative severing of the traditional ties between religion and the land. Certain cult centres remained sites of pilgrimage or objects of sentimental attachment; but the old beliefs in national deities and the inextricable relationship of the deity to certain sacred places was weakened. Rather than a god who dwelt in his temple, the diasporic traditions evolved complicated techniques for achieving visions, epiphanies (manifestations of a god), or heavenly journeys to a transcendent god. This led to a change from concern for a religion of national prosperity to one for individual salvation, from focus on a particular ethnic group to concern for every man. The prophet or saviour replaced the priest and king as the chief religious figure. In the diasporic centres, as is generally characteristic of immigrant groups, there were two circles. The first (or inner circle) was composed of devout, full-time adherents of the cult for whom the deity retained a separate and decisive identity (*e.g.*, Yahweh alone, there is one Zeus Sarapis, Isis is queen over all). Its membership was drawn from the ethnic group for whom the deity was indigenous, and the group tended to continue to speak the native language. The second (or outer circle) was composed of either second and third generation immigrants or converts from groups for whom the religion was not native. These individuals tended to speak Greek, and this began the lengthy process of reinterpretation of the archaic religion. Ancient sacred books were translated or paraphrased into Greek—*e.g.*, the 4th–3rd-century-BC Babylonian priest Berosus' version of Babylonian materials, the 4th–3rd-century-BC Egyptian priest Manetho's Egyptian accounts, the Jewish Septuagint (Greek version of the Old Testament), or the 1st-century-AD Jewish historian Josephus' *Antiquities of the Jews*, and the ethnic histories of the 1st-century-BC Greek writer Alexander Polyhistor. In each case the material was reinterpreted both in light of common Hellenistic ideals and in accord with the special traditions and needs of the diasporic community. Both the inner and outer circles fostered esotericism (secrets to be known only by initiates)—the former by its use of native language and its oral recollection of traditions from the homeland; the latter by its use of allegory and other similar methods to radically reinterpret the sacred texts. The difference between these groups was responsible for many shifts in the character of the religion. Most notable was the shift from elements characteristic of native religion in its definition of religion (*e.g.*, local tradition and custom, informal knowledge orally transmitted, and birth) to formulated dogma, creeds, law codes, and rules for conversion and admission that were characteristic of diasporic religion. It was a shift from "birthright" to "convinced" religion.

The history of Hellenistic religions is rarely the history of genuinely new religions. Rather it is best understood as the study of archaic Mediterranean religions in their Hellenistic phase within both their native and diasporic settings. It is usually by concentrating on the diaspora that the Hellenistic character of a cult has been described.

History. *Religion from the death of Alexander to the reformation of Augustus: 323–27 BC.* The conquests of Alexander opened the way for religious interchange between East and West; the political structures left behind by Alexander and continued by his successors provided strong incentives for the hellenization of native religions. Characteristic of this first period of Hellenistic religious history were the following developments: (1) the introduction of Oriental cults into the West, especially those associated with female deities who were either worshipped in frenzied rites of self mutilation (*e.g.*, the Phrygian Cybele, brought to Rome in 204 BC; the Syrian Atargatis; or the Cappadocian Ma-Bellona) or in adoring contemplation of their beneficence and gentle rites of divine rebirth (*e.g.*, the Egyptian Isis, whose cult was widespread in the Greco-Roman world by the middle of the 2nd century BC); (2) the hellenization of native cults (most famously that of the archaic Egyptian god Sarapis whose Greek form was promulgated by Ptolemy I, the founder of the Egyptian Ptolemaic dynasty in 305 BC); (3) the development of the ideology of divine kingship based on Oriental kingship traditions; and (4) the rise of nationalistic and messianic movements directed against internal and external hellenization; *e.g.*, the Maccabean rebellion led by Judas Maccabeus against Jewish hellenizing parties and the Syrian overlords in 167–165 BC, and the numerous Egyptian rebellions, especially that led by the Egyptian independence leader Harmakhis in Thebais in 207/6 BC.

Religious interchange between East and West

Religion from the Augustan reformation to the death of Marcus Aurelius: 27 BC–180 AD. Oriental cults underwent their most significant expansion westward during this period. Particularly noticeable was the success of a variety of prophets, magicians, and healers—*e.g.*, John

the Baptist, Jesus, Simon Magus, Apollonius of Tyana, Alexander of Abonouteichos, and the cult of the healer Asclepius—whose preaching corresponded to the activities of various Greek and Roman philosophic missionaries. A developing tension between these "new" Eastern religions and the archaic Greco-Roman traditions was expressed internally in the attempt by the emperor Augustus to revive traditional Roman religious practices. Attempts were made to expel foreigners or to suppress foreign worship—*e.g.*, the suppression of the Bacchic mysteries (salvation cults devoted to the god Dionysus, or Bacchus) in Rome in 186 BC, or the numerous attempts to prohibit the worship of the Egyptian goddess Isis in Rome, beginning in 59 BC. The Augustan reformation also restored Roman sacred books and Greek temples. Externally, the developing tension was expressed in wars, riots, and persecutions, such as the Jewish-pagan riots in Alexandria in AD 38 and 115–116, the Jewish-Roman wars of AD 66–70 and 132–135, or the beginning of the persecution of Christians under the Roman emperor Nero in AD 64. Another cause of tension was the elaboration of a full-blown cult of "Emperor worship," beginning with the deification of Augustus (September 17, 14 AD).

Developing tensions between the Greco-Roman and the Eastern religions

Religion from Commodus to Theodosius I: AD 180–395. After the death of the "philosopher-king" Marcus Aurelius (121–180), his son Commodus (161–192) became emperor, and a period of political instability began. The dominant feature of the concluding period of the Hellenistic Age—and shortly thereafter—was the rapid growth of Christianity throughout the Roman Empire, culminating in the conversion to Christianity of the emperor Constantine (313) and the religious legislation of the emperor Theodosius (347–395) affirming in 380 the dogmas of the Christian Council of Nicaea, which was convened in 325 under the auspices of Constantine, and prohibiting paganism in a decree of 392. In this period the various Hellenistic cults were victims of active hostilities, which were expressed through prohibition, acts of violence, and theological polemics between pagans and Christians (*e.g.*, the pagan philosophers Maximus of Tyre and Celsus, and the Christian philosophical theologians Irenaeus, Tertullian, and St. Clement of Alexandria, all of the 2nd century); but there were also brief periods of pagan vitality. The Neoplatonic school (based on a complicated system of levels of reality) of Plotinus (203/4–269) and Porphyry (*c.* 234–*c.* 305) represented the culmination of Hellenistic religious philosophy. The Syrian solar cults of Sol Invictus (the unconquerable Sun) and Jupiter Dolichenus played an important role under the emperors Antoninus (86–161), the Severans—Septimius (146–211) and Alexander (208–235)—and Elagabalus (reigned 218–222) and were hailed as the supreme deities of Rome under Aurelian (d. 275), whose Sun temple was dedicated in 274. From Parthia, the dualistic and spiritual teachings of the Iranian prophet Mani (*c.* 215–274) were widely disseminated throughout the Empire. The Persian cult of the ancient Iranian god of light, Mithra, spread rapidly throughout the western and northern Empire during the 3rd through 5th centuries. Although these various traditions enjoyed brief imperial patronage under Julian (reigned 361–363), they eventually were subsumed under the political and religious hegemony of Christianity (see below).

Mythology and theology. The archaic religions of the Mediterranean world were primarily religions of etiquette. At the centre of these religions were complex systems governing the interrelationships between gods and men, individuals and the state, living men and their ancestors. The entire cosmos was conceived as a vast network of relationships, each component of which, whether divine or human, must know its place and fulfill its appointed role. The model for this all-encompassing system was the divine society of the gods, and the map of this system was the order of the planets and stars. Through astrology, divination, and oracles, man discerned the unalterable patterns of destiny and sought to bring his world (the microcosm) into harmony with the divine cosmos (the macrocosm; see also ASTROLOGY).

The centrality of the relationship between the gods and men

This archaic pattern of affirming and celebrating the order of the cosmos was expressed in the typical creation myth of the Near Eastern and Mediterranean world, a creation by combat between the forces of order and chaos. Order was understood to be something won in the beginning by the gods, and it was this primordial act of salvation that was renewed and re-experienced in the cult.

In the Hellenistic period a new religious world was experienced that required new religious expressions. The old religions of conformity and place no longer spoke to this new religious situation and its questions. What if the law and order of the cosmos was no longer seen as the creative expression of limits and the delineation of roles, but rather as an evil, perverse, confining structure from which man and the cosmos must escape? Rather than the archaic structures of celebration and conformity to place, the new religious mood spoke of escape and liberation from place and of salvation from an evil, imprisoned world. The characteristic religion of the Hellenistic period was dualistic. Man sought to escape from the despotism of this world and its rulers (exemplified by the seven planetary spheres) and to ascend to another world of freedom. Hellenistic man saw himself as an exile from his true home, the Beyond, and he sought for ways to return. He strove to regain his place in the world beyond this world where he truly belonged, to encounter the god beyond the god of this world who was the true god, and to awaken that part of himself (his soul or spirit) that had descended from the heavenly realm by stripping off his body, which belonged to this world. The questions that the religions of the Hellenistic period sought to answer may be seen in a fragment from the 2nd-century Anatolian Gnostic teacher Theodotus: "What liberates is the knowledge of who we were [before our earthly existence] and what we have become [on earth]; where we were [the Beyond] and the place to which we have been thrown [the world]; where we are going and from what are we redeemed; what is birth and what is rebirth" (preserved in Clement of Alexandria, *Excerpta ex Theodoto*, 78.2).

Salvation as liberation through knowledge of one's origin, identity, and destiny

The gods. In the Greco-Roman world during the Hellenistic period, archaic deities were transformed in part because of the new spirit of the age and in part by foreign influences. A number of the old chthonic (underworld) and agricultural (fertility) gods and the old agricultural mysteries (corporate renewal religions related to fertility concepts) fundamentally altered their character. Rather than an expression of the alternation of life and death, of fertility and sterility, and a celebration of the promise of renewal for the land and the people, the seasonal drama was homologized to a soteriology (salvation concept) concerning the destiny, fortune, and salvation of the *individual* after death. The collective agricultural rite became a mystery, a salvific experience reserved for the elect (such as the Greek mystery religion of Eleusis). Other traditions even more radically re-interpreted the ancient figures. The cosmic or seasonal drama was interiorized to refer to the divine soul within man that must be liberated. Such cults were dualistic mysteries distinguishing sharply between the body and soul. They taught that it is the soul alone that was initiated by passing through death or the Underworld, or by being dismembered so that it might be freed from the body and regain its rightful mode of spiritual existence (such as the Orphic—mystical—reinterpretation of the role of the agricultural god Dionysus). In the gnostic mysteries (the esoteric dualistic cults that viewed matter as evil and the spirit good), this process was carried further through the identification of the experiences of the soul that was to be saved with the vicissitudes of a divine but fallen Soul, which had to be redeemed by cultic activity and divine intervention. This view is illustrated in the concept of the paradoxical figure of the saved saviour, *salvator salvandus*.

Other deities, who had previously been associated with national destiny (*e.g.*, Zeus, Yahweh, Isis), were raised to the status of transcendent, supreme deities whose power and ontological status (relating to being or existence) far surpassed the other gods, who were understood as their servants or antagonists. The religious man sought to make contact with, or to stand before, this one, true god

Individual piety

of the Beyond. The piety of the individual was directed either toward preparing himself to ascend up through the planetary spheres to the realm of the transcendent god or toward calling the transcendent god down that he might appear to him in an epiphany or vision. These techniques for achieving ascent or a divine epiphany make up the bulk of the material that has usually been termed magical, theurgic (referring to the art of persuading a god to reveal himself and grant salvation, healing, and other requests), or astrological and that represents the characteristic expression of Hellenistic religiosity.

Mythology. The cosmogonies (dealing with the origins of the world) and cosmologies (dealing with the ordering of the world) of the Hellenistic period centred around the problem of accounting for the distance between this world and the Beyond, or on accounting for the evil nature of this world and its gods. Many mythic schema were employed regarding the origin and ordering of this world. It was viewed as being: the result of the conscious or unconscious emanation from the transcendent realm; the result of the fall of a deity from the Beyond; the creation of a hostile, ignorant, or evil deity; or a joke or mistake. The purpose of this speculation was both pragmatic and soteriological: if one could determine how this creation came into being, one could reverse it or overcome it and be saved.

Institutions. The temples and cult institutions of the various Hellenistic religions were repositories of the knowledge and techniques necessary for salvation and were the agents of the public worship of a particular deity. In addition, they served an important sociological role. In the new, cosmopolitan ideology that followed Alexander's conquests, the old nationalistic and ethnic boundaries had broken down and the problem of religious and social identity had become acute. The Hellenistic Age was characterized by the rapid growth of private religious societies (*thiasoi*). Though some were organized according to national origin or trade, the majority were dedicated to the worship of a particular deity. In many instances these groups began as immigrant associations (*e.g.*, an Egyptian association of devotees of Amon was chartered in Athens at the beginning of the 3rd century BC); but they often transcended these origins and became a new form of religious organization in which citizens of various countries, freemen and slaves, could be united by their common devotion and share in a common religious heritage. Admission to such groups was voluntary (in contradistinction to the archaic national or familial religious organizations) and demanded the payment of dues, submission to collective authority, and the acceptance of strict codes of morality. Most of these groups had regular meetings for a communal meal that served the dual role of sacramental participation (referring to the use of material elements believed to convey spiritual benefits among the members and with their deity) and the social function of brotherhood; *i.e.*, the security of membership in a group and a shared sense of identity.

Hellenistic religions as voluntary associations

The influence of Hellenistic religions. The archaic gods worshipped during the Hellenistic period possessed a remarkable longevity. The Eleusinian Mysteries, founded in the 15th century BC, ceased in the 4th century AD; Dionysus, whose name first appears on tablets dated to *c.* 1400 BC, was last celebrated in the beginning of the 6th century AD; the last temple of Isis, whose cult extended back to the 2nd millennium in Egypt, was closed in 560 AD. Yet even after these ceased as objects of devotion in the post-Constantinian period, they continued to exercise their influence. Hellenistic philosophy (Stoicism, Cynicism, Neo-Aristotelianism, Neo-Pythagoreanism, and Neoplatonism) provided the key formulations for Jewish, Christian, and Muslim philosophy, theology, and mysticism through the 18th century. Hellenistic magic, theurgy, astrology, and alchemy remained influential until modern times in both East and West. Theosophy and other forms of the occult, especially since the Renaissance, drew their inspiration from the Hellenistic mystery cults, Hermeticism (Greco-Egyptian astrological, magical, and occultic movement), and Gnosticism. Various Jewish, Christian, and Muslim sectarian groups continued the theologies of many of the Hellenistic religions (especially dualistic modes of thought). Hellenistic sacred art and architecture has remained the basis of Christian and Jewish iconography and architecture to the present day. Figures such as Alexander the Great generated a vast body of religious literature, especially in the Middle Ages. Many of the symbols and legends associated with Hellenistic deities persisted in folk literature and hagiography (stories of saints and "holy" persons). The basic forms of worship of both the Jewish and Christian communities were heavily influenced in their formative period by Hellenistic practices, and this remains fundamentally unchanged to the present time. Finally, the central religious literature of both traditions—the Jewish Talmud (an authoritative compendium of law, lore, and interpretation), the New Testament, and the later patristic literature of the Early Church Fathers—are characteristic Hellenistic documents both in form and content.

BIBLIOGRAPHY. The most useful cultural and political history containing valuable discussions of controversial issues with full bibliography is R. COHEN, *La Grèce et l'hellenisation du monde antique*, 2nd ed. (1948). W.W. TARN and G.T. GRIFFITH, *Hellenistic Civilisation*, 3rd ed. (1952); and M.I. ROSTOVTZEFF, *The Social and Economic History of the Hellenistic World*, 3 vol. (1941), remain the standard English works. M.P. NILSSON, *Geschichte der griechischen Religion*, 2nd ed., vol. 2 (1961); and K. PRUMM, *Religionsgeschichtliches Handbuch für den Raum der altchristlichen Umwelt* (1954), are indispensible as general handbooks and for the rich bibliographies they contain. The magnificent encyclopaedia now in progress, *Reallexikon für Antike und Christentum* (1950–), will be, when completed, the best single resource for the study of Hellenistic and early Christian religion.

A varied selection of texts in good English translation may be found in E.R. BEVAN, *Later Greek Religion* (1927); F.C. GRANT (ed.), *Hellenistic Religions: The Age of Syncretism* (1953); E. BARKER (ed.), *From Alexander to Constantine* (1956); and R.M. GRANT (ed.), *Gnosticism: A Sourcebook of Heretical Writings from the Early Christian Period* (1961).

Important general interpretations include: F. CUMONT, *Oriental Religions in Roman Paganism* (Eng. trans. from 2nd French ed., 1911), the 4th ed. *Les religions orientales dans le paganisme romain* (1929), is considerably expanded; P. WENDLAND, *Die hellenistisch-römische Kultur in ihren Beziehungen zu Judentum und Christentum* (1912); H.R. WILLOUGHBY, *Pagan Regeneration* (1929); A.J. FESTUGIERE, *L'Ideal religieux des Grècs et l'Evangile* (1932), *Personal Religion among the Greeks* (1960); A.D. NOCK, *Conversion* (1933); M.P. NILSSON, *Greek Piety* (Eng. trans. 1948); E.R. GOODENOUGH, *Jewish Symbols in the Greco-Roman Period*, 13 vol. (1953–70); S.K. EDDY, *The King is Dead: Studies in the Near Eastern Resistance to Hellenism 334–31 B.C.* (1961); E.B. DODDS, *Pagan and Christian in an Age of Anxiety* (1965); A.J. TOYNBEE (ed.), *The Crucible of Christianity* (1969); and J. FERGUSON, *The Religions of the Roman Empire* (1970). In addition to these works (all of which contain full bibliographies), see the individual volumes in the important series in progress, ed. by M.J. VERMASEREN, *Études préliminaires aux religions orientales dans l'Empire romain.*

(J.Z.S.)

Helmholtz, Hermann von

One of the greatest scientists of the 19th century, Hermann Ludwig Ferdinand von Helmholtz made fundamental contributions to physiology, optics, electrodynamics, mathematics, and meteorology, but is best known for his statement of the law of the conservation of energy. In addition, he brought to his laboratory research the ability to analyze the philosophical assumptions on which much of 19th-century science was based, and he did so with clarity and precision.

Born at Potsdam near Berlin, on August 31, 1821, Helmholtz' delicate health confined him to home for his first seven years. His father was a teacher of philosophy and literature at the Potsdam Gymnasium, and his mother was descended from William Penn, the founder of Pennsylvania. The young Helmholtz thrived in this intellectual, though impecunious, home atmosphere, where his father taught him Latin, Greek, Hebrew, French, English, Italian, and Arabic. His father also introduced him to the prevailing Hegelian "Nature philosophy," which deduced conclusions from philosophical constructs instead of empirical data. Pretending to contain the key



Helmholtz.
By courtesy of the Ruprecht-Karl-Universitat, Heidelberg, Germany

to the secrets of nature, this transcendental view became an important influence on his career.

After graduating from the Gymnasium, Helmholtz in 1838 entered the Friedrich Wilhelm Medical Institute in Berlin, where he received a free medical education on the condition that he serve eight years as an army doctor. At the institute he did his first research under the distinguished physiologist, Johannes Müller. He attended physics lectures, worked his way through the standard textbooks of higher mathematics, and learned to play the piano with a skill that later helped him in his work on the sensation of tone.

On graduation from medical school in 1843, Helmholtz was assigned to a regiment at Potsdam. Since his army duties were few, he did experiments in a makeshift laboratory he set up in the barracks; at that time he also married Olga von Velten, daughter of a military surgeon.

Early success

Before long, Helmholtz' obvious scientific talents led to his release from military duties. In 1848 he was appointed assistant at the Anatomical Museum and extraordinary professor at the Academy of Fine Arts in Berlin, moving the next year to Königsberg in East Prussia (now Kaliningrad), to become assistant professor and director of the Physiological Institute. But Königsberg's harsh climate was injurious to his wife's health, and, in 1855, he became professor of anatomy and physiology at the University of Bonn, moving in 1858 to Heidelberg. During these years, his scientific interests progressed from physiology to physics. His growing scientific stature was further recognized in 1871 by the offer of the Chair of Physics at the University of Berlin; in 1882, by his elevation to the nobility; and, in 1888, by his appointment as first Director of the Physico-Technical Institute at Berlin, the post he held for the rest of his life.

The variety of positions he held reflects his interests and competence but does not reflect the way in which his mind worked. He did not start out in medicine, move to physiology, then drift into mathematics and physics. Rather, he was able to coordinate the insights he had acquired from his experience in these disciplines and to apply them to every problem he examined. His great work, *Handbook of Physiological Optics* (1867), was characterized—like all of his scientific works—by a keen philosophical insight, molded by exact physiological investigations, and illustrated with mathematical precision and sound physical principles.

Attack on "Nature philosophy"

The general theme that runs through most, if not all, of Helmholtz' work may be traced to his rejection of "Nature philosophy," and the violence of his rejection of this seductive view of the world may well indicate the early attraction it had for him. "Nature philosophy" derived from the German philosopher Immanuel Kant, who in the 1780s had suggested that the concepts of time, space, and causation were not products of sense experience but mental attributes by which it is possible to perceive the world. Therefore, the mind did not merely record order in nature, as the Empiricists insisted; rather, the mind organized the world of perceptions, so that, reflecting the divine reason, it could deduce the system of the world from a few basic principles. Helmholtz opposed this view by insisting that all knowledge came through the senses. Furthermore, all science could and should be reduced to the laws of classical mechanics, which, in his view, encompassed matter, force, and, later, energy, as the whole of reality.

Helmholtz disagreed with his mentor in physiology, Müller, who, influenced by the biological component of "Nature philosophy," assumed that the organism as a whole was greater than the sum of its physiological parts. Above all, Müller believed, organs were endowed with a nonmaterial vital force that coordinated physiological action to produce the harmonious organic behaviour that characterized the living creature; and since this vital force was not subject to experimental investigation, physiological experiment was impossible. This attitude was anathema to Helmholtz and he set out to destroy it. His doctor's thesis in 1842 on the connection between nerve fibres and nerve cells of invertebrates introduced him to physiological processes and stimulated him to investigate the problem of the nature and origin of animal heat. His measurement of the velocity of the nerve impulse in 1850 followed immediately upon Müller's remark that because this impulse was an example of "vitalism," it would never be measured. There was no "vital force," he was convinced, because organisms employed the same forces as those in inorganic nature, forces that he now considered in terms of classical mechanics. His resulting paper, "On the Conservation of Force," gave him claim to be one of the discoverers of the principle of the conservation of energy. His later physiological investigations continued his attack on both vitalism—nonmaterial cause of many living processes—and "Nature philosophy."

Discovery of the conservation of energy

The eye, furthermore, considered by vitalists to be an exquisite example of the divine mind at work, was shown by Helmholtz to be a rather imperfect piece of workmanship. While conducting these researches, he invented the ophthalmoscope and the ophthalmometer, the latter instrument permitting the measurement of accommodation of the eye to changing optical circumstances.

In 1863 Helmholtz published his masterly work *On the Sensation of Tone as a Physiological Basis for the Theory of Music*. Here, fundamental physiological and anatomical researches were combined with the mathematical and physical analysis of wave motion to produce a firm foundation for the aesthetic analysis of music. Helmholtz now used the physics and physiology of sound to attack Kant's theory of perception. Similarly, Helmholtz' views on non-Euclidean geometry appear to have been stimulated by Kant's statements that Euclidean geometry reflected the basic structure of the human mind.

Helmholtz' work in electricity and magnetism reveals his conviction that classical mechanics was probably the best mode of scientific reasoning. He was one of the first German scientists to appreciate the work in electrodynamics of the British scientists Michael Faraday and James Clerk Maxwell. Faraday had appeared to strike at the foundation of Newtonian physics by his unorthodox rejection of action at a distance; that is, action between two bodies in space without alteration of the medium between them. Maxwell, however, by interpreting the mathematics of Faraday's laws, was able to show that there was no contradiction between Newtonian physics and classical mechanics. Helmholtz further developed the mathematics of electrodynamics. He spent his last years in an unsuccessful attempt to reduce the entire field of electrodynamics to a minimum set of mathematical principles, an attempt in which he had to rely increasingly on the mechanical properties of the "ether," a weightless substance then thought to pervade all space. Though he was unsuccessful in his original purpose, he was almost able to deduce all electromagnetic effects from the ether's supposed properties. The discovery of radio waves by his pupil, Heinrich Hertz, in 1888, was viewed as the experimental confirmation of the theories of Faraday, Maxwell, and Helmholtz. The special and general theories of relativity, proposed by Albert Einstein, destroyed Helmholtz's theories by eliminating the ether.

Helmholtz' early work on sound and music had led him

to the study of wave motion. His work on the conservation of energy familiarized him with the problems of energy transfer. These two areas coalesced in his later years in his studies of meterology, but the phenomena were so complex that he could do little more than point the way to future areas of research.

Helmholtz was the end product of the development of classical mechanics. He pushed it as far as it could go. When he died in Berlin on September 8, 1894, the world of physics was poised on the brink of revolution. The discovery of X-rays, radioactivity, and relativity led to a new kind of physics in which Helmholtz' achievements, although impressive, had little to offer the new generation.

BIBLIOGRAPHY. There are two biographies of Helmholtz available to the reader of English. LEO KOENIGSBERGER, *Hermann von Helmholtz*, 3 vol. (1902–03; abr. Eng. trans., 1906, reprinted 1965), is often technical and sometimes difficult to understand. J.G. MCKENDRICK, *Hermann Ludwig Ferdinand von Helmholtz* (1899), deals only with Helmholtz' medical career. RICHARD M. and ROSLYN P. WARREN have published a collection of Helmholtz' writings on perception, entitled *Helmholtz on Perception: Its Physiology and Development* (1968), with critical comments. Helmholtz' *Popular Lectures on Scientific Subjects*, 2 vol. (1873; 2nd series, 1881), are excellent introductions to his thought.

(L.P.W.)

Helsinki

The capital of Finland since 1812, and that nation's leading seaport, Helsinki, despite a turbulent history stretching back for more than four centuries, is one of the most modern of European cities in terms of character and appearance. Founded in 1550, it was ravaged by a plague in 1710 and burned to the ground in 1713. Large parts of the city were destroyed by another great fire in 1808, and there is little physical evidence of its early history. Its development has also mirrored the strife-torn history of Finland: centuries of competitive Swedish and Russian domination, linked with national aspirations of marked individuality. Its destiny was controlled by a succession of visiting kings and tsars before it emerged, in 1917, as the administrative centre of the new Republic of Finland, and, as such, developed into the Helsinki of today. By the 1970s, it had become a vital economic and cultural centre, whose regional and international importance was well indicated by the popular titles given the city—"the daughter of the Baltic" and "the city between east and west." At intervals—for example, as host city to the 1952 Olympic Games—Helsinki has been known as an international sports centre, but from the 1960s onward, its name has equally often been connected with attempts to lessen international tension, as in the case of the American–Soviet talks on the limitation of strategic arms, begun in 1969. A further apellation—"the white city of the north"—draws attention to the physical appearance of the city, built largely of local light-coloured granite, and to the fact that, after Reykjavík, in Iceland, it is the world's most northerly capital. It has a spacious and well-planned atmosphere, and a stark but beautiful setting.

Historical development. For many centuries Finland, a part of the Swedish realm, was the scene of repeated wars between that nation and the growing Russian state. It was against this background that King Gustav Vasa of Sweden founded Helsinki, on June 12, 1550. The young settlement—whose Swedish name is Helsingfors—was intended to compete, in the economic field, with the city of Reval (now Tallinn, in the Estonian S.S.R.) situated on the opposite side of the Gulf of Finland. Helsinki was originally located on the Vantaa estuary but moved down to its present location, a promontory with a fine, island-sheltered harbour, in 1640. The trading centre numbered fewer than 2,000 when the plague of 1710 killed most of the inhabitants. Three years later it was burned by Swedish forces retreating from an attack by the fleet of the Russian tsar Peter the Great, and development was hindered by further Russian attacks later in the 18th century. Another fire, in 1808, added to the city's difficulties.

Early misfortunes

In 1809, Finland was ceded to Russia. The national capital during the period of Swedish hegemony had been Turku (Åbo) in the southwest of the country, facing Sweden, and thus both political and military motives played a part in the decision of the Russian tsar Alexander I to move the capital of the new grand Duchy from Turku to Helsinki, in 1812. In 1748 the settlement became much more secure when a fortress—once called the Gibraltar of the north—was constructed on a group of small islands outside the harbour of Helsinki, although the fortress (called Sveaborg by the Swedes and later Suomenlinna by the Finns) was bombarded by a French–British fleet in 1855 during the Crimean War.

Meanwhile, the centre of Helsinki had been completely reconstructed in an impressive imperial and monumental style, under the influence of a German-born architect, C.L. Engel, active in the then Russian capital, St. Petersburg. The completed city centre was comprised of a number of distinctive buildings still to be seen in the 1970s. These include government buildings, the main building of Helsinki University, and the cathedral, known as the Great Church, completed in 1852. All of them surround the broad expanse of Senate Square, which is often considered to be one of the most beautiful in the whole of Europe. Not far from the square rise the cupolas of the Uspenski Orthodox Cathedral, one of the few recognizable reminders of the period of Russian rule. Tsar Nicholas I visited Helsinki in person in 1833, and the health spa that subsequently opened attracted members of the Russian upper class from Moscow and St. Petersburg over a period of decades.

Reconstruction and growth

In the 1860s and 1870s industrialization brought great changes to the social and economic fabric of the city and started a development that made Helsinki Finland's largest industrial city. Population—a mere 4,000 in 1810—increased rapidly and was over 22,000 by 1860, 60,000 by 1890, and 111,000 by 1904. Although the policy of Tsar Alexander II toward Finland and its capital was liberal, by the end of the 19th century the Russians saw social changes within the grand duchy, including a national Finnish movement, as a threat to their position. Nicholas II and his governor general, Nikolay Bobrikov, initiated a russification program. Bobrikov, who had assumed dictatorial powers, was shot by a young Finn in 1904, and Helsinki became the centre of a group of activists working for the liberation of Finland. A national declaration of independence was proclaimed by the parliament in Helsinki in December 1917, and the capital was occupied by Finnish red guards and Russian units during the ensuing short but bloody civil war between "white" and "red" zones. Conditions soon became more stabilized, with the Helsinki parliament electing Finland's first president in 1919. The following half century saw the full development of Helsinki into an important centre of trade, industry, and culture, a process interrupted only by the war years of the 1940s.

Modern architecture

A distinctive architectural style, modern in concept and graceful in execution, achieved prominence during this period, complementing the buildings of a century earlier. The railway station (designed by Eliel Saarinen, 1914) and the National Pensions Institute (by Alvar Aalto, 1956) were notable examples of the work of Finnish architects with international reputation. At the turn of the century the population had numbered less than 100,000; by the early 1970s greater Helsinki—including a number of modern suburbs acquired during the 1940s—had more than 815,000 inhabitants. The spacious and well-planned atmosphere of the city was nevertheless successfully maintained.

The site. Helsinki is surrounded by the sea on three sides, and only 68 square miles (177 square kilometres) of its total administrative area of 173 square miles (448 square kilometres) consist of land. The city can expand only to the north, northwest and northeast, and it is in these areas that the 17 new suburbs of the mid-20th century are located. Helsinki is located at 60° N latitude (comparable to the location of Greenland's Kap Farvel, Nunivak Island of Alaska, and the upper Kamchatka Peninsula of the Soviet Union), but the Baltic Sea influ-

ence contributes to a rather mild climate. The average summer temperature is 61° F (+16° C) and the average winter temperature 21° F (−6° C). Icebreaker activity keeps at least part of the extensive harbour facilities open during the winter months. Helsinki is mainly built on rock, but there are only a few high points, while parts of the present centre of the city are actually built on filled-in seabed.

The people. Helsinki proper had 525,628 inhabitants in 1969, of whom a majority—291,000—were women. Greater Helsinki numbered about 815,000 persons in 1971, nearly 18 percent of Finland's total population. There was a continued movement of people into and out of the central city, with an average annual net gain of some 5,000 citizens by the late 1960s and early 1970s. Those leaving Helsinki proper went mostly to the suburbs of greater Helsinki. In the early 1970s the majority of the city's inhabitants—some 60 percent—had been born elsewhere. Reflecting its heritage, Helsinki is a bilingual city, with 85 percent of the population claiming Finnish, and 14 percent Swedish, as their mother tongues. As in the case of other northern European cities, the religious composition of Helsinki is overwhelmingly Protestant, with 85 percent adhering to the Lutheran faith. A small minority—about 2 percent—is Orthodox. Of the economically active population, 24 percent work in industry, 30 percent in the service professions, and 25 percent in commercial activities. About 22 percent of the population of Helsinki is in the 0–14 age group, with a tenth in the over-65 group.

Bilingualism

Economic life. Helsinki's economic life and development is based on its excellent harbours and on good railway and road connections to the extensive interior of the nation. More than half of Finland's total imports consequently pass through Helsinki, and are handled by the large wholesale enterprises in the city. Only 10 percent of the national exports, on the other hand, pass through Helsinki, as the largest export ports are elsewhere along the Finnish coast. In her capacity as Finland's largest industrial city, Helsinki is responsible for 17 percent of the nation's gross industrial production, with food and metal processing, printing, textiles, and clothing among the main industries. Many of those products exhibit the functional contemporary design for which the nation is famous. In connection with expansion programs, some of the larger industries had moved to the outer Helsinki region in the 1970s. The Wärtsilä shipyard (producing icebreakers, liners, and cable-laying ships) and the wares of the Arabia porcelain factory, one of the largest of its kind in Europe, are, in their distinctive ways, internationally famous.

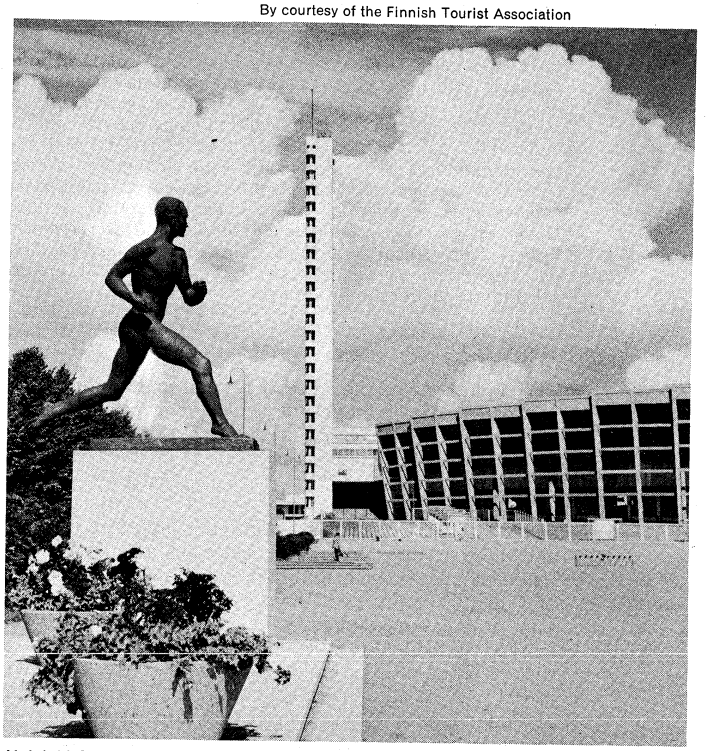
Political and governmental institutions. The parliament and the government of Finland meet and function in Helsinki. A city council has headed the municipal government since 1875. Elected every fourth year, it is the highest policy-making body of the capital. The city board and various committees prepare and execute the decisions of the council. The city board includes the mayor and six assistant city managers, together with 11 members who are elected each year by the City Council, on the basis of the relative strength of the political parties represented on that body. The city managers are responsible for matters concerning real estate, hospitals, public utilities, education and culture, construction and public works, and personal and social welfare, while the mayor is responsible for central administration and finance. The city board is assisted by some 50 administrative boards and commissions.

City managers

The presidential palace is located in Helsinki, which, as a capital, also houses the highest courts of the country, and the Bank of Finland.

Services. Many of the services provided for by the city of Helsinki—notably education in elementary and trade schools and in medical care—are free, or almost free. Of the total municipal expenditure, some 27 percent is used for medical care and social welfare functions, 10 percent for education, 9 percent for public works, and some 20 percent for such public utilities as electricity, water, gas, and transportation. Large sums are also invested in long-term transportation projects. In 1968 the City Council voted to construct Helsinki's first subway, which will total 7.2 miles, and of which 2.4 miles will be tunnel. The first Finnish-constructed "test train" for the subway was ready in 1972. Some 400,000 passengers are carried daily by municipally run buses and streetcars. The number of motor vehicles in the city is close to 100,000, of which more than 80,000 are private cars. The physical geography of Helsinki, constricting routes through the neck of a promontory, makes transportation and traffic planning a constant problem, even in this well laid out city.

By courtesy of the Finnish Tourist Association



Helsinki Stadium, with a statue (foreground) of Paavo Nurmi, the famous Finnish athlete.

Cultural life and recreation. Thirteen theatres, an opera and ballet company, and three symphony orchestras contributed to the year-round cultural activity of the Finnish capital in the early 1970s. An annual Helsinki festival features world famous orchestras and artists and a program of rich variety. In addition to museums and art galleries, the cultural interests of Helsinki citizens are furthered by a fine, modern city theatre, by Timo Penttilä, with an audience capacity of more than 1,000, and a new concert building, of a striking design, by Alvar Aalto. Helsinki University, with 23,000 students, is the largest university in Scandinavia.

Helsinki is a city for recreation. The sea, open or frozen, gives opportunity for sports ranging from yachting to a somewhat perilous car racing, and there are close to 300 separate sports grounds for both winter and summer events. The city provides close to 200 miles of ski routes, and there are numerous spacious parks, where young and old alike can enjoy the Helsinki air, which—thanks to the influence of sea breezes—is surprisingly clean for an industrial city. The many beaches and rocky islands are covered with sunbathers in the summer, while in winter the frozen bays are dotted with people fishing through the ice.

BIBLIOGRAPHY

Geography and topography: HEIKKI BROTHERUS, *Helsinki and Her People* (1966), is the fullest description of the city, with many illustrations. See also MATTI KURJENSAAR and PEKKA LOUNELA, *Helsinki in Color* (1966); FRED RUNEBERG, *Helsinki* (1964), mainly pictures and annotations; and G. MARTENSON (ed.), *Helsinki, Capital of Finland* (1950), also a well-illustrated account. LAURI AHO (ed.), *Helsinki: The Face of the Capital* (1959); and CLAIR AHO, *Helsinki,*

Daughter of the Baltic (1951), are almost entirely illustrations; KEYO PETAJA (ed.), *Helsinki* (1970), covers the city's architecture. OLIVER WARNER, *A Journey to the Northern Capitals* (1968), is a personal account with a section on Helsinki.

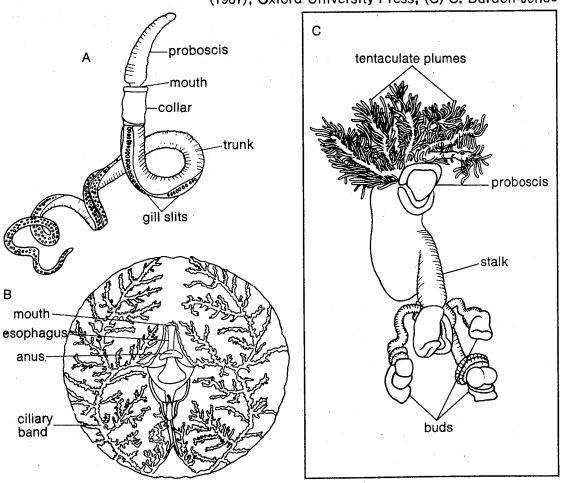
Scientific aspects and economics: The HELSINKI HARBOR BOARD, *Port of Helsinki* (1967), describes that particular aspect of the city; while W. SMITH and P. POLVINEN (eds.), *Helsinki Metropolitan Transportation Study*, bk. 1, *Helsinki* (1969), is an exhaustive survey. S.T.H. JAATINEN, *The Birthplace Field of Helsinki According to the Census of 1950* (1962), illustrates past population patterns.

(C.F.S.)

Hemichordata

The Hemichordata are a group of wormlike marine invertebrates closely related to the chordates. The term Hemichordata—from the Greek *hemi*, meaning "half," and *chorde*, meaning "string," thus, "half-chordate"—was first proposed because the buccal diverticulum, a tubular outgrowth from the mouth cavity forward into the proboscis, or "snout," resembled a rudimentary notochord—the dorsal, or back side, supporting axis of the more primitive vertebrates. This theory has since been rejected, however, because it has been determined that the diverticulum bears little resemblance in origin and function to the vertebrate notochord. The hemichordates are sometimes given phylum rank and are so treated in this article; some authorities place the group at various other taxonomic levels with the other so-called protochordates. The Hemichordata consist of three classes:

From (A) Borradaile *et al.*, *The Invertebrata* (1963); Cambridg University Press, (B) P. Meglitsch, *Invertebrate Zoology* (1967); Oxford University Press, (C) C. Burdon-Jones



Body plans of representative Hemichordata.
(A) Acorn worm, *Dolichoglossus kowalewskii*, (B) *Planctosphaera*, (C) *Cephalodiscus*.

Enteropneusta, Pterobranchia, and Planctosphaeroidea. Enteropneusta, or acorn worms, are solitary, wormlike, bilaterally symmetrical animals, often brilliantly coloured. They are known as acorn worms because of the appearance of the proboscis and collar. Pterobranchia are minute, colonial, tube-building forms. Planctosphaeroidea are known only from a few floating larvae.

Size range and distribution

Enteropneusts are common in the intertidal zones from the White Sea and Greenland south to New Zealand and the Cape of Good Hope; they are found offshore to depths of 400 metres (about 1,300 feet) or more. They vary in size from a few centimetres long (*Saccoglossus pygmaeus* of the North Sea) to two metres (about seven feet) or more (*Balanoglossus gigas* of Brazilian coastal waters). Most species of the pterobranch genus *Cephalodiscus* inhabit offshore waters at depths of about 165 to 2,000 feet in the Antarctic and south temperate seas. *Rhabdopleura*, also a pterobranch, is cosmopolitan on various sea bottoms at depths of about 16 to 1,650 feet from the sub-Arctic to the sub-Antarctic.

Natural history. *Reproduction and life cycle.* Among some Enteropneusta, the fertilized egg develops into a free-swimming larva—called a tornaria larva—similar to the starfish larva. The tornaria is barely visible as it spirals continuously through the water, propelled by the powerful beat of the telotroch, a belt of cilia, or tiny hairlike structures. These larvae may drift for weeks, feeding on plankton (microscopic plants and animals). Such larvae are called planktotrophic to distinguish them from lecithotropic larvae (*e.g.*, the larvae of *Saccoglossus* and other members of the family Harrimanidae), which are nourished by their own yolk reserves.

The tornaria gut is L-shaped. A short gullet leads from the mouth to a globular stomach; a straight intestine passes from the stomach to the anus. A thin-walled triangular vesicle, or sac, in front of the mouth leads to the outside through a pore on the dorsal surface; the vesicle develops into the proboscis coelom, or cavity, and the pore becomes the proboscis pore of the adult. As the enteropneust larva grows older, two pairs of vesicles, which develop on either side of the stomach, become the collar cavity and the trunk cavity. The free-swimming larva of the enteropneustan *Saccoglossus* does not develop a mouth or begin to feed until after it has settled to the sea bottom. It then develops a tail, which it uses to attach itself, and builds a burrow. Initially, the larva feeds at the surface by straining particles from the water drawn in through the mouth. This method is superseded by one that first entails selecting a surface to which to attach itself and then burrowing or foraging on that surface with its proboscis. The enteropneust tail is a relic of the stalk of more primitive relatives, the pterobranchs.

Mouthless larva of *Saccoglossus*

Among enteropneusts, the sexes are separate and fertilization is external. Asexual reproduction is known in one species of *Balanoglossus;* small pieces separated from the posterior end regenerate into mature adults.

The lecithotropic larvae of the pterobranchs, *Cephalodiscus* and *Rhabdopleura*, develop within the parent coenoecium (*i.e.*, a nonliving coat that is secreted by glands in the proboscis) and disperse during their short free-swimming stage. They rapidly develop the tripartite arrangement of the adult (*i.e.*, proboscis, collar, and trunk). The anus, at first near the posterior end, moves forward near the mouth as the intestine becomes U-shaped (an adaptation to tube dwelling). After settling, the larva of *Rhabdopleura* forms a coenoecium. It develops into a colony by a progressive process of budding and segregation from a creeping stolon (a rootlike extension) that remains attached to the bottom.

Behaviour and ecology. In general, Enteropneusta live in U-shaped, coiled, or sinuous burrows excavated beneath stones or live among the attachments of sea plants. *Balanoglossus gigas* burrows 75 centimetres (about 30 inches) or more. Adult enteropneusts use the proboscis and collar for burrowing. The proboscis penetrates the sand or mud and by rhythmic contraction and distention draws the collar into the burrow. Alternate elongation and dilation of the collar provide an anchoring effect for drawing the trunk forward. Movement within the burrow is accomplished by rhythmical contractions of muscle and by the action of cilia. Mucus secreted by certain cells provides lubrication. Mucus may have a cleansing and protective function, for large amounts are produced when the animals are disturbed by handling. Despite the mucus, bottom-feeding fishes eat the worms when they stretch from their burrows to forage.

Sand, mud, and organic debris collected by enteropneusts during burrowing and foraging pass into the gut. Water drawn into the mouth provides oxygen and delivers fine food particles that are strained from the current as it passes out through the gill slits.

Spawning

In the Harrimanidae (*e.g.*, *Saccoglossus*), development from egg to adult is direct; there is no tornaria larva. Eggs and sperm, released within the burrows, are subsequently expelled into the water, where fertilization takes place. The reproductive process is initiated by tidal and climatic factors, particularly a rise in water temperature at the onset of spring tides. The females spawn first, followed shortly by the males. The eggs and young drift freely for three to four days.

The minute animals, or zooids, comprising a *Rhabdopleura* colony are restricted in movement because they

are attached to each other by a stolon. The zooids of *Cephalodiscus* are not so attached and thus can move in and out of their tubes to feed at or on the surface. They feed on particles of organic matter trapped by the cilia and mucus secretions on their tentacles. *Atubaria* has no coenoecium and lives free with its stalk entwined around other organisms on the sea floor.

Form and function. The adult hemichordate is distinguished by the division of its body and body cavities, or coeloms, into three basic parts: the proboscis, collar, and trunk. A central nervous system is absent, but there is a concentration of nerve tissue in the collar, which is linked with a nervous system in the epidermis, or outer covering. The circulatory system usually includes a contractile heartlike vesicle, blood vessels, and sinuses. The pharynx may be perforated by numerous paired gill slits, or they may be absent.

The second region of the body, the collar, may bear two or more tentacle-like plumes, which may have a double row of ciliated tentacles well supplied with secretory cells. The tentacles are special adaptations for feeding on particles suspended in the water. The network of nerve cells and fibres lying within the epidermis is linked with two main nerve tracts that lie dorsally median (*i.e.*, toward the body midline on the upper side) and ventrally median (on the lower side). The dorsal side of the collar has a neurochord formed by an inpocketing of the epidermis; it may have a central lumen, or cavity, that opens to the exterior anteriorly and posteriorly, or it may have a series of lacunae, or spaces. The neurochord contains large nerve cells, extensions of which reach almost to the tip of the proboscis and into the ventral nerve cord. These cells probably facilitate rapid responses such as abrupt contractions of the anterior trunk when the proboscis is touched. The general body surface is innervated by a primitive receptor system, which consists of scattered sensory cells. There is no well-defined centre of stimuli and responses.

Role of the glomerulus in excretion

Excretion may be effected by means of the glomerulus. This organ forms a thick mass of cells and blood-filled cavities on either side of the heart vesicle. In some enteropneusts the cells of the glomerulus contain yellow or brown granules, thus suggesting the possible excretory function for the glomerulus. The blood is driven from the heart vesicle into the glomerulus, where it is presumably cleansed of wastes before passing into the blood vessels and sinuses.

The enteropneusts have five coelomic cavities: one in the proboscis, which opens to the exterior by a small pore on the proboscis stalk; two in the collar, which lie side by side and are indirectly connected to the exterior through small ciliated pores that open on either side into the first pair of gill chambers; and two in the trunk, which also lie side by side but have no internal or external connections.

The pharynx of *Cephalodiscus* and *Atubaria* is perforated by a pair of gill passages connecting the gut with the exterior. In *Rhabdopleura* these are represented by a pair of grooves, a simplification that may be a consequence of the animal's minute size and its adaptation to a tube-dwelling habit.

The sexes are separate. The reproductive organs consist of a large number of saclike ovaries or testes arranged in a row along each side of the anterior part of the trunk, below the gill pores. In some species they are lodged in prominent genital ridges, or wings. Each sac opens to the surface by a simple pore, through which the eggs or sperm are liberated.

Knowledge of the third class of hemichordates, the Planctosphaeroidea, is based solely on the descriptions of a few larvae, transparent spheres about one centimetre (0.4 inch) in diameter in which the viscera can be seen clearly. Ciliary bands that wind over the surface are highly branched and bilaterally symmetrical, as in the free-swimming tornaria, which in every respect this larva resembles. The planctosphaera larva differs only in having a pair of blind, boot-shaped inpocketings on either side of the intestine and stomach. Since the tornaria bears no resemblance to the adult, it is assumed that the planctosphaera larva is equally different from its adult form.

Paleontology and classification. *Evolution.* The hemichordates have remained a general primitive group, at a low level of evolution, closely linked with the echinoderms (*e.g.*, starfish, sea urchins) and sharing with them and the other protochordate groups a sessile (*i.e.*, fixed to a surface) or semi-sessile ancestor, which was bilaterally symmetrical and had a tripartite body and coelom. The comparatively simple larval development of the Hemichordata suggests that they have deviated less from the ancestral stock than have echinoderms or protochordates.

Annotated classification. The members of the phylum differ to such a great extent that early workers had considerable difficulty in interrelating them.

PHYLUM HEMICHORDATA

Solitary or colonial, more or less wormlike body with 3 divisions; proboscis, collar, and trunk; epidermal nervous system; open circulatory system with a contractile heart vesicle, blood vessels and sinuses; glomerulus, with or without gill slits; with or without tentacles on collar; fewer than 100 species.

Class Enteropneusta (acorn worms)

So called because of the resemblance of the proboscis and collar to an acorn; solitary, wormlike, frequently brilliantly coloured: yellow, orange, red, violet, or green; live in burrows in sand or mud; numerous gill slits, a straight alimentary canal; collar without tentaculate plumes; a few centimetres to about 2 m (about 6.5 ft) long; worldwide distribution not common, but often abundant locally.

Family Harrimanidae

Proboscis long, cylindrical, and flexible, often many times longer than the collar; gill slits numerous dorsally or dorsolaterally placed, opening directly to exterior; proboscis skeleton with 2 short or medium-length side arms; development direct by way of a lecithotrophic larva where known; a few cm to 0.5 m (about 20 in.) in length; found in northern and southern seas; *Saccoglossus.*

Family Spengelidae

Proboscis generally short, a few times longer than collar; proboscis skeleton with long side arms; gill silts may extend over most of the pharynx; length, a few to 30 cm (about 1 ft), in genus *Willeyia* some may reach 2 m; found in northern and southern seas.

Family Ptychoderidae

Proboscis short, often ovoid, only a few times longer than collar; proboscis skeleton with very short side arms, gill silts confined to dorsal and dorso-lateral region of pharynx; pharynx strongly constricted into dorsal respiratory and ventral digestive parts; a few cm to more than 2 m long; widespread throughout Indo-Pacific and Atlantic waters; *Balanoglossus.*

Class Pterobranchia

Small hemichordates, with or without gill silts, 2 or more tentaculate arms or plumes on collar; live in aggregations or colonies housed in externally secreted encasement attached to or embedded in some sediment; aggregations develop by budding from 1 sexually produced zooid; genus *Cephalodiscus* mostly Sub-Antarctic or Antarctic, south temperate regions around Australia, New Zealand, and South Africa, tropical seas around Indonesia, Sri Lanka (formerly Ceylon), India, and Indochina: genus *Rhabdopleura* common throughout coastal waters of northern Europe, on coral reefs in Sri Lanka, and southward to latitude 66° S.

Class Planctosphaeroidea

Knowledge based upon descriptions of a few specimens of 1 species: *Planctosphaera pelagica;* transparent, spherical, bilaterally symmetrical, with sinuous branched ciliary bands radiating over surface; believed to be the larva of an unknown hemichordate.

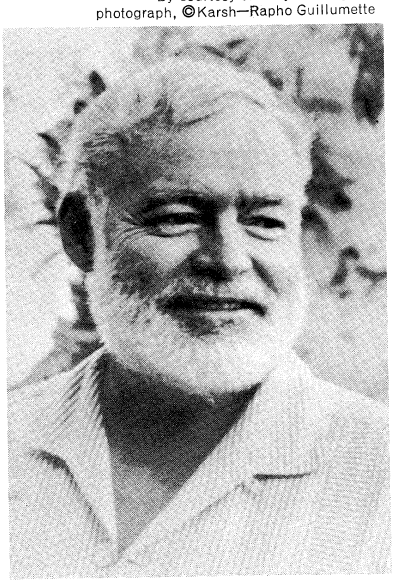
BIBLIOGRAPHY. L.H. HYMAN, *The Invertebrates*, vol. 5, *Smaller Coelomate Groups*, pp. 72–207 (1959), the most authoritative work in English to date; E.J.W. BARRINGTON, *The Biology of the Hemichordata and Protochordata* (1965), useful complementary reading to Hyman; C. BURDON-JONES, "Observations on the Spawning Behaviour of *Saccoglossus horsti.* Brambell & Goodhart, and of Other Enteropneusta," *J. Mar. Biol. Ass. U.K.*, 29:625–638 (1951); C.J. VAN DER HORST, "Hemichordata," in *Bronn's Klassen und Ordnungen des Tierreichs*, vol. 4 (1939), the best taxonomic and general account available (in German); "Adelochordate (Hemichordata)," in T.J. PARKER and W.A. HASWELL, *A Text-Book of Zoology*, vol. 2, 7th ed. rev. by A.J. MARSHALL (1963), a useful introduction.

(C.B.-J.)

Hemingway, Ernest

Ernest Hemingway, American novelist and short-story writer, who became notable for the intense masculinity of his writing as well as for his adventurous and widely publicized life, achieved a fame surpassed by few, if any, American authors of the 20th century. The virile nature of his writing, which attempted to re-create the exact physical sensations he experienced in big-game hunting, bullfighting, and warfare, in fact masked an aesthetic sensibility of great delicacy. A celebrity long before he reached middle age, his popularity has been validated by serious critical opinion.

By courtesy of Mary Hemingway; photograph, ©Karsh—Rapho Guillumette



Hemingway, 1959.

The first son of Clarence Edmonds Hemingway, a doctor of medicine devoted to hunting and fishing, and Grace Hall Hemingway, whose interests were artistic, Ernest Miller Hemingway was born on July 21, 1899, in Oak Park, Illinois, a suburb of Chicago. His life and career were to combine the divergent predilections of his parents, but the same life and career can be seen in large part as a never ended rebellion against the prudery and conventionalism of both father and mother and the place of his birth. He was educated in the public schools there and began to write in high school, where he was active and outstanding. But in his work he rejected Oak Park and what it stood for; as many of his short stories were later to indicate, the parts of his boyhood that mattered most were summers spent with his family on Walloon Lake in upper Michigan. On graduation from high school in 1917, impatient for a less sheltered environment, he did not enter college but went to Kansas City, where he was employed as a reporter for the *Star*, a leading newspaper of the time, which gave him valuable vocational training. Repeatedly rejected for military service because of a defective eye, he managed to enter World War I as an ambulance driver for the American Red Cross. On July 8, 1918, not yet 19 years old, he was injured on the Austro-Italian front at Fossalta di Piave. Decorated for heroism and hospitalized in Milan, he fell in love with a Red Cross nurse, Agnes von Kurowsky, who declined to marry him. These were experiences he was never to forget.

After recuperating at his home and in Michigan, Hemingway renewed his efforts at writing, for a while worked at odd jobs in Chicago, married Hadley Richardson, and sailed for France as a foreign correspondent for the *Toronto Star*. Advised and encouraged by other American writers in Paris—Scott Fitzgerald, Gertrude Stein, Ezra Pound—he began to see his nonjournalistic work appear in print there, and in 1925 his first important book, a collection of stories called *In Our Time*, was published in New York. The following year he published *The Sun Also Rises*, a novel with which he scored his first solid success. A pessimistic but sparkling book, it deals with a group of aimless expatriates in France and Spain—members of the postwar "lost generation," a phrase that Hemingway scorned while making it famous. This work also introduced him to the limelight, which he both craved and resented for the rest of his life.

First novels

The writing of books occupied him for most of the postwar years. Meanwhile, however, his first marriage, which produced a son, John, had failed, and he had married Pauline Pfeiffer, who was to bear his other two children, Patrick and Gregory. Based in Paris, he had also travelled widely for the skiing, bullfighting, fishing, or hunting that by then had become part of his life and formed the background for much of his writing. His position as a master of short fiction had been advanced by *Men Without Women* in 1927 and thoroughly established with *Winner Take Nothing* in 1933. At least in the public view, however, the novel *A Farewell to Arms* (1929) overshadowed both. Reaching back to his experience as a young soldier in Italy, he developed a grim but lyrical novel of great power, fusing love story with war story.

Many critics argue that by this stage Hemingway had done all of his best writing. The judgment is too harsh, but it is true that his success from this time on was mixed with unaccustomed failure and that his books were to appear less frequently. He once confessed that if he had spent less time hunting and fishing he might have written more. To these diversions might be added both his attraction to war, which was to absorb whole years of his life, and an energetic pursuit of personal enjoyment, for which he had an enormous capacity. Although he eventually drew on many of his adventures for books, he was at least as avid for the experience as for the literary use to which he might put it. His love of Spain and his passion for bullfighting resulted in *Death in the Afternoon* (1932), a learned study of a spectacle he saw more as tragic ceremony than as sport. Similarly, an African safari resulted in *Green Hills of Africa* (1935), an account of big-game hunting. Mostly for the fishing, he bought a house in Key West. Lured by the idea of catching the great marlin in the Gulf Stream of Cuba, he also bought his own fishing boat, "Pilar." A minor novel of 1937 called *To Have and Have Not* is set in and near Key West during the economic depression. Hemingway had now become concerned with social problems.

Activities during the Spanish Civil War

By now Spain was in the midst of civil war. Still deeply attached to that country, Hemingway made four trips there, once more a correspondent. He raised money for the Loyalists, who supported the government of the republic against the uprising of General Franco, and wrote a play called *The Fifth Column* (1938), set in besieged Madrid. As in many of his books, the protagonist of the play is based on the author, and his mistress on the writer and journalist Martha Gellhorn, whom Hemingway married when his divorce became final. Following his last visit to the Spanish war he purchased Finca Vigía (Lookout Farm), an unpretentious estate outside Havana, Cuba, and went with his bride to cover another war—the Japanese invasion of China.

The harvest of his considerable experience of Spain in war and peace was the novel *For Whom the Bell Tolls* (1940), the most successful of all his books as measured in sales. A substantial and often impressive work, it deals with a volunteer American guerrilla in the Spanish war who blows up a strategic bridge near Segovia in an attack that he knows is doomed to fail and in which he is left to die. All of his life Hemingway was fascinated by war—in *A Farewell to Arms* he focussed on its pointlessness, in *For Whom the Bell Tolls* on the comradeship it creates—and he was soon to become involved in World War II. For a long time he had bitterly predicted that the Spanish Civil War was a prelude to it. After he returned to Cuba he established what he called the Crook Factory, an informal but officially approved counterintelligence organization designed to deal with the influx of German spies in Cuba and with submarines off its coast. Also approved was a second plan, under which he outfitted his boat, "Pilar," in such a way as to attract U-boats, which he and his

crew might then destroy. He managed both operations efficiently, but before long he was disappointed at scoring no significant victories; and to get closer to the actual fighting he made his way to London, once again a journalist. He flew several missions with the Royal Air Force and crossed the English Channel with American troops on D-Day (June 6, 1944). Attaching himself to the 22nd Regiment of the 4th Infantry Division, he saw a good deal of action in Normandy and in the Battle of the Bulge. He also participated in the liberation of Paris and, although ostensibly a newsman, he impressed professional soldiers not only as a man of courage in battle but also as a real expert in military matters, guerrilla activities, and intelligence collection in particular.

Following the war in Europe, Hemingway returned to his home in Cuba and, his third marriage having gone the way of the others, he married his fourth wife—Mary Welsh, herself a correspondent whom he had met in London and with whom he was to spend the rest of his life. Established in Cuba at the Finca, where he began to work seriously again, they travelled widely, including one trip to Africa where they made a safari and were injured in two plane crashes. Soon after (in 1953), he received the Pulitzer Prize in fiction for a short, heroic novel—about an old Cuban fisherman who hooks and boats a giant marlin only to lose it to sharks—called *The Old Man and the Sea* (1952). This book, which played a role in gaining for him the Nobel Prize for Literature in 1954, was as enthusiastically praised as his previous novel, *Across the River and Into the Trees* (1950), the story of a professional army officer who dies while on leave in Venice, had been damned.

Pulitzer and Nobel prizes

By 1960 Fidel Castro's revolution in Cuba had driven Hemingway from his Finca. He had purchased another house in Ketchum, Idaho, and tried to lead his life and do his work as before. For a while he succeeded, but, anxiety-ridden and depressed, he was twice hospitalized at the Mayo Clinic in Rochester, Minnesota, where he received electroshock treatments. On July 2, 1961, two days after his return to the house in Ketchum, he took his life with a shotgun.

He left behind, however, a substantial amount of manuscript, some which has been published. *A Moveable Feast*, the memoir of his apprentice days in Paris, was issued in 1964. *Islands in the Stream*, three closely related novellas growing directly out of his peacetime memories of the Caribbean island of Bimini, of wartime Havana, and of searching for U-boats off Cuba, appeared in 1970.

A consummately contradictory man, Hemingway was witty, cheerful, irascible, by turns generous and selfish, expansive and egocentric. Hedonistic and dedicated, in love with life and yet by his own admission obsessed with death, an inveterate sportsman and omniverous reader, hard-drinking and early-rising, uninhibited, intricate, powerful, and damaged, he was himself a personification of the courage—which in a famous phrase he defined as "grace under pressure"—that mercilessly deserted him in the end.

MAJOR WORKS

NOVELS: *The Torrents of Spring* (1926); *The Sun Also Rises* (1926; English title, *Fiesta*, 1927); *A Farewell to Arms* (1929); *To Have and Have Not* (1937); *For Whom the Bell Tolls* (1940); *Across the River and Into the Trees* (1950); *The Old Man and the Sea* (1952); *Islands in the Stream* (1970).

SHORT STORIES: *In Our Time* (1925); *Men Without Women* (1927); *Winner Take Nothing* (1933).

OTHER WORKS: *Death in the Afternoon* (1932), on bullfighting; *The Green Hills of Africa* (1935), on big-game hunting; *The Fifth Column and the First Forty-Nine Stories* (1938), a play published with collected short stories; *A Moveable Feast* (1964), sketches of his early life in Paris.

BIBLIOGRAPHY. AUDRE HANNEMAN, *Ernest Hemingway: A Comprehensive Bibliography* (1967); PHILIP YOUNG and C.W. MANN, *The Hemingway Manuscripts: An Inventory* (1969), an exhaustive list of Hemingway manuscripts to be deposited in the Kennedy Library, Cambridge, Mass.; CARLOS BAKER, *Ernest Hemingway: A Life Story* (1969), the standard, authorized biography; *Hemingway: The Writers as Artist*, 3rd ed. (1963), and PHILIP YOUNG, *Ernest Hemingway: A Reconsideration*, rev. ed. (1966), two standard critical studies.

(P.Y.)

Henry II the Saint, Emperor

The last of the Saxon rulers of Germany, Henry II was canonized by Pope Eugenius III, more than 100 years after his death, in response to church-inspired legends. He was, in fact, far from saintly, but there is some truth in the legends concerning his religious character. Together with Henry III, he was the great architect of cooperation between church and state, following a policy inaugurated by Charlemagne and promoted by Otto the Great (German king 936–973). His canonization is sometimes justified on the grounds that he was a great representative of the medieval German priestly kings.

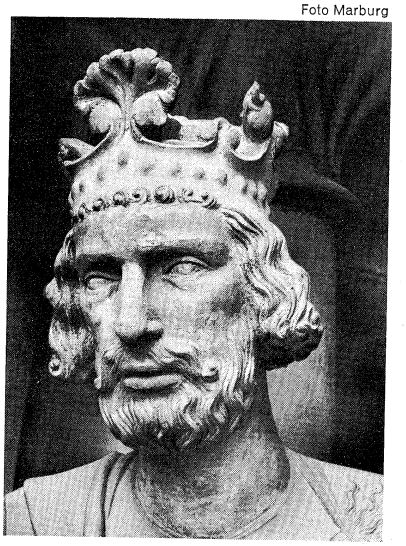
Born in Bavaria on May 6, 973, Henry II became king of Germany in 1002 and Roman Emperor in 1014. His father Henry, duke of Bavaria, having been in rebellion against two preceeding German kings, was forced to spend long years in exile from Bavaria. The younger Henry found refuge with Bishop Abraham of Freising and was later educated at the Cathedral School of Hildesheim. Exposed thus to strong church influence in his youth, religion influenced him strongly. Contemporaries observed an ironic trait in his character and were also impressed by his ability to intersperse his speeches with biblical quotations. Though devoted to church ritual and personal prayer, he was a tenacious and realistic politician, not adverse to alliances with heathen powers. Usually in poor health, he yet performed for 22 years the office of the itinerant king, riding on horseback through his dominions to judge and compose feuds, pursue rebels, and extend the power of the crown.

After the death of King Otto III in January 1002, Henry, aware of strong opposition to his succession, captured the royal insignia which were in the keeping of the dead king's companions. At Otto's funeral the majority of the princes declared against Henry, and only in June, with the assistance of Archbishop Willigis of Mainz, did Henry secure both election and coronation. It took another year before his recognition was final.

Succession to the throne

Henry turned his first attention to the east and made war against Polish king Bolesław I the Brave. After a successful campaign, he marched into northern Italy to subdue Arduin of Ivrea, who had styled himself king of Italy. His sudden interference led to bitter fighting and atrocities, and although Henry was crowned king in Pavia on May 15, 1004, he returned home without defeating Arduin, to pursue his campaigns against Bolesław. In 1003 Henry had made a pact with the Liutitian tribe against the Christian Bolesław, and allowed them to resist German missionaries east of the Elbe. Henry was more interested in political power than in spreading Christianity. Supported by his tribal allies he waged several campaigns against Poland, until in 1018, at Bautzen, he made a lasting compromise peace.

Foto Marburg



Henry II, detail from a statue, c. 1235. From a portal in the Bamberg Cathedral, Germany.

Sensitive to tradition and anxious to be crowned emperor, Henry decided in late 1013 on another expedition to Italy. He marched straight to Rome, where he was crowned by Pope Benedict VIII, on February 14, 1014. By May he was back in Germany, seeking to fulfill his duties to Italy by charging German officials with the administration of the country. He even convened an Italian imperial court at Strassburg in 1019. In 1020 Pope Benedict visited him in Germany and begged him to put in another appearance in Italy to fight the Greeks in the south and protect the papacy against the Lombard princes. Henry reluctantly responded the following year, fighting both Greeks and Lombards successfully; but he withdrew at the first opportunity.

Henry's main interest and success were concentrated on the consolidation of a peaceful royal regime in Germany. He spent much time and energy in elaborating the so-called Ottonian system of government. Inaugurated by Otto I, this system was based upon the principle that the lands and the authority of the bishops ought to be at the disposal of the king. Henry made generous grants to the bishops and, by adding to their territorial holdings, helped to establish them as secular rulers as well as ecclesiastical princes. He freely availed himself of the royal right to appoint faithful followers to these bishoprics. He insisted on episcopal celibacy—to make sure that on the death of a bishop the see would not fall into the hands of the bishop's children. In this way, he managed to create a stable body of supporters who made him more and more independent of rebellious nobles and ambitious members of his own family.

Foundation of the Bamberg bishopric

His greatest achievement was the foundation of the new bishopric of Bamberg. The upper region of the Main River was poorly populated, and Henry set aside large tracts of personal property to establish the new bishopric, much against the wishes of the bishop of Würzburg in the middle Main region. He obtained the consent of other bishops at a synod in Frankfurt in late 1007. The new bishop was consecrated on Henry's birthday in 1012. In 1020 Bamberg was visited by the pope, and it quickly developed into a splendid cathedral town where contemporary scholastic culture and art, as well as piety, found the support of Henry and his queen, Cunegunda.

During the last years of his reign Henry planned, in concert with Pope Benedict VIII, an ecclesiastical reform council at Pavia to seal the system of ecclesiastico-political order he had perfected in Germany. But he died suddenly in July 1024, before this could be done.

BIBLIOGRAPHY. G. BARRACLOUGH, *The Origins of Modern Germany*, 2nd rev. ed., ch. 3 (1947); and R. HOLTZMANN, *Geschichte der sächsischen Kaiserzeit, 900–1024*, 3rd ed., pp. 383–487 (1955), present Henry II according to accepted modern scholarly opinion. W.v.d. STEINEN, *Kaiser Heinrich der Zweite der Heilige* (1924), is romantic but unusually sensitive to later medieval opinion. T. SCHIEFFER, "Heinrich II und Konrad II," *Deutsches Archiv*, 8:384–437 (1951), supplies the necessary critical revision of modern research. For pictures, see P.E. SCHRAMM, *Die deutschen Kaiser und Könige in Bildern ihrer Zeit*, pp. 196 ff. (1928). Informative and summary treatments of Henry II may be found by A. DOVE in *Allgemeine deutsche Biographie*, vol. 11, pp. 376–384 (1880); by M.L. BULST-THIELE in B. GEBHARDT, *Handbuch der deutschen Geschichte*, 9th ed., vol. 1, pp. 284–298 (1970); and by H. APPELT in *Neue deutsche Biographie*, vol. 8, pp. 310–313 (1968).

(P.Mu.)

Henry III, Emperor

The German king and Holy Roman emperor Henry III was a powerful advocate of the Cluniac reform movement that sought to purify the Western Church in the 11th century. The last emperor who was able to dominate the papacy, he was troubled in the latter part of his reign by opposition to his rule both in Germany and Italy.

Youth and marriage

Henry was born on October 28, 1017, the son of the emperor Conrad II and Gisela of Swabia. He was more thoroughly trained for his office than almost any other crown prince before or after. With the Emperor's approval, Gisela had taken charge of his upbringing, and she saw to it that he was educated by a number of tutors and acquired an interest in literature.



Henry III between two abbots, miniature from his gospel, c. 1040. In the Universitätsbibliothek, Bremen, West Germany (Ms. b. 21).
By courtesy of the Universitatsbibliothek, Bremen, West Germany

In 1036 Henry married Gunhilda (Kunigunde), the young daughter of King Canute of England, Denmark, and Sweden. Because her father had died shortly before, the union with this frail and ailing girl brought with it no political advantages. She died in 1038, and the emperor Conrad died the following year.

His 22-year-old successor as German king resembled him in appearance. From his mother Henry inherited much, especially her strong inclination to piety and church services. His accession to the throne, unlike that of his two predecessors, did not lead to civic unrest, but his reign was burdensome from the beginning. Probably over questions of principle, the self-willed Emperor quarrelled with the aging Gisela during her last years.

He devoted his energies above all to the contemporary movement to bring an end to war among Christian princes, although his own policies were not always pacific. In possession of the duchies of Franconia, Bavaria, Swabia, and Carinthia, he had attempted to carry on his father's policy of supremacy in the east and, in fact, attained sovereignty over Bohemia and Moravia.

It may have been at this time that Henry, prematurely believing he had reached the zenith of his power, displayed openly, as if it were a matter of governmental policy, his leanings toward the clerical-reform party. Intending to re-create a theocratic age like that of Charlemagne, he failed to realize that this could be done only as long as the papacy was powerless.

Still a childless widower, he married Agnes, the daughter of William V of Aquitaine and Poitou, in 1043. The match must have been intended primarily to cement peace in the west and to assure imperial sovereignty over Burgundy and Italy, and Agnes' total devotion to the church reform advocated by the Cluniac monasteries probably confirmed Henry in his decision to take her for his wife. In November 1050 she bore him a son, who later became the emperor Henry IV. There followed another boy, Conrad, and three daughters. What Henry still lacked was the highest honour—his coronation as emperor at the hands of the pope.

When Henry reached Rome in 1046, three rivals were claiming the papacy. Henry wanted a pacified Italy, in which German supremacy was uncontested, and he wanted to receive the imperial crown from unsullied hands. He convoked a synod at Sutri, which, at his bidding, elected as the new pope a German, Suidger, bishop of Bamberg, who was inaugurated as Clement II. On the same day he crowned the imperial couple.

Control of the papacy

Rome became an imperial city, and the control over the church—*i.e.*, the decisive vote in future conclaves—passed into the hands of the German king. In succeeding

years Henry made use of this right to appoint a pope three more times. When the Normans were beginning their conquest of Calabria, Henry did not intervene to any extent in southern Italy; instead he left this problem to Pope Leo IX, who was defeated by the Normans.

Believing that the basis of his power was secure, the Emperor expected to be as successful with his internal projects as he had been in foreign affairs; but this was not to be the case. He could not carry out his ecclesiastical reforms in Germany or its neighbouring territories because he was virtually without friends among the clergy. He was increasingly opposed by the Scandinavian Church and by that of the Saxons. Also, he had to contend during most of his reign with Godfrey II, duke of Upper Lorraine, whom he repeatedly pardoned instead of disciplining.

There was unrest everywhere. In 1054–55, dukes Conrad of Bavaria and Welf III of Carinthia attempted to overthrow Henry's rule through a widely spread conspiracy, and only their demise saved him from great trouble. Conrad, who had fled to Hungary, managed to subvert that country to such an extent that German influence remained permanently weakened. Although resistance against him stiffened with time, Henry continued to rule with moderation. Perhaps because he was aware of a lessening of his powers, his actions became haphazard. Instead of holding on to duchies that he had inherited, he entrusted them to others; but he chose badly and seldom acted decisively against his disloyal feudatories. He no longer inspired fear in his opponents—the Saxon and south German lay nobility, the alliance between Lorraine and Tuscany, the increasingly independent papacy, and the adventure-seeking Normans.

Opponents of the Emperor's policy thought it was excessively indulgent toward the church and hostile toward the lay princes. Some of this criticism was voiced among the ranks of the ecclesiastical reformers. Matters had come to such an impasse that Henry no longer pleased anyone. His demands on the people to support his military strength were heavy from the beginning, and his revenues from inheritances and confiscations were also considerable. If the empire's basic wealth did not increase in his reign, it was because he used it to fulfill the demands of his clerical friends, even as he bestowed duchies on lay nobles in order to appease them. It is not surprising that, under these circumstances, he was compelled to find other sources of revenue by seeking credits, foreclosing mortgages, and looking after the interests of his treasury when conferring high imperial offices or church benefices. The abolition of simony (the sale of church offices) was difficult even for as high-principled a ruler as Henry, and, as a result, his enemies accused him of greed. According to some sources, Henry III was rumoured to have become "untrue to himself" and inaccessible to the common people in old age; he was reported to have refused to grant a judicial hearing to "the poor." In contrast, in the early years of his reign, he could not be praised enough for his zeal in the administration of justice.

Disintegration of the empire

His change of personality may have resulted from the blunders and failures of his rule. After 1046 this man, shaped partly by religious ideals and partly by the harsh realities of political life, saw all his gains being swept away: northeast Germany, Hungary, southern Italy, and Lorraine. Even the part of his work that he considered his very own, church reform, began to turn against him. A high priest among men, who did penance even while ruthlessly persecuting and even hanging heretics, Henry learned at the end of his days that clemency, goodness, and earthly justice do not necessarily benefit a prince.

On the other hand, it may have been a physical disease that changed Henry. In 1045 he was so tortured with illness that negotiations concerning the succession were begun. The bad tidings from all corners of the empire must have complicated his condition. In September 1056 he fell sick in his favourite residence, the imperial palace at Bodfeld near Goslar, and, having assured the succession of his son Henry, he died on October 5.

BIBLIOGRAPHY. There is no contemporary biography, English-language work, or detailed treatise on Henry III. ERNST STEINDORFF, *Jahrbücher des deutschen Reichs unter Heinrich III*, 2 vol. (1874–81), is still the only comprehensive treatment. See also *Die Urkunden Heinrichs III*, ed. by HARRY BRESSLAU and PAUL KEHR (1931), vol. 5 in the "Monumenta Germaniae historica Series"; PAUL KEHR, *Vier Kapital aus der Geschichte Kaiser Heinrichs III* (1931); GERHART LADNER, *Theologie und Politik vor dem Investiturstreit: Abendmahlstreit, Kirchenreform, Cluni und Heinrich III* (1936, reprinted 1968); ERNST MUELLER, *Das Itinerar Kaiser Heinrichs III, 1039 bis 1056* (1901); HEINRICH APPELT, "Heinrich III," *Neue deutsche Biographie*, vol. 8, pp. 313–315 (1969); CAROLINE M. RYLEY, *Cambridge Medieval History*, vol. 3, pp. 272–308 (1922).

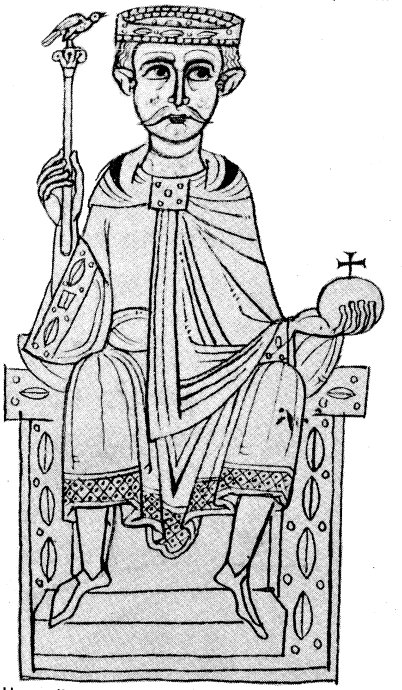
(H.L.M.)

Henry IV, Emperor

German king and Holy Roman emperor, Henry IV, in one of the most famous episodes in medieval history, engaged in a monumental contest of strength with the great reforming pope, Gregory VII. Gregory, who stood for the supremacy of the spiritual over the temporal order, sought to deprive the German kings of the right of investiture of ecclesiastical offices (*i.e.*, the right to confer on prelates the symbols of their spiritual authority). The power of the German crown, however, largely depended on this right, since it linked the high ecclesiastics to the crown as a counterweight against the territorial nobles, and Henry IV thus regarded retention of the right of investiture and the curtailment of the power of the nobles as his chief task.

Early years. The only surviving son of Emperor Henry III and Agnes of Poitou, he was born in Goslar (now in West Germany) on Nov. 11, 1050. Henry III, who had retained a firm hold on the church, had resolved a schism in Rome (1046), opening new activities for the reformers. At Easter 1051, the boy was baptized after the German princes had taken an oath of fidelity and obedience at Christmas 1050. On July 17, 1053, he was elected king at Tribur (modern Trebur, in West Germany) on condition that he would be a just king. In 1054 he was crowned king in Aix-la-Chapelle (modern Aachen, in West Germany), and the following year he became engaged to Bertha, daughter of the margrave of Turin. When the Emperor died in October 1056, at the age of 39, succession to the throne and survival of the dynasty were assured. The princes of the realm raised no objection when nominal government was handed over to the six-year-old boy, for whom his pious and un-

By courtesy of the Master and Fellows of Corpus Christi College, Cambridge; photograph, Courtauld Institute of Art, London



Henry IV, illumination from the manuscript *Ekkehardi historia*, c. 1113. In possession of Corpus Christi College, Cambridge.

worldly mother became regent. Yet the early death of Henry III was the beginning of a fateful change that marked all of his son's reign. In his will, the late Emperor had appointed Pope Victor II as counsellor to the Empress, and the Pope solved some of the conflicts between the princes and the imperial court that had endangered peace in the empire.

Regency

After Victor's early death (1057), however, the politically inept Empress committed a number of decisive mistakes. On her own, and without the benefit of the advice of a permanent group of counsellors, she readily yielded to various influences. She turned over the duchy of Bavaria, which Henry III had given to his son in 1055, to the Saxon count Otto of Nordheim, thus depriving the king of an important foundation of his power. She gave the duchy of Swabia to Count Rudolf of Rheinfelden—who married her daughter—and the duchy of Carinthia to Count Berthold of Zähringen; both of them eventually became opponents of Henry IV. The death of the Emperor also marked the disruption of German influence in Italy and of the close relationship between the king and the reform popes. Their independence soon became apparent in the elections of Stephen IX and Nicholas II, which were not influenced (as under Henry III) by the German court; in the new procedure for the election of the popes (1059); and in the defensive alliance with the Normans in southern Italy. This alliance was necessary for the popes as an effective protection against the Romans and was not directed against the German king. Yet the Normans were considered usurpers and enemies of the Holy Roman Empire, and thus the pact resulted in strained relations between the Pope and the German court, and these strains were aggravated by papal claims and disciplinary action taken by Nicholas II against German bishops. While the German king had so far been known as a supporter of the reformers, the Empress now imprudently entered into an alliance with Italian opponents of church reform and brought about the election of Cadalus, bishop of Parma, as antipope (Honorius II) against the reigning pope, Alexander II, who had been elected by the reformers. But since she did not give effective support to Honorius, Alexander was able to prevail. Her unwise church policy was matched by an obscurely motivated submissive policy at home, which, by unwarranted cession of holdings of the crown, weakened the material foundations of the king's power and, in addition, encouraged the rapacity of the nobles. Increasing discontent reached a climax in a conspiracy of the princes led by Anno, archbishop of Cologne, in April 1062. During a court assembly in Kaiserswerth he kidnapped the young king and had him brought to Cologne by ship. Henry's attempt to escape by jumping into the Rhine failed. Agnes resigned as regent and the government was taken over by Anno, who settled the conflict with the church by recognizing Alexander II (1064). Anno was, however, too dominating and inflexible a man to win Henry's confidence, so that Adalbert, archbishop of Bremen, granting more freedom to the lascivious young king, gained increasing and finally sole influence. But he used it for such unscrupulous personal enrichment that Henry, who was declared of age in 1065, had to ban him from court early in 1066. This incident marks the beginning of the King's own rule, for which he was badly prepared. Repeated changes in the government of the empire had an unsettling effect on the boy king and had, moreover, prevented him from being given a regular education. The selfishness of his tutors, the dissolute character of his companions, and the traumatic experience of his kidnapping had produced a lack of moral stability during his years of puberty. In addition, his love of power, typical of all the rulers of his dynasty, contributed to conduct often characterized by recklessness and indiscretion.

Henry's abduction by Anno

In 1069, after three years of marriage, he suddenly announced his intention of divorcing his wife, Bertha. Following protests by high church dignitaries, he dropped his plan but his mercurial behaviour incurred the displeasure of the reformers. At the same time he was faced with domestic difficulties that were to harass him throughout his reign. After his mother had freely dispensed of lands during her regency, he began to increase the royal possessions in the Harz Mountains and to protect them by castles, which he handed over to Swabian ministerials (higher civil servants directly responsible to the crown). Peasants and nobles in Saxony were stirred up by the ruthless repossession of former royal rights that had long ago been appropriated by nobility or had become obsolete and by the highhanded and severe measures of the foreign ministerials. Henry tried to stop the unrest by imprisoning Duke Magnus of Saxony and by depriving the widely respected Otto of Bavaria of his duchy, after having unjustly accused him of plotting the murder of the King (1070). Then a rebellion broke out among the Saxons, which in 1073 spread so rapidly that Henry had to escape to Worms. After negotiations with Welf IV, the new duke (as Welf I) of Bavaria, and with Rudolf, the duke of Swabia, Henry was forced to grant immunity to the rebels in 1073 and had to agree to the razing of the royal Harz Castle in the final peace treaty in February 1074. When the peasants, destroying the castle, also desecrated the church and the tomb of one of the King's sons, Henry declared the peace broken. This incident assured him of support from all over the empire, and in June 1075 he won an overwhelming victory that resulted in the surrender of the Saxons. It also forced the princes at Christmas to confirm on oath the succession of his one-year-old son Conrad.

Saxon wars

Role in investiture conflict. This rebellion affected relations between Henry and the Pope. In Milan a popular party, the Patarines, dedicated to reforming the city's corrupt higher clergy, elected its own archbishop, who was recognized by the Pope. When Henry countered by having his own nominee consecrated by the Lombard bishops, Alexander II excommunicated the bishops. Henry did not yield, and it was not until the Saxon rebellion that he was ready to negotiate. In 1073 he humbly asked the new pope, Gregory VII, to settle the Milan problem. The King having thus renounced his right of investiture, a Roman synod, called to strengthen the Patarine movement, forbade any lay investiture in Milan; henceforward Gregory regarded Henry as his ally in questions of church reform. When planning a crusade, he even put the defense of the Roman Church into the King's hands. But after defeating the Saxons, Henry considered himself strong enough to cancel his agreements with the Pope and to nominate his court chaplain as archbishop of Milan. The violation of the agreement on investiture called into question the King's trustworthiness, and the Pope sent him a letter warning him of the melancholy fate of King Saul (after breaking with his church in the person of the prophet Samuel) but offering negotiations on the investiture problem. Instead of accepting the offer, which arrived at his court on January 1, 1076, Henry, on the same day, deposed the Pope and persuaded an assembly of 26 bishops, hastily called to Worms, to refuse obedience to the Pope. By this impulsive reaction he turned the problem of investiture in Milan, which could have been solved by negotiations, into a fundamental dispute on the relations between church and state. Gregory replied by excommunicating Henry and absolving the King's subjects from their oaths of allegiance. Such action equalled dethronement. Many bishops who had taken part in the Worms assembly and had subsequently been excommunicated now surrendered to the Pope, and immediately the King was also faced with the newly aroused opposition of the nobility. In October 1076 the princes discussed the election of a new king in Tribur. It was only by promising to seek absolution from the ban within a year that Henry could reach a postponement of the election. The final decision was to be taken at an assembly to be called at Augsburg to which the Pope was also invited. But Henry secretly travelled to northern Italy and in Canossa did penance before Gregory VII, whereupon he was readmitted to the church. For the moment it was a political success for the King because the opposition had been deprived of all canonical arguments. Yet, Canossa meant a change. By doing penance Henry had admitted the legality of the

Penance at Canossa

Pope's measures and had given up the king's traditional position of authority equal or even superior to that of the church. The relations between church and state were changed forever.

The princes, however, considered Canossa a breach of the original agreement providing for an assembly at Augsburg and declared Henry dethroned. In his stead, they elected Rudolf, duke of Swabia, in March 1077, whereupon Henry confiscated the duchies of Bavaria and Swabia on behalf of the crown. He received support from the peasants and citizens of these duchies, whereas Rudolf relied mainly on the Saxons. Gregory watched the indecisive struggle between Henry and Rudolf for almost three years until he resolved to bring about a decision for the sake of continued church reform in Germany. At a synod in March 1080, he prohibited investiture, excommunicated and dethroned Henry again, and recognized Rudolf. The reasons for this act of excommunication were not as valid as those advanced in 1077, and many nobles who had so far favoured the Pope turned against him because they thought the prohibition of investiture infringed upon their rights as patrons of churches and monasteries. Henry now succeeded in deposing Gregory and in nominating Guibert, archbishop of Ravenna, as pope at a synod in Brixen (Bressanone). When the opposition of the princes was crippled by the death of Rudolf in October 1080, Henry, freed of the threat of enemies to the rear, went to Italy to seek a military decision in his struggle with the church. After attacking Rome in vain in 1081 and 1082, he conquered the city in March 1084. Guibert was enthroned as Clement III and crowned Henry emperor on March 31, 1084. Gregory, the legitimate pope, fled to Salerno, where he died on May 25, 1085. A number of cardinals joined Clement, and, feeling that he had won a complete victory, the Emperor returned to Germany. In May 1087 he had his son Conrad crowned king. The Saxons now made peace with him. Further, Henry replaced bishops who did not join Clement with others loyal to the king.

Crowned emperor

Later crises in Italy and Germany. The escape and death of Gregory VII and the presence of Clement III in Rome caused a crisis in the reform movement of the church, from which, however, it quickly recovered under the pontificate of Urban II (1088–1099). The marriage, arranged by Urban in 1089, of the 17-year-old Welf V of Bavaria with the 43-year-old countess Matilda of Tuscany, a zealous adherent of the cause of reform in the church, allied Henry's opponents in southern Germany and Italy. Henry was forced to invade Italy once more in 1090, but after initial success, his defeat in 1092 resulted in the uprisings in Lombardy; and the rebellion of his son Conrad, who was crowned king of Italy by the Lombards, led to general rebellion. The Emperor found himself cut off from Germany and besieged in a corner of northeastern Italy. In addition, his second wife Praxedis of Kiev, whom he had married in 1089 after the death of Bertha in 1087, left him, bringing serious charges against him. It was not until Welf V separated from Matilda, in 1095, and his father, the deposed Welf IV, was once more granted Bavaria as a fief, in 1096, that Henry was able to return to Germany (1097).

In Germany, sympathy for reform and the papacy no longer excluded loyalty to the Emperor. Gradually Henry was able to consolidate his authority so that in May 1098 the princes elected his second son, Henry V, king in place of the disloyal Conrad. But peace with the Pope, which was necessary for a complete consolidation of authority, was a goal that remained unattainable. At first a settlement was impossible because of Henry's support for Clement III, who had died in 1100. Paschal II (1099–1118), a follower of the reformist policies of Gregory VII, was unwilling to conclude an agreement with Henry. Finally, the Emperor declared that he would go on a crusade if his excommunication were removed. To prepare for the crusade, he forbade all feuds among the great nobles of the empire for four years (1103). But unrest started again when reconciliation with the church did not materialize and the nobles thought the Emperor was restricting their rights in favour of his son. Henry V feared a controversy with the princes. In alliance with Bavarian nobles he revolted against the Emperor in 1104 to secure his throne by sacrificing his father. The Emperor escaped to Cologne, but when he went to Mainz his son imprisoned him and on December 31, 1105, extorted his apparently voluntary abdication. Yet Henry IV was not yet prepared to give up. He fled to Liège and with the Lotharingians defeated Henry V's army near Visé on March 22, 1106. Henry IV suddenly died in Liège on August 7, 1106. His body was transferred to Speyer but remained there in an unconsecrated chapel before being buried in the family vault in 1111.

Judgment of Henry by his contemporaries differed according to the parties to which they belonged. His opponents considered the tall, handsome king a tyrant—the crafty head of heresy—whose death they cheered because it seemed to usher in a new age. His friends praised him as a pious, gentle, and intelligent ruler, a patron of the arts and sciences, who surrounded himself with religious scholars and who, in his sense of law and justice, was the embodiment of the ideal king. In his attempt to preserve the traditional rights of the crown, Henry IV was only partially successful, for while he strengthened the king's position against the nobles by gaining the support of the peasants, the citizens, and the ministerials, his continuing battles with the reforming church over investiture ultimately weakened royal influence over the papacy.

Assessment

BIBLIOGRAPHY. No critical biography has appeared. For the King's dispute with the Pope, see K.F. MORRISON, "Canossa: A Revision," *Traditio*, 18:121–148 (1962). All recent research and a full bibliography may be found in B. GEBHARDT and H. GRUNDMANN (eds.), *Handbuch der deutschen Geschichte*, 9th rev. ed., vol. 1 (1970). See also the *Cambridge Medieval History*, vol. 5 (1964).

(F.-J.Se.)

Henry V, Emperor

German king and Holy Roman emperor, last of the Salian dynasty, Henry V settled the controversy between the German kings and the papacy over the right to invest bishops with their offices begun by his father, Henry IV.

The struggle with the church

He was born in Utrecht on November 8, 1086, the second son of Henry IV and his first wife, Bertha of Turin. After his father became emperor, Henry's elder brother, Conrad, was elected German king; Henry succeeded him

By courtesy of the Master and Fellows of Corpus Christi College, Cambridge; photograph, Courtauld Institute of Art, London



Henry V (left) receiving the imperial insignia from Pope Paschal II at Rome, April 13, 1111, illumination from a German manuscript, *c.* 1114–25. In the possession of Corpus Christi College, Cambridge.

after Conrad had rebelled unsuccessfully against his father, being crowned on January 6, 1099. In 1104, in the conflict between the papacy and his father, he sided with the Bavarians and Saxons against his father. As a promoter of church reform willing to compromise with the papacy, he had the support of the church. He took his father prisoner and forced him to abdicate (December 31, 1105) but was not certain of his throne until his father's death on August 7, 1106. He had already sent messengers to Pope Paschal II inviting him to come to Germany; he was prepared to reach a settlement provided the Pope granted him full rights of investiture of bishops. The Pope rejected this condition. Henry was still able to consolidate his rule in Germany. Campaigns against Hungary (1108) and Poland (1109) failed, but Henry reasserted German lordship over Bohemia in 1110. In 1110 he became betrothed to Matilda, daughter of Henry I of England, marrying her in 1114.

An understanding with the Pope in the controversy over investiture was essential to Henry. The church possessed not only spiritual rights but secular rights as well. Henry journeyed to Rome in 1110 and again demanded the right of investiture. The Pope was willing to command the German churches to give back all lands and rights received from the crown if Henry would renounce the right to investiture, a bargain that was acceptable to Henry but not to the German bishops and princes. Henry then imprisoned the Pope, forcing him to grant the right of investiture. On April 13, 1111, the Pope crowned him emperor in St. Peter's. In the satisfaction that he had achieved what Henry IV had not, he arranged a memorial ceremony for his father in Speyer on August 7, 1111.

In Germany, Henry V followed his father's policy of favouring the class of civil servants known as *ministeriales* and also the towns, thus provoking the antagonism of the princes. Rebellion soon broke out; Archbishop Adalbert of Mainz fomented unrest in the upper Rhineland, and the revolt of Lothair of Supplinburg (later to become king as Lothair III and emperor as Lothair II) in Saxony ended in 1115 in a severe defeat for Henry.

There was also strong opposition to Henry within the church. While the Pope kept to his agreement with Henry, a council in Rome declared the privilege granted to Henry invalid. Papal legates in Germany pronounced Henry's excommunication, and consequently he lost the support of the German bishops. Despite this, he went to Italy in 1116 to take possession of the inheritance of Matilda of Tuscany, who had died in 1115. Further negotiations with the Curia over the investiture question were without success. When, in 1118, Pope Gelasius II was elected successor to Paschal II, Henry set up an antipope, Gregory VIII; but the move failed. Henry was called back from Italy in 1118 by an ultimatum from the German princes, who threatened to dethrone him. He was forced to make political concessions. When Gelasius II's successor, Calixtus II, offered to negotiate with him, Henry was prepared to drop his demand for full rights of investiture, but these negotiations failed. As his domestic difficulties increased, the princes finally took the initiative and negotiated the Concordat of Worms (1122). The King had to renounce the right to invest the bishops with ring and crozier and to accede to their canonical election, while the Pope granted the King the right to be present at the election, the right to a deciding voice if the election was indecisive, and the right to enfeoff the elected bishop with the temporalities of his see. This arrangement, however, applied only to Germany, whereas in Italy and in Burgundy the enfeoffment was to follow consecration and would therefore be a pure formality.

The deciding voice of the princes

Henry's subsequent struggle with the princes and, especially, with Lothair was without success. At the same time he became involved in the conflict between the English and the French. The death of the successor to the English throne had made Matilda, Henry's wife, the heiress and created the prospect of a German–English empire. Henry therefore supported his father-in-law in his conflict with France but could achieve nothing militarily. He died of cancer at Utrecht, in the Netherlands, on May 23, 1125, and was buried in Speyer Cathedral. Henry was childless. His successor was his former enemy Lothair III, duke of Saxony, who was elected king largely through the efforts of the church.

As a ruler, Henry V showed political skill, but his reach exceeded his grasp. He had dethroned his father by allying himself with the princes and presenting himself as a champion of the church's rights. Once in power, he took up his father's cause but was unable to force the church to grant him his demands. The settlement of 1122, which secured the King's influence over the German church, was brought about mainly by the German princes. By intervening in the conflict between the King and the church, they won a victory for themselves against the King, a fact that dominated the subsequent history of Germany.

BIBLIOGRAPHY. There is no really critical biography of Henry V. A summary of recent research and a full bibliography may be found in B. GEBHARDT and H. GRUNDMANN, *Handbuch der deutschen Geschichte*, vol. 1 (1970). See also *The Cambridge Medieval History*, vol. 5 (1964).

(F.-J.S.)

Henry I of England

The youngest and ablest of William the Conqueror's sons, and king of England from 1100 until his death in 1135, Henry I greatly strengthened the crown's executive powers and reunited the duchy of Normandy with the English kingdom.

By courtesy of the trustees of the British Museum



Henry I, miniature from a 14th-century manuscript. In the British Museum (Cottonian Claud D11 45 B).

Henry was born in England in 1069. He was crowned at Westminster, on August 5, 1100, three days after his brother, King William II, William the Conqueror's second son, had been killed in a hunting accident. Duke Robert Curthose, the eldest of the three brothers, who by feudal custom had succeeded to his father's inheritance, Normandy, was returning from the First Crusade and could not assert his own claim to the English throne until the following year. The succession was precarious, however, because a number of wealthy Anglo-Norman barons supported Duke Robert, and Henry moved quickly to gain all the backing he could. He issued an ingenious Charter of Liberties which purported to end capricious taxes, confiscations of church revenues, and other abuses of his predecessor. By his marriage with Matilda, a Scottish princess of the old Anglo-Saxon royal line, he established the foundations for peaceable relations with the Scots and support from the English. And he recalled St. Anselm, the scholarly archbishop of Canterbury whom his brother, William II, had banished.

When Robert Curthose finally invaded England in 1101 several of the greatest barons defected to him. But Henry, supported by a number of his barons, most of the Anglo-Saxons, and St. Anselm, worked out an amicable settlement with the invaders. Robert relinquished his claim to England, receiving in return Henry's own territories in Normandy and a large annuity.

Although a crusading hero, Robert was a self-indulgent, vacillating ruler who allowed Normandy to slip into chaos. Norman churchmen who fled to England urged Henry to conquer and pacify the duchy and thus pro-

vided moral grounds for Henry's ambition to reunify his father's realm at his brother's expense. Paving his way with bribes to Norman barons and agreements with neighbouring princes, in 1106 Henry routed Robert's army at Tinchebrai in southwestern Normandy. Robert was captured and remained Henry's prisoner for the rest of his life.

Church–state relations

Between 1104 and 1106 Henry had been in the uncomfortable position of posing, in Normandy, as a champion of the church while fighting with his own archbishop of Canterbury. St. Anselm had returned from exile in 1100 dedicated to reforms of Pope Paschal II, which were designed to make the church independent of secular sovereigns. Following papal bans against lay lords investing churchmen with their lands and against churchmen rendering homage to laymen, Anselm refused to consecrate bishops whom Henry had invested and declined to do homage to Henry himself. Henry regarded bishoprics and abbeys not only as spiritual offices but as great sources of wealth. Since in many cases they owed the crown military services, he was anxious to maintain the feudal bond between the bishops and the crown.

Ultimately, the issues of ecclesiastical homage and lay investiture forced Anselm into a second exile. After numerous letters and threats between king, pope, and archbishop, a compromise was concluded shortly before the battle of Tinchebrai and was ratified in London in 1107. Henry relinquished his right to invest churchmen while Anselm submitted on the question of homage. With the London settlement and the English victory at Tinchebrai, the Anglo-Norman state was reunified and at peace.

In the years following, Henry married his daughter Matilda (also called Maud) to Emperor Henry V of Germany and groomed his only legitimate son, William, as his successor. Henry's right to Normandy was challenged by William Clito, son of the captive Robert Curthose, and Henry was obliged to repel two major assaults against eastern Normandy by William Clito's supporters: Louis VI of France, Count Fulk of Anjou, and the restless Norman barons who detested Henry's ubiquitous officials and high taxes. By 1120, however, the barons had submitted, Henry's son had married into the Angevin house, and Louis VI—defeated in battle—had concluded a definitive peace.

The settlement was shattered in November 1120, when Henry's son perished in a shipwreck of the "White Ship," destroying Henry's succession plans. After Queen Matilda's death in 1118, he married Adelaide of Louvain in 1121, but this union proved childless. On Emperor Henry V's death in 1125, Henry summoned the empress Matilda back to England and made his barons do homage to her as his heir. In 1128 Matilda married Geoffrey Plantagenet, heir to the county of Anjou, and in 1133 she bore him her first son, the future King Henry II. On December 1, 1135, Henry I died at Lyons-la-Forêt in eastern Normandy, whereupon his favorite nephew, Stephen of Blois, disregarding Matilda's right of succession, seized the English throne. Matilda's subsequent invasion of England unleashed a bitter civil war that ended with King Stephen's death and Henry II's unopposed accession in 1154.

Assessment

Henry I was a skillful, intelligent monarch who achieved peace in England, relative stability in Normandy, and notable administrative advances on both sides of the Channel. Under Henry, the Anglo-Norman state his father had created was reunited. Royal justices began making systematic tours of the English shires, but although his administrative policies were highly efficient, they were not infrequently regarded as oppressive. His reign marked a significant advance from the informal, personal monarchy of former times toward the bureaucratized state that lay in the future. It also marked a shift from the wide-ranging imperialism of earlier Norman leaders to consolidation and internal development. In the generations before Henry's accession, Norman dukes, magnates, and adventurers had conquered Southern Italy, Sicily, Antioch, and England. Henry won his major battles but preferred diplomacy or bribery to the risks of the battlefield. Subduing Normandy in 1106, he contented himself with keeping domestic peace, defending his Anglo-Norman state against rebellion and invasion, and making alliances with neighbouring princes.

BIBLIOGRAPHY. There exists no biography of Henry I. A.L. POOLE, *From Domesday Book to Magna Carta, 1087–1216*, 2nd ed. (1955), contains a good sketch and bibliography of the reign. R.W. SOUTHERN, *Medieval Humanism*, ch. 11 (1970), deals perceptively with Henry's patronage system. On Henry's early years, see C.W. DAVID, *Robert Curthose, Duke of Normandy* (1920); on Henry's administration, H.G. RICHARDSON and G.O. SAYLES, *The Governance of Mediaeval England from the Conquest to Magna Carta* (1963); C.H. HASKINS, *Norman Institutions* (1960); and *Regesta Regum Anglo-Normannorum*, vol 2, ed. by H.A. CRONNE and CHARLES JOHNSON (1956).

(C.W.Ho.)

Henry II of England

Henry II, king of England from 1154 to 1189, a man of great intelligence and ability, increased by war and diplomacy the dominions that he inherited and controlled finally continental territories wider than those of any other English medieval monarch. His greatest achievement was to strengthen the financial administration of the country, to rationalize tenurial and judicial procedure, and thus to control crime, to ensure peaceful possession of land, and to establish regular and equitable courts applying customary law.

Early life. Henry was born at Le Mans, France, in 1133. The claim of his mother, Matilda, daughter of Henry I, to the English crown had been set aside by her cousin, King Stephen. After receiving a good literary education, part of it in England, Henry became duke of Normandy in 1150, and count of Anjou on the death of his father, Geoffrey Plantagenet, in 1151. In 1152 he advanced his fortunes by marrying the beautiful and talented Eleanor, recently divorced from King Louis VII of France, who brought with her hand the lordship of Aquitaine. Henry invaded England in 1153, and King Stephen agreed to accept him as coadjutor and heir. When Stephen died the following year Henry succeeded without opposition, thus becoming lord of territories stretching from Scotland to the Pyrenees.

Accession in 1154

The young king lacked visible majesty. Of stocky build, with freckled face, close-cut tawny hair, and gray eyes, he dressed carelessly and grew to be bulky; but his personality commanded attention and drew men to his service. He could be a good companion, with ready repartee in a jostling crowd, but he displayed at times the ungovernable temper of a furious animal and could be heartless and ruthless when necessary. Restless, impetuous, always on the move, regardless of the convenience of others, he was at ease with scholars, and his administrative decrees were the work of a cool realist. In his long reign of 34 years he spent an aggregate of only 14 in England.

Reign. His career may be considered in three aspects: the defense and enlargement of his dominions, the involvement in two lengthy and disastrous personal quarrels, and his lasting administrative and judicial reforms.

By courtesy of the trustees of the British Museum



Henry II (left) disputing with Becket (centre), miniature from a 14th-century manuscript. In the British Museum (Cotton MS. Claudius D.ii).

Dominions. His territories are often called the Angevin Empire. This is a misnomer, for Henry's sovereignty rested upon various titles, and there was no institutional or legal bond between different regions. Some, indeed, were under the feudal overlordship of the king of France. By conquest, through diplomacy, and the marriages of two of his sons, he gained acknowledged possession of what is now the west of France from the northernmost part of Normandy to the Pyrenees, near Carcassonne. During his reign, the dynastic marriages of three daughters gave him political influence in Germany, Castile, and Sicily. His continental dominions brought him into contact with Louis VII of France, the German emperor Frederick I (Barbarossa), and, for much of the reign, Pope Alexander III. With Louis the relationship was ambiguous. Henry had taken Louis's former wife and her rich heritage. He subsequently acquired the Vexin in Normandy by the premature marriage of his son Henry to Louis's daughter, and during much of his reign he was attempting to outfight or outwit the French king who, for his part, gave shelter and comfort to Henry's enemy, Thomas Becket, the archbishop of Canterbury. The feud with Louis implied friendly relations with Germany, where Henry was helped by his mother's first marriage to the emperor Henry V but hindered by Frederick's maintenance of an antipope, the outcome of a disputed papal election in 1159. Louis supported Alexander III, whose case was strong, and Henry became arbiter of European opinion. Though acknowledging Alexander, he continued throughout the Becket controversy to threaten transferrence of allegiance to Frederick's antipope, thus impeding Alexander's freedom of action.

Early in his reign Henry obtained from Malcolm III of Scotland homage and the restoration of Northumberland, Cumberland, and Westmorland; and later in the reign (1174) homage was exacted from William the Lion, Malcolm's brother and successor. In 1157 Henry invaded Wales and received homage, though without conquest. In Ireland, reputedly bestowed upon him by Pope Adrian IV, Henry allowed a expedition of barons from South Wales to establish Anglo-Norman supremacy in Leinster (1169), which the King himself extended in 1171.

Personal quarrels. His remarkable achievements were impaired, however, by the stresses caused by a dispute with Becket and by discords in his own family.

The quarrel with Becket, Henry's trusted and successful chancellor (1154–62), broke out soon after Becket's election to the archbishopric of Canterbury (May 1162, see BECKET, THOMAS). It led to a complete severance of relations and to the Archbishop's voluntary exile. Besides disrupting the public life of the church, this situation embroiled Henry with Louis VII and Alexander III; and though it seemingly did little to hamper Henry's activities, the time and service spent in negotiations and embassies was considerable, and the tragic dénouement in Becket's murder earned for Henry a good deal of damaging opprobrium.

More dangerous were the domestic quarrels, which thwarted Henry's plans and even endangered his life and which finally brought him down in sorrow and shame.

Family relations

Throughout his adult life Henry's sexual morality was lax; but his relations with Eleanor, 11 years his senior, were for long tolerably harmonious, and, between 1153 and 1167, she bore him eight children. Of these, the four sons who survived infancy—Henry, Geoffrey, Richard, and John—repaid his genuine affection with resentment toward their father and discord among themselves. None was blameless, but the cause of the quarrels was principally Henry's policy of dividing his dominions among his sons while reserving real authority for himself. In 1170 he crowned his eldest son, Henry, as coregent with himself; but in fact the young king had no powers and resented his nonentity, and in 1173 he opposed his father's proposal to find territories for the favoured John (Lackland) at the expense of Geoffrey. Richard joined the protest of the others and was supported by Eleanor. There was a general revolt of the baronage in England and Normandy, supported by Louis VII in France and William the Lion in Scotland. Henry's prestige was at a low ebb after the murder of Becket and recent taxation, but he reacted energetically, settled matters in Normandy and Brittany, and crossed to England, where fighting had continued for a year. On July 12, 1174, he did public penance at Canterbury. The next day the king of Scots was taken at Alnwick, and three weeks later Henry had suppressed the rebellion in England. His sons were pardoned, but Eleanor was kept in custody for 11 years.

A second rebellion flared up in 1181 with a quarrel between his sons Henry and Richard over the government of Aquitaine, but young Henry died in 1183. In 1184 Richard quarrelled with John, who had been ordered to take Aquitaine off his hands. Matters were eased by the death of Geoffrey (1186), but the King's attempt to find an inheritance for John led to a coalition against him of Richard and Philip II Augustus, the young king of France. Henry was defeated and forced to give way, and news that John also had joined his enemies hastened the King's death near Tours in 1189 (July 6).

Reforms. In striking contrast to the checkered pattern of Henry's wars and schemes, his governance of England displays a careful and successful adaptation of means to a single end—the control of a realm served by the best administration in Europe. This success was obscured for contemporaries and later historians by the varied and often dramatic interest of political and personal events; and, not until the 19th century, when the study of the public records began and when legal history was illuminated by the British jurist Frederic William Maitland and his followers, did the administrative genius of Henry and his servants appear in its true light.

At the beginning of his reign Henry found England in disorder, with royal authority ruined by civil war and the violence of feudal magnates. His first task was to crush the unruly elements and restore firm government, using the existing institutions of government, with which the Anglo-Norman monarchy was well provided. Among these was the King's council of barons, with its inner group of ministers who were both judges and accountants and who sat at the Exchequer, into which the taxes and dues of the shires were paid by the King's local representative the sheriff (shire-reeve). The council contained an unusually able group of men—some of them were great barons, such as Richard de Lucy and Robert de Beaumont, earl of Leicester; others included civil servants, such as Nigel, bishop of Ely, Richard Fitzneale and his son, Richard of Ilchester. Henry took a personal interest in the technique of the Exchequer, which was described at length for posterity in the celebrated *Dialogus de scaccario*, whose composition seemed to Maitland "one of the most wonderful things of Henry's wonderful reign." How far these royal servants were responsible for the innovations of the reign cannot be known, though the development in practice continued steadily, even during the King's long absences abroad.

In the early months of the reign the King, using his energetic and versatile chancellor Becket, beat down the recalcitrant barons and their castles and began to restore order to the country and to the various forms of justice. It was thus, a few years later, that he came into conflict with the bishops, then led by Becket, over the alleged right of clerks to be tried for crime by an ecclesiastical court. A result of this was the celebrated collection of decrees—the Constitutions of Clarendon (1164)—which professed to reassert the ancestral rights of the King over the church in such matters as clerical immunity, appointment of bishops, custody of vacant sees, excommunication, and appeals to Rome. The Archbishop, after an initial compliance, refused to accept these, and they were throughout the controversy a block to an agreement. The quarrel touched what was to be the King's chief concern—the country's judicial system.

The judicial system

Anglo-Saxon England had two courts of justice—that of the hundred, a division of the shire, for petty offenses, and that of the shire, presided over by the sheriff. The feudal regime introduced by the Normans added courts of the manor and of the honour (a complex of estates). Above all stood the royal right to set up courts for important pleas and to hear, either in person or through

his ministers, any appeal. Arrest was a local responsibility, usually hard upon a flagrant crime. A doubt of guilt was settled by ordeal by battle; the accused in the shire underwent tests held to reveal God's judgment. Two developments had come in since William the Conqueror's day: the occasional mission of royal justices into the shires and the occasional use of a jury of local notables as fact finders in cases of land tenure.

Henry's first comprehensive program was the Assize of Clarendon (1166), in which the procedure of criminal justice was established; 12 "lawful" men of every hundred, and four of every village, acting as a "jury of presentment," were bound to declare on oath whether any local man was a robber or murderer. Trial of those accused was reserved to the King's justices, and prisons for those awaiting trial were to be erected at the King's expense. This provided a system of criminal investigation for the whole country, with a reasonable verdict probable because the firm accusation of the jury entailed exile even if the ordeal acquitted the accused. In feudal courts the trial by battle could be avoided by the establishment of a concord, or fine. This system presupposed regular visits by the King's justices on circuit (or, in the technical phrase, "on eyre"), and these tours became part of the administration of the country. The justices formed three groups: one on tour, one "on the bench" at Westminster, and one with the King when the court was out of London. Those at Westminster dealt with private pleas and cases sent up from the justices on eyre.

Equally effective were the "possessory assizes." In the feudal world, especially in times of turmoil, violent ejections and usurpations were common, with consequent vendettas and violence. Pleas brought to feudal courts could be delayed or altogether frustrated. As a remedy Henry established the possessory writ, an order from the Exchequer, directing the sheriff to convene a sworn local jury at petty assize to establish the fact of dispossession, whereupon the sheriff had to reinstate the defendant pending a subsequent trial at the grand assize to establish the rights of the case. This was the writ of Novel Disseisin (*i.e.*, recent dispossession). This writ was returnable; if the sheriff failed to achieve reinstatement, he had to summon the defendant to appear before the King's justices and himself be present with the writ. A similar writ of Mort d'Ancestor decided whether the ancestor of a plaintiff had in fact possessed the estate, whereas that of Darrein Presentment (*i.e.*, last presentation) decided who in fact had last presented a parson to a particular benefice. All these writs gave rapid and clear verdicts subject to later revision. The fees enriched the treasury, and recourse to the courts both extended the King's control and discouraged irregular self-help. Two other practices developed by Henry became permanent. One was scutage, the commutation of military service for a money payment; the other was the obligation, put on all free men with a property qualification by the Assize of Arms (1181), to possess arms suitable to their station.

The ministers who engaged upon these reforms took a fully professional interest in the business they handled, as may be seen in Fitzneale's writing on the Exchequer and that of the chief justiciar, Ranulf de Glanville, on the laws of England, and many of the expedients adopted by the King may have been suggested by them. In any case, the long-term results were very great. By the multiplication of a class of experts in finance and law Henry did much to establish two great professions, and the location of a permanent court at Westminster and the character of its business settled for England (and for much of the English-speaking world) that common law, not Roman law, would rule the courts and that London, and not an academy, would be its principal nursery. Moreover, Henry's decrees ensured that the judge-and-jury combination would become normal and that the jury would gradually supplant ordeal and battle as being responsible for the verdict. Finally, the increasing use of scutage, and the availability of the royal courts for private suits, were effective agents in molding the feudal monarchy into a monarchical bureaucracy before the appearance of Parliament.

Results of Henry's measures

Significance. Henry II lived in an age of biographers and letter writers of genius. John of Salisbury, Thomas Becket, Giraldus Cambrensis, Walter Map, Peter of Blois, and others knew him well and left their impressions. All agreed on his outstanding ability and striking personality and also recorded his errors and aspects of his character that appear contradictory, whereas modern historians agree upon the difficulty of reconciling its main features. Without deep religious or moral conviction, Henry nevertheless was respected by three contemporary saints, Aelred of Rievaulx, Gilbert of Sempringham, and Hugh of Lincoln. Normally an approachable and faithful friend and master, he could behave with unreasonable inhumanity. His conduct and aims were always self-centred, but he was neither a tyrant nor an odious egoist. Both as man and ruler he lacked the stamp of greatness that marked Alfred the Great and William the Conqueror. He seemed also to lack wisdom and serenity; and he had no comprehensive view of the country's interest, no ideals of kingship, no sympathetic care for his people. But if his reign is to be judged by its consequences for England, it undoubtedly stands high in importance, and Henry, as its mainspring, appears among the most notable of English kings.

BIBLIOGRAPHY. W.L. WARREN'S *Henry II* (1973) is the one full biography (with bibliography). The best short accounts are still those of KATE NORGATE in the *Dictionary of National Biography*, vol. 26 (1891); and DORIS M. STENTON in the *Cambridge Medieval History*, vol. 5, ch. 17 (1929), both with full bibliographies. The classical essay by WILLIAM STUBBS, his introduction to the *Gesta Henrici* ("Rolls Series," 1867), was reprinted by A.H. HASSALL in his collection of *Historical Introductions to the Rolls Series*, pp. 89–172 (1902). Many contemporary sources are translated in D.C. DOUGLAS and G.W. GREENAWAY (eds.), *English Historical Documents II* (1952), including the whole of the *Dialogue of the Exchequer* (*Dialogus de Scaccario*), of which the best edition, with translation, is that by CHARLES JOHNSON (1950). For Henry's judicial reforms, the best account is still that in F. POLLOCK and F.W. MAITLAND, *The History of English Law Before the Time of Edward I*, 2nd ed. (1898). See also D.M. STENTON, *English Justice Between the Norman Conquest and the Great Charter, 1066–1215* (1965).

(M.D.K.)

Henry V of England

In his nine-year reign (1413–22), King Henry V of England renewed the Hundred Years' War (*q.v.*), reduced France to his mercy, and made England the strongest kingdom in Europe. The hero of the Battle of Agincourt (1415), Henry was popular as a warrior and among the ablest of England's medieval monarchs. Henry, eldest son of Henry, earl of Derby (afterward Henry IV), by Mary de Bohun, was born at Monmouth in 1387, probably on September 16. On his father's exile in 1398, Richard II took the boy into his own charge, treated him kindly, and knighted him in 1399. Henry's uncle, Henry Beaufort, bishop of Winchester, seems to have been responsible for his training, and despite his early entry into public life, he was well educated by the standards of his time. He grew up fond of music and reading and became the first English king who could both read and write with ease in the vernacular tongue. On October 15, 1399, after his father had become king, Henry was created earl of Chester, duke of Cornwall, and prince of Wales, and soon afterward, duke of Aquitaine and Lancaster. From October 1400 the administration of Wales was conducted in his name, and in 1403 he took over actual command of the war against the Welsh rebels, a struggle that absorbed much of his restless energy until 1408. Thereafter he began to demand a voice in government and a place on the council, in opposition to his ailing father and Thomas Arundel, archbishop of Canterbury. The stories of Prince Henry's reckless and dissolute youth, immortalized by Shakespeare, and of the sudden change that overtook him when he became king, have been traced back to within 20 years of his death and cannot be dismissed as pure fabrication. This does not involve accepting them in the exaggerated versions of the Elizabethan playwrights, to which the known facts of his conduct in war and council



Henry V, painting by an unknown artist.
In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

provide a general contradiction. Probably they represent no more than the natural ebullience of a young man whose energies found insufficient constructive outlet. The most famous incident, his quarrel with the chief justice, Sir William Gascoigne, was a Tudor invention, first related in 1531.

Henry succeeded his father on March 21, 1413. In his early years his position was threatened by an abortive Lollard rising (January 1414) and by a conspiracy (July 1415) of Richard of York, earl of Cambridge, and Henry, Lord Scrope of Masham, in favour of Edmund Mortimer, earl of March. On each occasion Henry was forewarned and the opposition was suppressed without mercy. Neither incident long distracted him from his chief concern: his ambitious policy toward France. Not content with a demand for possession of Aquitaine and other lands ceded by the French at the Treaty of Calais (1360), he also laid claim to Normandy, Touraine, and Maine (the former Angevin empire) and to parts of France never in English hands. Although such demands were unlikely to be conceded even by the distracted government of France, Henry seems to have convinced himself that his claims were just, and not a merely cynical cover for calculated aggression. Yet if "the way of justice" failed, he was ready to turn to "the way of force," and warlike preparations were well advanced long before the negotiations were finally broken off in June 1415.

The French wars

Henry V's true genius is revealed in the planning and execution of his subsequent campaigns for the conquest of France. Before hostilities began, his diplomatic skill was exerted in an effort to secure the support or at least the neutrality of John the Fearless, duke of Burgundy. His attempts to deprive France of maritime assistance show an awareness of the importance of sea power unusual in medieval kings, and after the Battle of the Seine (August 1416), England's naval mastery of the Channel was not seriously disputed. At home, Henry turned to the systematic financing of his projected invasion, partly through large-scale borrowing, partly through parliamentary taxation, the generosity of which reflects his success in arousing national enthusiasm for the war. Henry began the struggle with the wholehearted support of the magnates and the backing of a united nation. His military strategy was conceived with equal ability. It stands in marked contrast with the haphazard and spasmodic operations of the English in France in the previous century. His main objective, to which the winning of battles was largely irrelevant, was the systematic reduction of the great towns and fortresses of northern France. These, kept as headquarters of permanent English garrisons, would become focal points for the subjection of the surrounding countryside; behind the soldiers were to come administrators and tax collectors, who would make the war pay for itself. Despite the forethought and grasp this plan displayed, its execution took longer than Henry had anticipated. It absorbed his energies for seven years and brought him to an early grave.

His first campaign brought the capture of Harfleur (September 1415) and the great victory of Agincourt (October 25, 1415). This resounding triumph made Henry the diplomatic arbiter of Europe: it won him a visit (1416) from the Holy Roman emperor Sigismund, with whom he made a treaty of alliance at Canterbury (1416) and whose influence was used to detach Genoa from its naval alliance with France. The cooperation of the two rulers led directly to the ending of the papal schism through the election of Martin V (1417), an objective that Henry had much at heart. Thereafter he returned to the long, grim war of sieges and the gradual conquest of Normandy. Rouen, the capital of northern France, surrendered in January 1419, and the murder of Duke John of Burgundy by the Dauphin's partisans in September 1419 brought him the Burgundian alliance. These successes forced the French to agree to the Treaty of Troyes on May 21, 1420. Henry was recognized as heir to the French throne and regent of France, and Catherine, the King's daughter, was married to him on June 2. He was now at the height of his power: but his triumph was short-lived. His health grew worse at the sieges of Melun and Meaux. He died of camp fever at Bois de Vincennes, August 31, 1422.

Character and ability

Henry's character is by no means wholly admirable. Hard and domineering, he was intolerant of opposition and could be ruthless and cruel in pursuit of his policy. His lack of chivalrous qualities deprives him of any claim to be regarded as "the typical medieval hero." Yet contemporaries united in praising his love of justice, and even French writers of his own day admired him as a brave, loyal, and upright man, an honourable fighter, and a commanding personality in whom there was little of the mean and the paltry. Although personally lacking in warmth, he had the capacity to inspire devotion in others, and he possessed high qualities of leadership. His piety was genuine, and on his deathbed he expressed a last wish that he might live to rebuild the walls of Jerusalem in a new crusade. In respect of ability, he must rank high among English kings. His achievement was remarkable: it has been rightly observed that "he found a nation weak and drifting and after nine years left it dominant in Europe." The tragedy of his reign was that he used his great gifts not for constructive reform at home but to commit his country to a dubious foreign war. His premature death made success abroad unlikely and condemned England to a long, difficult minority rule by his successor.

BIBLIOGRAPHY. C.L. KINGSFORD, *Henry V: The Typical Mediaeval Hero*, 2nd ed. (1923), the first modern scholarly biography; J.H. WYLIE and W.T. WAUGH, *The Reign of Henry the Fifth*, 3 vol. (1914–29), vol. 1 and 2 minutely detailed, vol. 3 more judicious; E.F. JACOB, *Henry V and the Invasion of France* (1947), and *The Fifteenth Century, 1399–1485* (1961), critical of Henry's achievements; H.F. HUTCHINSON, *Henry V* (1967), a popular account, mostly about the French War; C.T. ALLMAND, *Henry V* (1968), a useful short modern reappraisal; R.A. NEWHALL, *The English Conquest of Normandy, 1416–1424* (1924), the best assessment of Henry V as a soldier.

(C.D.R.)

Henry VII of England

By his victory over King Richard III at Bosworth Field in 1485, Henry Tudor, earl of Richmond, became king of England as Henry VII and thereby ended a long period of dynastic civil wars between the houses of Lancaster and York, later known as the Wars of the Roses. Henry VII (reigned 1485–1509) restored stable and efficient government, and by his caution and thriftiness laid the foundation for the achievements of the Tudor dynasty.

Henry, son of Edmund Tudor, earl of Richmond, and Margaret Beaufort, was born at Pembroke Castle on January 28, 1457, nearly three months after his father's death. His father was the son of Owen Tudor, a Welsh squire, and Catherine of France, the widow of King Henry V. His mother was the great-granddaughter of John of Gaunt, duke of Lancaster, whose children by



Henry VII, painting by an unknown artist, 1505. In the National Portrait Gallery, London.

By courtesy of the National Portrait Gallery, London

Catherine Swynford were born before he married her. Henry IV had confirmed Richard II's legitimation (1397) of the children of this union but had specifically excluded the Beauforts from any claim to the throne (1407). Henry Tudor's claim to the throne was, therefore, very weak and of no importance until the deaths in 1471 of Henry VI's only son, Edward, of his own two remaining kinsmen of the Beaufort line, and of Henry VI himself, which suddenly made him the sole surviving male representative of the House of Lancaster.

Early life. As his mother was only 14 when he was born and soon married again, Henry was brought up by his uncle Jasper Tudor, earl of Pembroke. When the Lancastrian cause crashed to disaster at the Battle of Tewkesbury (May 1471), Jasper took the boy out of the country and sought refuge in the duchy of Brittany. The House of York then appeared so firmly established that Henry seemed likely to remain in exile for the rest of his life. The usurpation of Richard III (1483), however, split the Yorkist party and gave Henry his opportunity. His first chance came in 1483 when his aid was sought to rally Lancastrians in support of the rebellion of Henry Stafford, duke of Buckingham, but that revolt was defeated before Henry could land in England. To unite the opponents of Richard III, Henry had promised to marry Elizabeth of York, eldest daughter of Edward IV; and the coalition of Yorkists and Lancastrians continued, helped by French support, since Richard III talked of invading France. In 1485 Henry landed at Milford Haven in Wales and advanced toward London. Thanks largely to the desertion of his stepfather, Lord Stanley, to him, he defeated and slew Richard III at the Battle of Bosworth on August 22, 1485. Claiming the throne by just title of inheritance and by the judgment of God in battle, he was crowned on October 30 and secured parliamentary recognition of his title early in November. Having established his claim to be king in his own right, he married Elizabeth of York on January 18, 1486.

Battle of Bosworth

Yorkist plots. Henry's throne, however, was far from secure. Many influential Yorkists had been dispossessed and disappointed by the change of regime, and there had been so many reversals of fortune within living memory that the decision of Bosworth did not appear necessarily final. Yorkist malcontents had strength in the north of England and in Ireland and had a powerful ally in Richard III's sister Margaret, dowager duchess of Burgundy. All the powers of Europe doubted Henry's ability to survive, and most were willing to shelter claimants against him. Hence, the King was plagued with conspiracies until nearly the end of his reign.

The first rising, that of Lord Lovell, Richard III's chamberlain, in 1486 was ill prepared and unimportant; but in 1487 came the much more serious revolt of Lambert Simnel. Claiming to be Edward, earl of Warwick, the son of Richard III's elder brother, George, duke of Clarence, he had the formidable support of John de la Pole, earl of Lincoln, Richard III's heir designate, of many Irish chieftains, and of 2,000 German mercenaries paid for by Margaret of Burgundy. The rebels were defeated (June 1487) in a hard-fought battle at Stoke (East Stoke, near Newark in Nottinghamshire), where the doubtful loyalty of some of the royal troops was reminiscent of Richard III's difficulties at Bosworth. Henry, recognizing that Simnel had been a mere dupe, employed him in the royal kitchens.

Then in 1491 appeared a still more serious menace: Perkin Warbeck, a handsome young Fleming who had been coached by Margaret to impersonate Richard, the younger son of Edward IV. Supported at one time or another by France, by Maximilian I of Austria, regent of the Netherlands (Holy Roman emperor from 1493), by James IV of Scotland, and by powerful men in both Ireland and England, Perkin three times invaded England before he was captured at Beaulieu in Hampshire in 1497. Apparently the Spanish monarchs, Ferdinand and Isabella, were still so doubtful of Henry's survival that in 1499 he had to resort to the judicial murder of Perkin Warbeck and the unfortunate Earl of Warwick before they would allow their daughter Catherine to come to England to marry Prince Arthur.

Even after this Henry was worried by the treason of Edmund de la Pole, earl of Suffolk, the eldest surviving son of Edward IV's sister Elizabeth, who fled to the Netherlands (1499) and was supported by Maximilian. Doubtless the plotters were encouraged by the deaths of Henry's sons, the infant Edmund in 1500 and Arthur in 1502, and of his wife, Elizabeth, in 1503. It was not until 1506, when he imprisoned Suffolk in the Tower of London, that Henry could at last feel safe. When he died on April 21, 1509, his only surviving son, Henry VIII, succeeded him without a breath of opposition.

Foreign policy. In the early years of his reign, in a vain attempt to prevent the incorporation of the duchy of Brittany into France, Henry found himself drawn along with Spain and the Holy Roman emperor into a war against France. But he realized that war was a very hazardous activity for one whose crown was both impoverished and insecure; and in 1492 he made peace with France on terms that brought him recognition of his dynasty and a handsome pension. Thereafter, French preoccupation with adventures in Italy made peaceful relations possible; but the support that Maximilian and James IV gave to Warbeck led to sharp quarrels with the Netherlands and Scotland. The economic importance of England for the Netherlands enabled Henry to induce Maximilian and the Netherlands to abandon the pretender in 1496 and to conclude a treaty of peace and freer trade (the Intercursus Magnus) that was beneficial to both countries.

Efforts to avoid war

With Scotland, the long tradition of hostility was harder to overcome; but Henry eventually succeeded in concluding in 1499 a treaty of peace, followed in 1502 by a treaty for the marriage of James IV to Henry's daughter Margaret. James's consent to the match may have been fostered by the arrival in England of Catherine of Aragon for her marriage with Prince Arthur in 1501. Spain had recently sprung into the first rank of European powers, so a marriage alliance with Spain enhanced the prestige of the Tudor dynasty, and the fact that in 1501 the Spanish monarchs allowed the marriage to take place is a tribute to the growing strength of the Tudor regime in the eyes of the European powers.

After Arthur's death in 1502, Henry was in a strong position to insist on the marriage of Catherine to his surviving son, Henry (later King Henry VIII), since he had possession both of Catherine's person and of half her dowry, and Spain needed English support against France. Indeed, in these last years of his reign, Henry had gained such confidence in his position that he indulged in some wild schemes of matrimonial diplomacy. But the caution

of a lifetime kept him from involvement in war, and his foreign policy as a whole must not be judged by such late aberrations. He had used his diplomacy not only to safeguard the dynasty but to enrich his country, using every opportunity to promote English trade by making commercial treaties (including agreements with Spain, the Netherlands, Florence, and Denmark). He made his country so prosperous and powerful that he was able to betroth his daughter Mary to the archduke Charles (afterward Emperor Charles V), the greatest match of the age.

Government and administration. In home affairs Henry achieved striking results largely by traditional methods. Like Edward IV, Henry saw that the crown must be able to display both splendour and power when occasion required. This necessitated wealth, which would also free the king from embarrassing dependence on Parliament and creditors. Solvency could be sought by economy in expenditure, such as avoidance of war and promotion of efficiency in administration, and by increasing the revenue. To increase his income from customs dues, Henry tried to encourage exports, protect home industries, help English shipping by the time-honoured method of a navigation act to ensure that English goods were carried in English ships, and find new markets by assisting John Cabot and his sons in their voyages of discovery. More fruitful was the vigorous assertion of royal fiscal rights, such as legal fees, fines and amercements, and feudal dues. This was largely achieved by continuing Yorkist methods in ordering most of the royal revenue to be paid into the chamber of the household, administered by able and energetic servants and supervised by the king himself, instead of into the royal exchequer, hidebound by tradition. So efficient and ruthless were Henry's financial methods, especially in his latter years, that he left a fortune to his successor and a legacy of hatred for some of his financial ministers.

Growth of crown's wealth

In restoring order after the civil wars, Henry used more traditional methods than was once thought. Like the Yorkist kings, he made use of a large council, presided over by himself, in which lawyers, clerics, and lesser gentry were active members. Sitting as the Court of Star Chamber, the council dealt with judicial matters, but less than was formerly thought. Nearly all the heavy fines levied for the illegal retaining of armed men toward the end of his reign were imposed in the Court of King's Bench and by the justices of assize. Special arrangements were made for hearing poor men's causes in the council and for trying to promote better order in Wales and the North by setting up special councils there; and more powers were entrusted to the justices of the peace. The King, moreover, could not destroy the institution of retainers since he depended on them for much of his army and society regarded them as natural adjuncts of rank. So Henry's government was conservative, as it was in its relations with Parliament and with the church.

Character. The whole of Henry's youth had been spent in conditions of adversity, often in danger of betrayal and death, and usually in a state of poverty. These experiences, together with the uncertainties of his reign, taught him to be secretive and wary, to subordinate his passions and affections to calculation and policy, to be always patient and vigilant. There is evidence that he was interested in scholarship, that he could be affable and gracious, that he disliked bloodshed and severity; but all these emotions had to give way to the needs of survival. The extant portraits and descriptions suggest a tired and anxious-looking man, with small blue eyes, bad teeth, and thin white hair. His experiences and needs had also made him acquisitive, a trait that increased with age and success, and one that was opportune for both the crown and the realm.

BIBLIOGRAPHY. R.L. STOREY, *The Reign of Henry VII* (1968), is a good clear summary of recent research, with a useful bibliography. *English Historical Documents*, vol. 5, *1485–1558*, ed. by C.H. WILLIAMS (1967); and *The Reign of Henry VII from Contemporary Sources*, ed. by A.F. POLLARD, 3 vol. (1913–14), provide valuable collections of documents for the reign. J.D. MACKIE, *The Earlier Tudors, 1485–1558* (1952); and G.R. ELTON, *England Under the Tudors* (1955), offer interesting interpretations of the reign and useful bibliographies.

(A.R.M.)

Henry VIII of England

As king of England from 1509 to 1547, Henry VIII presided over the beginnings of the English Reformation, unleashed by his own matrimonial involvements, even though he never abandoned the fundamentals of the Roman Catholic faith. Exceptionally well served by a succession of brilliant ministers, he turned upon them all; those he elevated, he invariably cast down again. Attracted to humanist learning and something of an intellectual himself, he was responsible for the deaths of the outstanding English humanists of the day. Six times married, he left a minor heir and a dangerously complicated succession problem. Of his six wives, two joined a large tally of eminent persons executed for alleged treason; yet otherwise his regime observed the law of the land with painful particularity. Formidable in appearance, in memory, and in mind, and fearsome of temper, he yet attracted genuine devotion and knew how to charm people. Monstrously egotistical and surrounded by adulation, he nevertheless kept a reasonable grasp on the possible; forever taking false steps in politics, he emerged essentially unbeaten and superficially successful in nearly everything he attempted to do. It is no wonder that this embodiment of personal monarchy, so mixed in all that touched him, has divided opinion ever since his death; and the debate is far from over.

By courtesy of the Duke of Rutland; photograph the Royal Academy of Arts, London



Henry VIII, painting by Hans Holbein the Younger, *c.* 1538. In the collection of the Duke of Rutland.

Accession to the throne. Henry was born at Greenwich on June 28, 1491, the second son of Henry VII, first of the Tudor line, and Elizabeth, daughter of Edward IV, first king of the short-lived line of York. When his elder brother, Arthur, died in 1502, Henry became the heir to the throne; of all the Tudor kings, he alone spent his childhood in calm expectation of the crown, which helped give an assurance of majesty and righteousness to his self-willed and ebullient character. A child of

Childhood and first marriage

quick-witted intelligence, he excelled in book learning as well as in the physical exercises of an aristocratic society, and, when in 1509 he ascended the throne, great things were expected of him. Six feet tall, powerfully built, a tireless athlete, huntsman, and dancer, he promised to bring to England the joys of spring after the long winter of Henry VII's reign. Yet those who lavishly praised his milk-and-roses beauty, his lively mind, and his affability overlooked the suspiciousness lurking at the back of his close-set, small eyes and the ready temper flaring up at even small checks.

Observers ought also to have taken note of the manner in which this early popularity was turned to political use. Henry and his ministers exploited the dislike inspired by his father's energetic pursuit of royal rights by sacrificing, without a thought, some of the unpopular institutions and some of the men that had served his predecessor. Yet the unpopular means for governing the realm soon reappeared because they were necessary. Soon after his accession, Henry married Catherine of Aragon, Arthur's widow, and the court embarked on a round of lavish entertainments that ate into the modest reserves accumulated in the previous reign.

More serious was Henry's determination to engage in the other proper avocation of kings, military adventure. Europe was being kept on the boil by rivalries between the French and Spanish kingdoms, mostly over Italian claims; and, against the advice of his older councillors, Henry found occasion in 1512 to enter upon the scene as the ally of his father-in-law, Ferdinand II of Aragon, against the ancient enemy, France, and ostensibly in support of a threatened Pope, to whom the devout King for a long time paid almost slavish respect.

The war was something of a farce. Henry himself displayed no military talent in a campaign that included the fortuitous rout of a French force and led to the capture of Tournai in northern France, occupied until 1518, while a real victory was won by the Earl of Surrey at Flodden (1513) against a Scottish invasion. Despite the obvious pointlessness of the fighting, the appearance of success was popular. Moreover, in Thomas Wolsey, who organized his first campaign in France, Henry discovered his first outstanding minister. By 1515 Wolsey was archbishop of York, lord chancellor of England, and a cardinal of the church; more important, he was the King's good friend, to whom he gladly left the active conduct of affairs. Henry never altogether abandoned the positive tasks of kingship and often interfered in business; though the world might think that England was ruled by the Cardinal, the King himself knew that he possessed perfect control any time he cared to assert it, and Wolsey only rarely mistook the world's opinion for the right one.

Wolsey's ascendancy

Nevertheless, the years from 1515 to 1527 were marked by Wolsey's ascendancy, and his initiatives set the scene. The Cardinal had some occasional ambition for the papal tiara, and this Henry supported; Wolsey at Rome would have been a powerful card in English hands. In fact, there was never any chance of this happening, any more than there was of Henry's election to the imperial crown, briefly mooted in 1519 when the emperor Maximilian I died, to be succeeded by his grandson Charles V. That event altered the European situation. In Charles, the crowns of Spain, Burgundy (with the Netherlands), and Austria were united in an overwhelming complex of power that reduced all the dynasties of Europe, with the exception of France, to an inferior position. In celebrating, in 1518, the London treaty of universal peace—a multilateral covenant based on the new concept of collective security—Wolsey had been able to display an effective English initiative; a year later, the new situation so far circumscribed freedom of action that Henry's allegedly independent role ceased to convince the experts. The fact was briefly disguised when he was wooed by the Emperor and the King of France as they moved into position for a showdown; in 1520 and 1521 a series of public and private negotiations, including the lavish splendour of a meeting with Francis I of France at the so-called Field of Cloth of Gold near Calais (1520), seemed to show Henry as a monarch equal to the greatest that Europe could produce. But there was little meaning in it. From 1521, Henry became an outpost of Charles V's imperial power, which at Pavia (1525), for the moment, destroyed the rival power of France. In a desperate attempt to escape from this domination, Wolsey tried to reverse alliances at this unpropitious moment. This policy brought reprisals against the vital English cloth trade with the Netherlands and lost the advantages that alliance with the victor of Pavia might have had. It provoked a serious reaction in England, nor did the fact that he had helped to inspire it prevent Henry from concluding that Wolsey's usefulness might be coming to an end.

Loss of popularity. While the greatness of England in Europe was being shown up as a sham, the regime was also losing popularity at home. The fanciful expectations of the early days could not, of course, endure; some measure of reality was bound to intrude. As it was, those influential shapers of posterity's opinions—journalists and writers—continued to be full of hope for a king who, from 1517, commanded the services of a new councillor, Sir Thomas More, one of the outstanding minds of the day. But More soon discovered that Henry found it easy to keep his enjoyment of learned conversation apart from the conduct of policy. Nothing for the moment could dent the strength of Wolsey's ascendancy, and this had serious drawbacks for the King who supported him. The country was showing increasing signs of its unwillingness to endure priestly rule; there were riots against the clergy, foreign merchants, tax gatherers, and exploiting landlords; and Wolsey's efforts to remedy grievances—always promising fairer than he could perform—only exasperated men of influence without bringing satisfaction to the poor. In the early 1520s, plague and trade depression afflicted the country, and Wolsey's always expensive foreign policy absorbed large sums in the equipping of futile expeditions to France (1522) and larger ones in subsidies to unreliable allies. Feelings came to the boil in the years 1523–24. Although he disliked Parliaments, Wolsey had to agree to the calling of one in 1523, but the taxes voted—with great difficulty—were well below what was required. Next year, the attempt to levy a special tax led to such fierce resistance that Henry rescinded it, he and the Cardinal both trying to take the credit for the remission of what they had been jointly responsible for imposing. While he had Wolsey to take the blame, Henry could afford such fiascoes; the Cardinal could not.

By 1527 a government policy that, though seemingly Wolsey's, was really the King's was facing bankruptcy; ineffective abroad, unpopular at home, it made the regime look as empty of positive purpose as in fact it was. At this point, the King entered affairs unmistakably and spectacularly, to prove that 16th-century politics were bound to revolve around the personal affairs of monarchs. Among his failures so far had been his or Catherine's inability to provide a male heir to the throne; several stillbirths and early deaths had left only a girl, the princess Mary (born in 1516), to carry on the line, and no one relished the thought of a female succession with all the dynastic and political uncertainties it would bring. Being the man he was, Henry could not suppose the fault to be his. His rapidly growing aversion to Catherine, six years older than himself and aged even more by her experiences, received powerful assistance from his infatuation with one of the ladies of the court, Anne Boleyn, the sister of one of his earlier mistresses. Henry was no profligate, and the stories of his sexual adventures are wildly exaggerated. Indeed, he had a strong streak of prudery and is thought by some to have been sexually timid. But he was a passionate man in the prime of physical condition, and he sought the occasional relief from marriage to a worthy but ailing wife to which princes have generally been held entitled. In Anne he met his match; this 20-year-old girl, brought up in a tough school of courtly intrigue, would be more than a king's mistress. It took Henry, who in any case needed to marry her if the expected issue was to solve the succession problem, some six years to achieve their joint purpose. Inadvertently, he provoked a revolution.

Anne Boleyn

From 1527 Henry pursued what became known as "the

Divorce from Catherine

King's great matter": his divorce from Catherine. He convinced himself that his first marriage had been against the divine law; that is, against the biblical injunction (Lev.) forbidding marriage with a brother's widow. The deaths of the children proved God's judgment on the union. With his characteristic readiness to convert his own desires into the law of God, Henry rapidly assured himself that he was living in mortal sin with Catherine and must find relief if he was again to become acceptable to God. So quickly did he arrive at this immovable position that he bypassed without notice a neat device of Wolsey's that might have solved his trouble painlessly but on a lower plane of conscientious conviction. He appealed to Rome for a declaration of annulment. Popes had usually obliged kings in such matters, but Henry had picked both his time and his case badly. He was asking Pope Clement VII to help him discard the Emperor's aunt, but Clement, the Emperor's prisoner in 1527–28, never thereafter dared resist Charles, whose powerful feelings of familial honour and public prestige barred any concession to Henry's wishes. Moreover, the Pope's reluctance was increased by the fact that he was being asked to declare illegal an earlier exercise of papal power —which had licensed Henry's marriage to his brother's widow—of a kind that brought a good deal of money to the papal coffers.

Thus Henry's attempts to solve his dilemma in the accepted legal way were doomed from the start. Wolsey, in a worse dilemma, since only success in the impossible could keep him in power, obtained a trial of the case in England, but this was frustrated by his fellow judge, Cardinal Campeggio, on orders from Rome (1529). Within weeks, Wolsey was ousted; 15 years of high-handed activity rose up against him, and, though the King spared his life, the Cardinal was stripped of power and possessions, to die a year later while being brought to London for further persecution. But Wolsey's disappearance solved nothing, and the councillors who succeeded him could offer little help to their king, who knew only what he wanted, not how to get it.

The chancellorship went to Thomas More, who had told Henry that he did not approve of the divorce and who wished to devote himself to a fight against Lutheran heresy, the alleged dangers of which had driven everything else from a mind once dedicated to rational reform. His policy of assisting the bishops and maintaining the privileges of the church was hampered by the violent hostility of the influential part (nobility, gentry, and mercantile interest) of the nation, now effectively represented in the Parliament that Henry called in November 1529, in order to have backing for his battle with the Pope. The anticlericalism of the Commons increased the pressure on Rome, threatened with loss of power and revenue, even possibly with secession, but it conflicted with the renewal of bigoted orthodoxy urged on by More. Confusion, in fact, was the keynote of policy for some three years, while the King dithered between hope that Rome might yet be forced to let the formal trial of his first marriage take place in England and stirrings of a more radical nature—to reject Rome outright. But though he occasionally talked of doing just that, neither he nor anyone else knew how to convert talk into action.

The breach with Rome. Action called for a revolution, and the revolution required a man who could conceive and execute it. That man was Thomas Cromwell, who, in April 1532, won the battle for control of the council and thereafter remained in rarely troubled command for some eight years. The revolution consisted of the decision that the English Church should separate from Rome, becoming effectively a spiritual department of state under the rule of the king as God's deputy on earth. It was accomplished by a massive burst of parliamentary legislation in the years 1533–36. The revolution that he had not intended gave the King his wish: in January 1533 he married Anne Boleyn; in May a new archbishop, Thomas Cranmer, presided over the formality of a trial that declared the first marriage annulled; in September the princess Elizabeth was born. The Pope retaliated with a sentence of excommunication; it troubled no one.

Head of Church of England

The supreme headship on earth over the Church of England, though he had not sought it, represented Henry's major achievement. It had very wide ranging consequences, but those that immediately concerned the King were two. In the first place, the new title consolidated his own concept of kingship, his conviction that (as he once said) he had no superior on earth. It rounded off the majestic image of divinely instituted royal rule that it was Henry's constant ambition to present to an awed and obedient world. But, in the second place, it created a real personal problem for a king who in 1521, in his book against Luther (*Assertio septem sacramentorum adversus Martinum Lutherum*, certainly Henry's own work), had expressed a profound devotion to the papacy and been rewarded with the title of Defender of the Faith. Now he had thrust out the Pope, and the language used was even more abusive than that employed before had been respectful. In the 1530s, turning against the pope was equal to encouraging the Protestant Reformation, a thing attractive to Cranmer and Cromwell (and perhaps Anne Boleyn), but not to Henry, who never forgot Luther's rude contempt for him. The religion of the newly independent church was for its head to settle: where was he to take his stand? For the rest of his life, Henry, who justly prided himself on his theological learning, was to give much time and thought to the nature of the true religion. With the exception of the papal primacy, he never gave up the main tenets of the faith in which he had grown up, but he changed his mind on details and arrived at an eclectic amalgam of his own in which transubstantiation (the doctrine that the bread and wine used in Holy Communion are changed into the body and blood of Christ) and clerical celibacy mingled with radical views about the worldly authority of the church and man's ability to seek salvation without the aid of priests.

Domestic reforms. Cromwell's decade, the 1530s, was the only period of the reign during which a coherent body of policies was purposefully carried through. Cromwell's work greatly enlarged Henry's power, especially by transferring to the crown the wealth of the monasteries, dissolved in 1536–40, and new clerical taxes; but it also, more explicitly than ever, subjected the King to the law and to the legislative supremacy of Parliament. Since Henry knew how to work with parliaments, the immediate effect was to make him appear more dominant than ever and to give to his reign a spurious air of autocracy—spurious because in fact the rule of law remained to control the sovereign's mere will. The appearance of autocracy was misleadingly emphasized by the fact that all revolutions have their victims. As heads rolled, the King's earlier reputation as a champion of light and learning was permanently buried under his enduring fame as a man of blood. Old friends such as More, refusing to accept the new order, fell before the onslaught, as did some 50 other men caught by the treason laws. Between 1538 and 1541 the families of Pole and Courtenay were destroyed by the axe for treasons linked with efforts abroad to reverse the course of events in England but mainly because they could claim royal blood and represented a dynastic danger to the unprolific Tudor line.

Later marriages

The King now embarked on the series of matrimonial adventures that made him appear both a monster and a laughingstock. He soon tired of Anne, who failed to produce a male heir; in 1536 she was executed, with other members of the court, for alleged treasonable adultery. Catherine of Aragon, rejected but unbowed, had died a little earlier. Free at last, Henry immediately married Jane Seymour, who bore him his son Edward but died in childbirth (1537); she not only escaped the fates of her fellow wives but remained the sentimental King's one remembered attachment. The next three years were filled with attempts to replace her, and the bride chosen was Anne, sister of the Duke of Cleves, a pawn in Cromwell's policy for a northern European alliance against dangers from France and the Emperor. But Henry hated the first sight of her and at once demanded to be freed, an end achieved by the divorce to which the lady obligingly consented, in return for a sizable estate, upon which she settled in happy obscurity.

Physical and mental decline. The Cleves fiasco destroyed Cromwell; it enabled his many enemies to turn the King against him, and in July 1540 his head fell on the scaffold. There is no doubt that Henry had by now become truly dangerous. He had always been secretive and suspicious, trusting no one and determined to know the inmost thoughts of all other men. Now he was beginning to show paranoiac tendencies. Convinced that he controlled everyone, he was in fact readily manipulated by those who knew how to feed his suspicions and pander to his self-righteousness. Full of experience—the oldest king in Europe—and increasingly competent in the routine of rule, he lacked the comprehensive vision and large spirit that would have made him a great man, instead of the image of greatness painted on a Tudor canvas. His temperamental deficiencies were aggravated by what he regarded as his undeserved misfortunes and by ill health. Ever since a fall in his last tournament in 1536, he had suffered from headaches. In the late 1530s he developed the ulcer on his leg (not syphilitic) that was to keep him in agony for most of his remaining days. No longer able to subject his splendid body to the strenuous exercises to which it had been accustomed, he grew enormously fat. Fifty was a good age in the 16th century, but even so Henry aged faster than he should have done. His mind did not weaken; there was no senility. But he grew restless, peevish, and totally unpredictable; often melancholy and depressed, he was usually out of sorts and always out of patience. In 1540–42 he briefly renewed his youth in marriage to the 20-year-old Catherine Howard, a promiscuous young woman whose folly in continuing her adventures, even as queen, brought her to the block. The blow finished Henry. Thereafter, he was really a sad and bitter old man, and, though he married once more, to find a measure of peace with the calm and obedient Catherine Parr, his physical ruin was complete, and behind the trappings of capricious authority there was only a sick man frightened of the coming dissolution.

But he was still the king and, from Cromwell's fall (which he regretted too late), the only maker of policy. Policy in the hands of a sick, unhappy, violent man was not likely to be either sensible or prosperous, and so it proved. Left to himself, Henry concentrated on keeping the realm united, despite the growing strife between the religious factions, and on keeping before the world his own image as the glorious monarch of the age. The first resulted in frequent explosions against the ingratitude of subjects who dared maintain their petty squabbles when he had told them what to believe and in frequent outbursts against his councillors, whose careers in the 1540s became unprecedentedly precarious and variable. The second brought him back to his first love—war and conquest, the sport of kings.

Military successes

In 1542 the Emperor and the King of France resumed hostilities. After a pretense of independence, Henry again joined in on the side of the former, which promptly brought in the Scots on the side of their allies, the French. Now it was success in war that was to cause trouble. The Scots were routed at Solway Moss (1542), and their king died soon after: this opened the possibility of subjugating that country permanently by means of a marriage alliance between the infant heirs to both thrones. In France, an English army conquered Boulogne. But the Scottish dream quickly collapsed as Henry's crude handling of that nation gave control to a pro-French party, determined to resist even an alliance with England; physical conquest was beyond the King's means. Boulogne, a useless prize, was expensive to keep but—since honour was involved—almost impossible to give up. Henry personally managed both the war and the subsequent negotiations, and he displayed amazing energy for so sick a man. But energy is not the same thing as competence. The war proved ruinous. Heavy taxation and borrowing did not suffice to cover its cost. Thus, money had to be raised by selling off the monastic lands, which had brought a good income, so that the financial improvement made by Cromwell was quickly lost. The desperate expedient of debasing the coinage, though it brought temporary succour, led to a violent inflation that made things worse. Yet, even after the Emperor made peace with France behind his back (1544), Henry would not let go; it was only two years later that he could be brought to end the pointless struggle.

As the year 1546 drew to a close, it was apparent to all observers that the King had not long to live. Not that it was clear to the man most concerned; he continued as before, lamenting religious dissension, attending to the business of government, continuing the pretense of deathless majesty, destroying the powerful Howard family, whom he suspected of plotting to control his successor. Conscious almost to the very end, he died on January 28, 1547. He left the realm feeling bereft and the government the more bewildered because, to the last, he had refused to make full arrangements for the rule of a boy king.

Assessment

Immediately and ever since, the impact of that overpowering personality maintained the achievement of which he himself was proudest: Henry VIII has always seemed the very embodiment of true monarchy. Even his evil deeds, never forgotten, have been somehow amalgamated into a memory of greatness. He gave his nation what it wanted: a visible symbol of its nationhood. He also had done something toward giving it a better government, a useful navy, a start on religious reform and social improvement. But he was not a great man in any sense. Although a leader in every fibre of his being, he little understood where he was leading his nation. But, if he was neither statesman nor prophet, he also was neither the blood-stained monster of one tradition nor the rowdy *bon vivant* of another. Cold, self-centred, ungiving, forever suspicious of the ways of the world, he could not live down to the second stereotype; despite a ruthlessness fed by self-righteousness, he never took the pleasure in killing required of the first. Simply, he never understood why the life of so well-meaning a man should have been beset by so many unmerited troubles.

BIBLIOGRAPHY. J.J. SCARISBRICK, *Henry VIII* (1968), supersedes all earlier biographies, especially that by A.F. POLLARD, *Henry VIII* (1902, reprinted 1966). LACEY BALDWIN SMITH, *Henry VIII: The Mask of Royalty* (1971), though inclined to overdramatize, interestingly discusses the King's last years. The often discussed medical problems are best studied in A.S. MACNALTY, *Henry VIII: A Difficult Patient* (1953). A.F. POLLARD, *Wolsey* (1929, reprinted 1965), still gives the most searching account of the first half of the reign; G.R. ELTON, *The Tudor Revolution in Government: Administrative Changes in the Reign of Henry VIII* (1953), and *Policy and Police* (1972), provide the best analysis of the Cromwell era. For the religious setting, see especially A.G. DICKENS, *The English Reformation*, rev. ed. (1964).

(G.R.E.)

Henry IV of France

The first of the Bourbon kings of France, Henry IV brought unity and prosperity to the country after the ruinous 16th-century Wars of Religion. Though not a great strategist, his gallantry made him a great military leader; though never an efficient administrator, his political insight and understanding of his people (he is credited with having promised them a chicken to eat every Sunday) made him an efficient ruler. He is one of the most popular figures in French history, for his amorous as well as his political achievements. He deserved to be called Henry the Great and good king Henry.

Prince of Béarn. Henry de Bourbon-Navarre was born on December 13, 1553, at Pau, near the Franco-Spanish border. The son of Antoine de Bourbon, duc de Vendôme, and Jeanne d'Albret, queen of Navarre from 1555, Henry, through his father, was in the sole legitimate line of descent from the Capetian kings of France. It was scarcely to be expected, however, that he would one day succeed to the throne of France, since Catherine de Médicis had already borne three sons to the reigning king, Henry II, and would soon bear him a fourth. Prince Henry spent most of his early childhood in Béarn. From 1561 to 1567 he lived with his second cousins, the children of the King of France, among whom was his future wife Marguerite.



Henry IV of France by Frans Pourbus the Younger (1569–1622). In the Louvre.
Giraudon

The religious crisis between Roman Catholic and Protestant forces was then coming to a head, leading to the long period of civil war. Antoine de Bourbon temporarily allied himself with the Protestants but changed sides and was mortally wounded in battle against them. Henry's mother, Jeanne d'Albret, held firm and publicly announced her Calvinism on Christmas 1560.

Henry had just turned 13 when his mother brought him back to Béarn. At a crucial age in his intellectual development, he was brought up in the strict principles of Protestantism. About the same time he began his military education. In the autumn of 1567, he served as nominal head of a punitive expedition launched against the rebellious Catholic gentry of lower Navarre, which ended in an easy victory. Less than a year later Jeanne d'Albret, who had remained neutral during the first two religious wars, readily entered the third. She proceeded to La Rochelle, where she put her son into the charge of her brother-in-law Louis I de Bourbon, prince of Condé, who was leader of the Protestant (Huguenot) forces.

The Protestants were surprised and defeated near Jarnac on March 13, 1569, by the duc d'Anjou, the future Henry III, and Condé was killed. Jeanne d'Albret immediately hastened to the scene and proclaimed her son head of the army, though actual command was exercised by Gaspard de Coligny, from whom Henry would receive his military education. Henry and his cousin, Henri, the young prince of Condé, were present on October 3 at the Battle of Moncontour, when the Protestants were again defeated. Coligny, however, forbade Henry to draw his sword.

The 16-year-old general received his baptism of fire near Arnay-le-Duc on June 26, 1570, when he led the first charge of the Huguenot cavalry. The long campaign through the ravaged provinces, extending from Poitou to the heart of Burgundy, forged in him the soldierly spirit that he would retain throughout his life and made him reflect on the disaster that had befallen the kingdom.

King of Navarre. Peace was concluded in August 1570, and a very liberal edict was granted the Protestants. Many persons, including Catherine de Médicis, hoped the civil war had come to an end. In order to strengthen the peace, the Queen sought to arrange a marriage between Prince Henry and her daughter, Marguerite de Valois, a project that had first been considered during Henry II's lifetime, when the future couple were still children. After difficult negotiations, which lasted until the spring of 1572, the Queen Mother and the Queen of Navarre reached an agreement. The latter, who preceded her son to Paris, had scarcely arrived when she died June 9 of a respiratory illness. Prince Henry thereby became king of Navarre and sovereign lord of Béarn. He and Marguerite exchanged vows August 18 before the main portal of Notre Dame, but only Marguerite attended mass with the royal family.

St. Bartholomew's Day Massacre

On August 24 came the St. Bartholomew's Day Massacre, in which thousands of French Protestants were cut down by royal forces. The marriage was publicly styled the "scarlet nuptials" because of the bloodshed. Ordered by his brother-in-law Charles IX to abjure his Protestant faith, Henry yielded. His conversion was obviously of dubious sincerity, and he was therefore held for three and a half years at the courts of Charles IX and then Henry III. Careful to restrain his impatience, he hid his forceful personality from his detainers. The wily Catherine de Médicis was deceived along with everyone else, and it was with her secret consent that her son-in-law escaped from the court at the beginning of 1576.

Once free, the Prince displayed his sharp intellect and political acumen in his role as protector of the Protestant churches. His common sense—one of his outstanding traits, except in love affairs—manifested itself when civil war broke out anew at the end of 1576. The Huguenots fared badly, and Henry, evaluating the situation, was able to persuade his coreligionists to give up the struggle and accept the Treaty of Bergerac on September 17, 1577, despite the sacrifices it imposed on them.

In the meantime, Catherine de Médicis went to Guyenne to return Marguerite to her husband, who had left her with alacrity, and, more important, to seek a lasting peace. While this subject was being negotiated at Nérac, Henry learned that the Catholics had seized the Château de La Réole on the Garonne. His response, a crushing surprise attack on Fleurance, revealed him to be a born military leader. In the spring of 1580, after five days of furious street fighting, which delighted him, he subdued Cahors, emerging from battle "all blood and powder," in his own words. When his town of Mont-de-Marsan was not returned to him, he stormed it by night (November 1583) without firing a shot.

War and politics were not enough, however, for this restless man, and Henry enjoyed romantic as well as territorial conquests. In the summer of 1583, he began a love affair that lasted for several years with Corisande d'Andouins, comtesse de Guiche; a member of the highest nobility of Béarn, she was a young and beautiful widow, saluted by the essayist Michel de Montaigne as "the great Corisande."

Heir presumptive to the throne. On the death of Henry III's brother, François, duc d'Anjou, in 1584, Henry de Bourbon-Navarre became the heir presumptive to the throne of France. He was irrevocably opposed, however, by the militant Catholics of the Holy League, who were unwilling to accept a Protestant king, and by the Pope, who excommunicated him and declared him devoid of any right to inherit the crown. Headed by Henri, duc de Guise, and his brothers, the League claimed to be the defender of the ancestral faith of France, but its increasing reliance on Spanish support rapidly became a serious threat to French independence. Henry III lacked the strength to contain the League's overwhelming influence.

The Holy League

The King of Navarre himself could rely only on the grudging support—peppered with shrewd demands—of Elizabeth of England and on the distant German Protestant princes. In France, in addition to the Huguenots, a few enlightened and liberal Catholics, such as Montaigne, supported him. Ultimately he could depend only on himself. In this crucial episode of French history, in which the very independence of the nation was at stake, Henry's activity was the essential factor. Although too prone in peace to neglect public affairs for private pleasure, he was an unrivalled leader in times of peril. Quick to grasp the significance of every situation, he was equally prompt to act, and victory was invariably the reward of his bold swiftness. He was not a brilliant strategist but had the ability to inspire men to action, as much by his own example as by the spoken or written word. Four centuries later, his notes and speeches still have the impact and clarity of a clarion call.

The military situation drew and held Henry within an extended radius of his base at La Rochelle. Though this town was a symbol of Protestant resistance, as well as a fortress, its extreme religious atmosphere could not keep Henry away from amorous escapades.

War of the three Henrys

In the fall of 1587, the outcome of the war hinged on the impending encounter between Henry and the army of the French king Henry III, who came increasingly under the influence of the League. The King placed another brother-in-law, Anne, duc de Joyeuse, in command of the army. The battle took place on October 20 on the outskirts of Coutras about 30 miles northeast of Bordeaux. Joyeuse was killed, and his troops were cut to pieces. The significance of the event was more political than military, for two of the young Catholic stepbrothers of the Prince of Condé had fought with the King of Navarre.

It became clear that the struggle, initially a religious conflict, had become dynastic and national when the League accepted the daughter of Philip II of Spain and Elisabeth de Valois as the next ruler of France. Henry III had the merit to grasp the full meaning of the situation for the future of France and, after long hesitation, had the Duc de Guise assassinated. The logic of the events left him with no choice but reconciliation with the King of Navarre, which took place April 30, 1589.

The united forces of the two princes laid siege to Paris on July 30, but on August 1 Henry III, the last of the Valois, was stabbed in his headquarters at Saint-Cloud. He died the next day, after staunchly proclaiming the head of the House of Bourbon, first prince of the blood, as his successor in accordance with dynastic law.

Henry IV. Henry IV was now king of France, but it would take him nine years to conquer his kingdom. Many of the Catholic gentry who had remained loyal to Henry III deserted him, and his army was growing exhausted. He had to withdraw from the outskirts of Paris. A few weeks after his accession, he fought Charles de Lorraine, duc de Mayenne, in sight of the Château d'Arques in Normandy. Mayenne, who had become head of the League on the death of his brother Guise, suffered a more serious defeat at Ivry on March 14, 1590. It was at Ivry that Henry IV issued his famous call to his troops: "If your cornets fail you, rally to my white plume—you will find it on the path to honour and victory!" He was unable to break the resistance of the capital, however, despite the terrible famine that came in the wake of his long siege. He then laid siege to Chartres, which capitulated April 10, 1591. During the siege of Chartres, Henry IV became involved with Gabrielle d'Estrées, who of all his women was to exert the most influence over him.

Soon after Chartres, Henry IV captured Noyon, but he could not conquer Rouen. The war dragged on interminably, and the King realized that it had to be ended at any cost. After long hesitation, he abjured Calvinism on July 25, 1593, in the basilica of Saint-Denis, the sepulchre of the kings of France.

Abjuration of Calvinism

Though many remained unconvinced of his sincerity, the monarch's conversion still brought quick results. Important towns, notably Orléans and Lyons, submitted in growing numbers. Reims remained loyal to the League, so the King was crowned at Chartres on February 27, 1594. The conversion removed all legitimate pretext for resistance, and on March 22 he finally entered Paris. Whether or not he made the comment attributed to him —"Paris is well worth a mass!"—he went, amid cheers, to hear the Te Deum at Notre Dame. He was, as he wrote, "in his triumphal chariot."

During the spring and summer, many League towns hastened to acknowledge royal authority. Laon, however, submitted only after a difficult siege. Even after Pope Clement VIII removed the ban of excommunication from Henry IV on September 17, 1595, Brittany remained in the hands of Philippe-Emmanuel de Lorraine, duc de Mercoeur, the younger brother of Guise and Mayenne.

Brittany was held by the League only through Spanish support. In order to bring this situation to an end, the King had declared war on Philip II in January 1595. On June 5 at Fontaine-Française in Burgundy, he overwhelmed the Spanish cavalry. In return, Philip II's forces seized Cambrai and then Calais and Ardres. Henry succeeded, after a six-month siege, in taking La Fère, but, on the night of March 11–12, 1597, panic swept Paris, for Spain had taken Amiens by surprise.

Henry quickly responded to the situation, exclaiming, "Enough of being king of France, it is time to be king of Navarre!" He lacked both men and money but was able to raise both as if by a miracle; as one of his chief ministers, Nicolas de Neufville, seigneur de Villeroy, wrote in a private letter, "The king has enough courage for everyone." He succeeded in forcing the Spaniards to surrender Amiens on September 19, and the treaty between France and Spain was concluded at Vervins on May 2, 1598. Spain kept Cambrai, which was not returned to France until 1677.

Henry also prepared to unseat the duc de Mercoeur in Brittany. He was able to come to a bloodless agreement by arranging a marriage between Françoise de Lorraine and César de Vendôme, the eldest of his two sons by Gabrielle d'Estrées; at that time César was about four years old. The high point of Henry IV's visit to Brittany was at Nantes, where Gabrielle gave birth to a daughter, and, on April 13, 1598, Henry signed the famous edict bearing the name of the city. The Edict of Nantes proclaimed freedom of conscience and granted many places of worship and nearly a hundred places of refuge to the Protestants.

Edict of Nantes

Henry IV had united the kingdom and achieved peace at home and abroad. During the remainder of his brief reign, he devoted himself to healing France's wounds. He faced the risk of having his work undone, however, should he die without leaving a universally accepted successor. Blinded by an unlucky passion and exploited by the coterie that had gathered around Gabrielle, Henry nearly married his mistress, which would have made César, the son of their double adultery, heir to the crown of Saint Louis. Such an act could not have failed to provoke a war of succession when the King died. Gabrielle's death in childbirth on April 10, 1599, however, removed all reason for Pope Clement VIII's reluctance to annul the marriage of Henry IV and Marguerite de Valois. The annulment made it possible for the King to marry the princess of Tuscany, Marie de Médicis, in October 1600, despite the protests of his new mistress, Henriette d'Entragues. Continuing his effort to unify the kingdom, Henry went on to subdue Charles-Emmanuel, duc de Savoie, forcing him to cede Bresse and Bugey to France. The new queen gave birth on September 27, 1601, to the Dauphin, the future Louis XIII, and eventually to four other children. Henriette d'Entragues, Jacqueline de Bueil, and Charlotte des Essarts also bore him many children.

Even after unification of the kingdom, Henry IV's reign was not tranquil. More than a century of impassioned religious conflict had torn France asunder, and peace was not to be easily regained. Political intrigue was closely intertwined with the religious struggle, as in the treason case of the marshal de Biron, who was decapitated in 1602, and the conspiracy of the Entragues and Charles, comte d'Auvergne, the illegitimate son of Charles IX. There was more than one attempt on Henry's life. In December 1594, Jean Châtel wounded the King with a dagger. Châtel was tortured, and his crime was used as a pretext to expel the Jesuits from the kingdom. The regicide attempts were born of the still-widespread belief that, behind the mask of his hypocritical conversion, Henry intended to ensure the triumph of heresy.

Nonetheless, France flourished anew. Maximilien de Béthune, duc de Sully, the most renowned of Henry's councillors, reorganized the national finances, stabilized the economy, and the country prospered. The groundwork was laid for the splendour of the next two reigns. Although Henry IV never became a sedentary or reflective man, his remarkable political insight enabled him to govern efficiently. His interest in maritime and colonial expansion led him to support Samuel de Champlain's exploration in Canada.

Achievement of the reign

Henry's folly in affairs of the heart continued to stand in contrast to his wisdom in affairs of state. He finally

grew tired of the malicious and scheming Henriette d'Entragues, now marquise de Verneuil, or perhaps he was motivated by a predilection for young girls: at the age of 55, he became infatuated with Charlotte de Montmorency, then 15 years old, the daughter of his close friend the constable of France. Henry conceived of marrying Charlotte to his cousin the young Condé, Henri II, who was supposedly not interested in women. Once married, however, the Prince prevented the King from seeing his wife by taking her to Brussels, the capital of the Spanish Netherlands. Henry IV became furious. The argument made by many historians that the war he then prepared to launch against the Habsburgs had its cause in this unfortunate love affair is, however, unconvincing. Difficulties had arisen with the Clèves–Jülich succession. After some hesitation, Henry finally decided on a military expedition to expel the imperial troops from Jülich. He had long decided on May 19, 1610, as the date of his departure to take command of the army of the east, one of three armies he had formed for this expedition to the Rhine. Early on the afternoon of May 14, he entered his coach with the intention of visiting the ailing Sully. The coach had not gone far from the Louvre when a traffic congestion in the narrow Rue de la Ferronnerie obliged the coachman to slow down. Suddenly a certain François Ravaillac, a fanatic, dashed out, leaped onto Henry's carriage, and stabbed the King twice with a long knife. Henry died on the way back to the palace.

Henry IV died a victim of the fanaticism he wanted to eradicate. Centuries ahead of his own time, he said, "Those who follow their consciences are of my religion, and I am of the religion of those who are brave and good." Too often misunderstood during his lifetime, his tragic end seemed finally to have opened the eyes of his people. They soon bestowed on him the appellation Henry the Great.

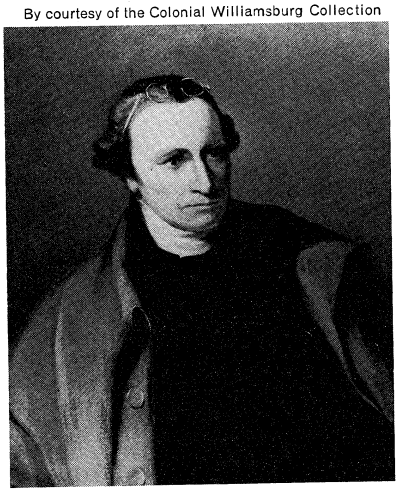
BIBLIOGRAPHY. After those of Napoleon I, the Revolution, and Louis XIV, the bibliography concerning Henry IV is the most abundant of any in French history, to the point that it fills an entire volume of the classic work of HENRI HAUSER, *Les Sources de l'histoire de France au XVIe siècle*, vol. 4 (1915). PIERRE DE VAISSIERE, *Henri IV* (1928), is a study that remains today the best informed in its entirety; it may be supplemented by the recent work of the DUC ANTOINE DE LEVIS-MIREPOIX, *Henri IV, roi de France et de Navarre* (1971), a brilliant and lively evocation of the personality and of the great events of his life and reign; and that of ROLAND MOUSNIER, *L'Assassinat d'Henri IV, 14 mai 1610* (1964). Among the testimonies of contemporaries are: P. DE L'ESTOILE, *Mémoires-Journaux*, 12 vol. (Brunet edition, 1875–96), by a Parisian, curious and informed about everything, who has left us the richest treasure of anecdotes and miscellaneous facts relating to Henry IV; the *Chronologie septenaire* (1605) and the *Chronologie novenaire*, 3 vol. (1608), by P.V. PALMACAYET, who had been the Calvinist minister attached to the household of Catherine de Bourbon, the sister of Henry IV, and knew the prince well; J.A. DE THOU, *Histoire universelle . . .*, 16 vol. (1734), which continues to be the most serious and the most fruitful source for works of all kinds about Henry IV; and RAYMOND RITTER (ed.), *Lettres du cardinal de Florence sur Henri IV et sur la France (1596–1598)* (1955), the day-to-day notes, valuable because of their sincerity, of a Medici, a papal legate, who understood and defined the exceptional merits of the monarch. Particular points of the political history of the reign are discussed in MARTIN PHILIPPSON, *Heinrich IV and Philipp III*, 3 vol. (1870–73); L. ANQUEZ, *Henri IV et l'Allemagne* (1887); E. ROTT, *Henri IV, les Suisses et la haute Italie* (1882); and L.A. PREVOST-PARADOL, *Elisabeth et Henri IV (1595–1598)* (1855), which may be supplemented by J. NOUAILLAC, *Villeroy* (1909). More instructive are the writings of Henry himself in the *Recueil des lettres missives de Henri IV*, 9 vol. (1843–76), by M. BERGER DE XIVREY (finished by J. GUADET), and the numerous documents of the same nature that have followed and continued since the last quarter of the 19th century. Little has been published on Henry IV in English; however, HESKETH PEARSON, *Henry of Navarre: The King Who Dared* (British title, *Henry of Navarre: His Life;* 1963), is of interest.

(R.Ri.)

Henry, Patrick

Patrick Henry was one of the outstanding orators and civil leaders of the American Revolutionary period. He was largely responsible for the passage of the Bill of Rights—the first ten amendments to the Constitution, guaranteeing individual rights and limiting the power held by government over individuals.

By courtesy of the Colonial Williamsburg Collection



Patrick Henry, portrait by Thomas Sully, 1815. In the Colonial Williamsburg Collection.

Patrick Henry was born at Studley, Hanover County, Virginia, on May 29, 1736, the son of John Henry, a well-educated Scotsman who served in the colony as a surveyor, colonel, and justice of the Hanover County Court. Before he was ten, Patrick received some rudimentary education in a local school, later reinforced by tutoring from his father, who was trained in the classics. As a youth, he failed twice in seven years as a storekeeper and once as a farmer; and during this period he increased his responsibilities by marriage, in 1754, to Sarah Shelton. The demands of a growing family spurred him to study for the practice of law, and in this profession he soon displayed remarkable ability. Within a few years after his admission to the bar in 1760 he had a large and profitable clientele. He was especially successful in criminal cases, where he made good use of his quick wit, his knowledge of human nature, and his forensic gifts.

Meanwhile, his oratorical genius had been revealed in the trial known as the Parson's Cause (1763). This suit grew out of the Virginia law, disallowed by King George III, that permitted payment of the Anglican clergy in money instead of tobacco when the tobacco crop was poor. Henry astonished the audience in the courtroom with his eloquence in invoking the doctrine of natural rights, the political theory that man is born with certain inalienable rights. Two years later, at the capitol in Williamsburg, where he had just been seated as a member of the House of Burgesses (the lower house of the colonial legislature), he delivered a speech opposing the British Stamp Act. The act was a revenue law requiring certain colonial publications and documents to bear a legal stamp. Henry offered a series of resolutions asserting the right of the colonies to legislate independently of the English Parliament, and he supported these resolutions with great eloquence. "Caesar had his Brutus, Charles the First his Cromwell, and George III . . ." Here he was interrupted by cries of "Treason! treason!" But he concluded, according to a likely version, ". . . may profit by their example. If *this* be treason, make the most of it."

During the next decade Henry was an influential leader in the radical opposition to the British government. He was a member of the first Virginia Committee of Correspondence, which aided intercolonial cooperation, and a delegate to the Continental congresses of 1774 and 1775. At the second Virginia Convention, on March 23, 1775, in St. John's Church, Richmond, he delivered the speech that assured his fame as one of the great advocates of liberty. Convinced that war with Great Britain was inevitable, he presented strong resolutions for equipping the Virginia militia to fight against the British and defended them in a fiery speech with the famed peroration, "I know

Speech at Virginia Convention

not what course others may take, but as for me, give me liberty or give me death."

The resolutions passed, and Henry was appointed commander of the Virginia forces, but his actions were curbed by the Committee of Safety; in reaction, he resigned on February 28, 1776. Henry served on the committee in the Virginia Convention of 1776 that drafted the first constitution for the state. He was elected governor the same year, and was re-elected in 1777 and 1778 for one-year terms, thereby serving continuously as long as the new constitution permitted. As wartime governor, he gave Gen. George Washington able support, and during his second term he authorized the expedition to invade the Illinois country under the leadership of George Rogers Clark.

After the death of his first wife, Henry married Dorothea Dandridge and retired to life on his estate in Henry County. He was recalled to public service as a leading member of the state legislature from 1780 to 1784 and again from 1787 to 1790. From 1784 to 1786 he served as governor. He declined to attend the Philadelphia Constitutional Convention of 1787 and in 1788 was the leading opponent of ratification of the U.S. Constitution at the Virginia Convention. This action, which has aroused much controversy ever since, resulted from his fear that the original document did not secure either the rights of the states or those of individuals, as well as from his suspicion that the North would abandon to Spain the vital right of navigation on the Mississippi River.

Bill of Rights

Henry was reconciled, however, to the new federal government, especially after the passage of the Bill of Rights, for which he was in great measure responsible. Because of family responsibilities and ill health, he declined a series of offers of high posts in the new federal government. In 1799, however, he consented to run again for the state legislature, where he wished to oppose the Kentucky and Virginia resolutions, which claimed that the states could determine the constitutionality of federal laws. During his successful electoral campaign, he made his last speech, a moving plea for American unity. He did not take his seat, for he died on June 6 at Red Hill, his last home, near Brookneal, Virginia.

BIBLIOGRAPHY. ROBERT D. MEADE, *Patrick Henry*, 2 vol. (1957–69), a full, authoritative treatment; WILLIAM WIRT HENRY, *Patrick Henry: Life, Correspondence and Speeches*, 3 vol. (1891, reprinted 1969), still valuable; MOSES COIT TYLER, *Patrick Henry* (1915, reprinted 1966); GEORGE MORGAN, *The True Patrick Henry* (1907).

(R.D.M.)

Henry the Lion, Duke of Saxony

Henry the Lion, duke of Saxony from 1142 and of Bavaria from 1156 to 1180, was the most powerful German prince of his time, apart from the Holy Roman emperor Frederick I Barbarossa. His aim in life was above all to build a large territorial state in northern Germany extending from the Rhine to the Baltic Sea. But the almost kingly position he achieved during his reign menaced the feudal order of the Holy Roman empire and led to a rupture with Frederick Barbarossa, who had long been an advocate of Henry's policies. Henry's eventual downfall led to the territorial dissolution of northern Germany.

The only son of Henry the Proud, duke of Saxony and Bavaria, and Gertrude, the daughter of the emperor Lothair III, Henry the Lion was born in 1129 or 1130. The place and the exact date of his birth are unknown. In May 1142 he recovered Saxony, one of the two duchies of which his father had been divested by Conrad III, the first Hohenstaufen German king. In 1147 Henry laid claim to Bavaria, which Conrad III had granted to Henry II Jasomirgott, margrave of Austria, and, in 1151, he tried in vain to take possession of the duchy. In 1147 or 1148 he married Clementia, the daughter of Conrad, duke of Zähringen, but this marriage was dissolved in 1162.

When Frederick I Barbarossa of Hohenstaufen, his cousin, was elected king of Germany in 1152, the Hohenstaufen made peace with the rival dynasty, the Welfs. In 1154 Frederick granted Henry the right to invest the bishops of the new bishoprics beyond the Elbe and also recognized his territorial claims to Bavaria. In September 1156 Henry secured possession of the duchy at Ratisbon; Austria was subsequently separated from Bavaria and given to Henry Jasomirgott and elevated into a duchy.



Henry the Lion, sandstone figure from his tomb, 1227. In the Cathedral of St. Blasius, Brunswick.
Foto Marburg

Alliance with Frederick Barbarossa

Henry, in turn, for 20 years supported Frederick Barbarossa. He accompanied him with a large army on his first Italian campaign (1154/1155) and, after Frederick's coronation as emperor, suppressed a rising of the Romans. In 1157 he took part in Frederick's expedition against the Poles. During Frederick's second Italian campaign, Henry provided valuable assistance to the Emperor at the siege of Crema in 1160 and in the war against the Milanese cities in 1161.

One year after recovering Bavaria, he laid the foundations of the city of Munich by establishing a new market on the Isar River. But his main effort was directed toward expanding the Duchy of Saxony, especially in the lands beyond the Elbe. In 1159 he refounded the city of Lübeck on territory he had taken from Adolf II, count of Holstein, who had first founded Lübeck in 1143. By treaties with the merchants of Gotland and the princes of Sweden and Novgorod, he considerably enhanced Lübeck's position as a commercial centre. In 1160 the bishopric of Oldenburg was also transferred to that city. From 1158 on Henry had subdued the Slavic Obodrites in several expeditions, extending his power all over Mecklenburg and thus opening the way for its Christianization and colonization.

In 1160 Schwerin became the seat of the bishopric of Mecklenburg and was granted the privileges of a city. Even the princes of western Pomerania temporarily acknowledged his feudal sovereignty. When Valdemar I, king of Denmark, conquered the island of Rügen, in the Baltic Sea, a long, drawn-out struggle broke out between him and Henry that lasted until 1171, when the dispute was settled and Henry's daughter married Valdemar's son.

In those years Henry also consolidated his position in Saxony by seizing the properties of several extinct dynasties without regard to the hereditary claims of other families. He made Brunswick his capital and in front of the castle he had built he erected the statue of a lion as a symbol of his family and a sign of his sovereignty. But Henry's arrogant nature and his propensity for aggrandizement evoked growing opposition. Beginning in the middle 1150s, several Saxon princes entered into alliances against him. Ten years later, a great coalition led by Al-

bert I the Bear, margrave of Brandenburg, and the Archbishop of Cologne posed a serious threat to him. It was only after the Emperor intervened in 1168 that peace was restored in Saxony.

At that time, Henry was at the zenith of his power. In early 1168 he married Matilda, the daughter of Henry II of England, and soon afterward was sent to France and England as ambassador of Frederick I on a mission to arrange an armistice between both nations. In 1172 he went on a pilgrimage to Jerusalem with a large following and was received with great ceremony by the Byzantine emperor Manuel I Comnenus at Constantinople (now Istanbul).

When in 1176 Frederick Barbarossa asked for support against the Lombard cities in northern Italy, Henry's price for aiding the emperor was the important imperial city of Goslar, together with its silver mines. But Frederick refused to cede it, and his old alliance with Henry came to an end.

Henry's decline

When fighting broke out again in Saxony in 1177, Frederick, after his return to Germany in 1178, instituted proceedings based on the charge of the Saxon nobles against Henry for breach of the king's peace. Henry, who had refused to answer the charges in the King's court, was deprived of his two duchies and of all imperial fiefs, in 1180. The emperor then proceeded to break up Henry's former domain. In the same year the Saxon duchy was divided into two parts. The lands of the two bishoprics of Cologne and Paderborn were given to the Archbishop of Cologne as the new Duchy of Westphalia; the eastern part of Saxony was given as a fief to a son of Albert the Bear of Brandenburg. The Duchy of Bavaria was granted to an ally of Frederick's, Otto von Wittelsbach.

Henry was at first able to maintain his position against Barbarossa in northern Saxony, but in the summer of 1181 he had to submit. Allowed to retain his hereditary lands of Brunswick and Lüneburg, he was exiled for several years to the court of his father-in-law, Henry II of England. On his return in 1185 he tried to regain his influence in Saxony. For his refusal to participate in the Third Crusade or to renounce his claims to Saxony, he was again banished, in 1189, rejoining Henry II in Normandy.

After Frederick Barbarossa's death in 1190, Henry returned once more to Saxony. King Henry VI of Germany now took the field against him but made peace with him at Fulda in Judy 1190. After Henry the Lion renewed the fighting during Henry VI's campaign in Italy, the Emperor and Henry became reconciled at a meeting in 1194. On August 6, 1195, Henry the Lion died in Brunswick. He was buried in the cathedral he had built there, at the side of his wife.

BIBLIOGRAPHY. A.L. POOLE, *Henry the Lion* (1912), the only biography in English; KARL JORDAN (ed.), *Die Urkunden Heinrichs des Löwen* (1949–57), a complete collection of documents and letters; *Die Bistumsgründungen Heinrichs des Löwen* (1939, reprinted 1962), a treatment of Henry's policies in the lands beyond the Elbe; THEODOR MAYER, *Friedrich I. und Heinrich der Löwe* (1957), a good summary of the problems of the period; PETER MUNZ, *Frederick Barbarossa* (1969), the most recent treatment of Henry's times in English.

(K.J.)

Henry the Navigator

Henry, a prince of Portugal (1394–1460), later called by English historians "Henry the Navigator" because of his contribution to the science of navigation, sponsored the voyages of discovery that led to the foundation of the Portuguese empire overseas (see Portugal, History of). Under his auspices, the sailing vessel known as the Portuguese caravel was developed, the techniques of cartography were advanced, navigational instruments were improved, and commerce by sea was vastly stimulated. He was also the designer of a grand strategy—not to be brought to fulfillment until after his death—whereby Christian Europe outflanked the power of Islām by establishing contact with Africa south of the Sahara and with Asia.

Henry was born on March 4, 1394, at Porto, the third son of King John I (*q.v.*) and Philippa of Lancaster, the daughter of John of Gaunt of England. Henry and his older brothers, the princes of Duarte and Pedro, were educated under the supervision of their parents; they were taught soldiering, statecraft, and the appreciation of literature.

Capture of Ceuta

The starting point of Henry's career was the capture of the Moroccan city of Ceuta in 1415. According to Henry's enthusiastic biographer, Gomes Eanes de Zurara, the three princes persuaded their still-vigorous father to undertake a campaign that would enable them to win their knightly spurs in genuine combat instead of in the mock warfare of a tournament. King John consented and, with Ceuta in mind, began military preparations, meanwhile spreading rumours of another destination, in order to lull the Moroccan city into a feeling of false security.

Although a plague swept Portugal and claimed the queen as a victim, the army sailed in July 1415. King John found Ceuta unprepared, as he had hoped, and its capture unexpectedly easy. Though Zurara later claimed the principal role in the victory for Henry, it would seem

By courtesy of the Museu Nacional de Arte Antiga, Lisbon



Henry the Navigator, detail of a triptych attributed to Nuno Gonçalves, *c.* 1465–70. In the Museu Nacional de Arte Antiga, Lisbon.

that the experienced soldier-king actually directed the operation. That Henry distinguished himself, however, is indicated by his immediate appointment as governor of Ceuta, which did not require his permanent residence there but obligated him to see that it was adequately defended.

An emergency arose in 1418, when the Muslim rulers of Fez (Fès) in Morocco and the kingdom of Granada in Spain joined in an attempt to retake the city. Henry hastened to the rescue with reinforcements but on arrival found that the Portuguese garrison had beaten off the assailants. He then proposed to attack Granada, despite reminders that this would antagonize the kingdom of Castile, on whose threshold it lay. But his father, who had spent years fighting the attempts of the Castilians to annex Portugal, wanted peace with them and sent peremptory orders to return home.

As governor of Ceuta, Henry always had ships at his command and by 1418 had begun to sponsor voyages in a small way. In that year, two squires of his, João Gonçalves Zarco and Tristão Vaz Teixeira, rediscovered the islands of Porto Santo near Madeira and a little later Madeira itself, both of which had been visited by Genoese in the previous century.

Upon his return to Portugal, Henry had been made duke of Viseu and lord of Covilhã. In 1419 he retired

from the court and became governor of the Algarve, the southernmost province of Portugal. There, on the rocky promontory of Sagres, at the southwest tip of Portugal, he founded a small court of his own, to which he attracted seamen, cartographers, astronomers, shipbuilders, and instrument makers.

Maritime research at Sagres

In 1420, at the age of 26, he was made grand master of the Order of Christ, the supreme order sponsored by the pope, which had replaced the crusading order of the Templars in Portugal. While this did not oblige him to take religious vows, it did oblige him to dedicate himself to a chaste and ascetic life. (He had, however, not always refrained from worldly pleasures; as a young man he had fathered an illegitimate daughter.) The funds made available through the order largely financed his great enterprise of discovery, which also had as its object the conversion of the pagans to Christianity. It was for this reason that all of Henry's ships bore a red cross on their sails.

From 1420 onward he began to dispatch expeditions from the nearby port of Lagos, first with the aim of discovering more of the Moroccan Atlantic coast and later—when he began to think in terms of continents—with the aim of discovering the southerly route to India, in order to introduce Christianity there and to foster commerce. Little is known of the Prince's private life in the 1420s. Duarte and Pedro both married, but Henry remained single to the end of his life.

When Duarte succeeded King John in 1433, he did not hesitate to lecture and reprove Henry for such shortcomings as extravagance, unmethodical habits, failure to keep promises, and lack of scruples in the raising of money. This rebuke is not supported by the traditional account of the Navigator as a lofty, ascetic person, indifferent to all but religion and the furtherance of his mission of discovery.

Henry unquestionably was also—although in a different way—influenced by his older and perhaps more brilliant brother, Prince Pedro. In 1425 Pedro set out on a long tour of Europe on which he visited England, Flanders, Germany, Hungary, and the principalities of Moldavia and Walachia (now Romania) before returning home through Italy, Aragon, and Castile. In eastern Europe, he was close enough to Ottoman Turkey to appreciate the Muslim danger. The travels stimulated his interest in geography, which was further whetted in Italy, the home of most European travellers to distant parts. From Italy Pedro brought home to Portugal, in 1428, a copy of Marco Polo's travels that he had translated for Prince Henry's benefit.

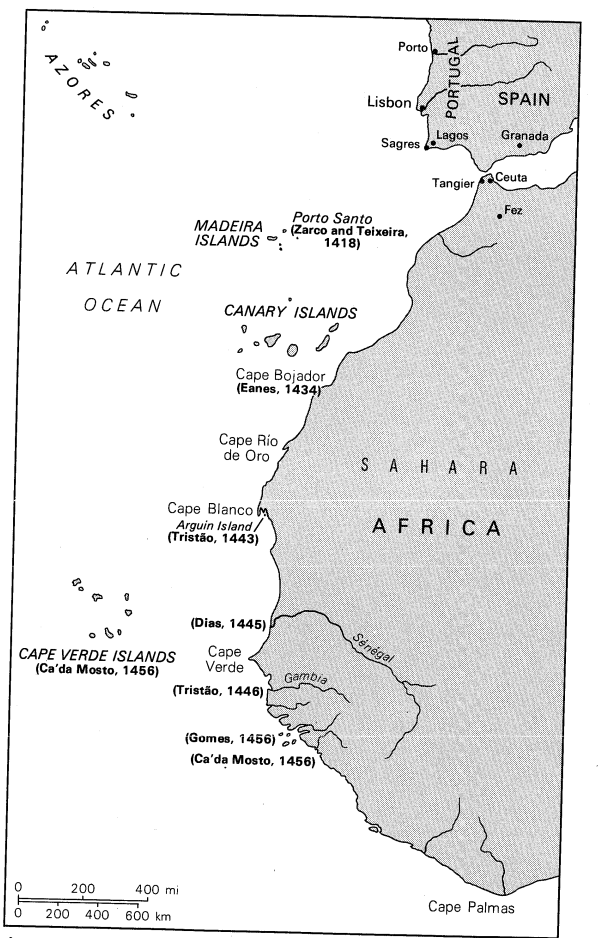
African exploration

During the five years of his brother Duarte's reign, Henry was able to persuade his captains to venture farther down the African coast. The most important achievement was the rounding of Cape Bojador by Gil Eanes in 1434, overcoming a superstition that had previously deterred seamen. During the next years, Henry's captains pushed southward somewhat beyond the Rio de Oro. They also began the colonization of the recently discovered Azores, through the orders of both Henry and Pedro.

In 1437 Henry and his younger brother, Fernando, gained Duarte's reluctant consent for an expedition against Tangier. Ceuta had proved an economic liability, and they believed that possession of the neighbouring city would both insure Ceuta's safety and provide a source of revenue. Pedro opposed the undertaking as he felt it meant deviation from Portugal's true mission, which to him was prosecution of further discovery. Henry and Fernando nevertheless attacked Tangier and met with disaster; Henry had shown poor generalship and mismanaged the enterprise. The Portuguese army would have been unable to re-embark had not Fernando been left as hostage. Henry offered himself as hostage, but as the army refused to lose its commander, Fernando remained in captivity to later die of ill treatment at Fez in 1443.

King Duarte died in 1438, shortly before Henry's return. His heir, Afonso V, was only six at the time, and Pedro assumed the regency over the bitter opposition of the boy's mother, Leonor of Aragon, who hated her brother-in-law and would willingly have accepted Henry. But Henry had no wish to govern Portugal and attempted unsuccessfully to bring about peace in the family. He felt satisfied with Pedro as regent and for himself wished only to return to Sagres and resume his maritime work. The queen-mother somewhat eased matters by leaving the country, and for most of the next decade Pedro and Henry worked in harmony, though their illegitimate half brother, Afonso, count of Barcelos, dissatisfied with his inferior position in the family, attempted to sow discord and eventually succeeded.

During these years, Henry's mission of discovery, encouraged and aided by the regent, progressed rapidly. One of his immediate aims was to find an African gold supply—the existence of which he is thought to have learned from the Moors of Ceuta—to strengthen the Portuguese economy and to make the voyages pay for themselves. In 1441 a caravel returned from the West African coast with some gold dust and slaves, thus silencing the growing criticism that Henry was wasting money on a profitless enterprise. One of Henry's voyagers, Dinís Dias, in 1445 reached the mouth of the Sénégal (then taken for a branch of the Nile); and a year later Nuno Tristão, another of Henry's captains, discovered the Gambia River. By 1448 the trade in slaves to Portugal had become sufficiently extensive for Henry to order the building of a fort and warehouse on Arguin Island; this installation was, in fact, the first European trading post established overseas.



Areas reached by explorers under the sponsorship of Henry the Navigator.

Afonso V attained his legal majority at the age of fourteen in 1446. His embittered mother had meanwhile died in Castile, and although the young king presently married Pedro's daughter, Isabel, his relations with the regent were nonetheless bad. Afonso of Barcelos now came to work on the boy's susceptible mind. His task was ren-

dered easier by the obvious reluctance with which Pedro turned full power over to the youth, whose weaknesses were already apparent.

Henry, who wished only to be a peacemaker, left Sagres and tried, unsuccessfully, to establish harmony between his brother Pedro and his nephew King Afonso. Armed conflict between the two became inevitable, and Henry in the end felt obliged to side with the King, though he remained as much as possible in the background. He took no part in a skirmish at Alfarrobeira in May 1449, in which Pedro was killed by a chance shot from a crossbowman. There is reason to believe that after this sad termination of the family feud, Henry wished to go into exile at Ceuta and spend his remaining days fighting Moors but that the King refused him permission. A historian writing 50 years later gave the impression that Henry had deserted his brother when he might have saved him. Henry's biographer, Zurara, on the other hand, declared that his hero did everything possible to prevent Pedro's death and promised to explain the circumstances further in later writings; but if he did so, the account is lost.

Final maritime ventures

After Alfarrobeira, Henry spent most of his time at Sagres, though he did not altogether abandon public life. He was accorded by the King the sole right to send ships to visit and trade with the Guinea coast of Africa. He appeared occasionally at the Lisbon court and in 1450 helped arrange for the marriage of the King's sister, to the emperor Frederick III. During most of his last decade, Henry concentrated on the sponsorship of voyages. These accomplished only minor discoveries as the Prince now seemed mainly interested in trade with the regions already contacted. The last two important mariners sent out by Henry were the Venetian Alvise Ca' da Mosto (Cadamosto) and the Portuguese Diogo Gomes, who between them discovered several of the Cape Verde Islands.

The farthest point south along the African coast reached during Henry's lifetime is generally considered to have been Sierra Leone, though one piece of evidence suggests that his seamen progressed to Cape Palmas (off the Ivory Coast), some 400 miles beyond. So great was his investment in exploration that, despite his great revenues, Henry died heavily in debt.

Afonso V had small interest in discovery but great zeal for crusading and knight-errantry. Resuming the old attempt at Moroccan conquest, he led an expedition in 1458 against Alcácer Ceguer (now Ksar es-Shrhir), in which Henry accompanied him. The Prince, now 64, did well in the fighting; and when the town capitulated, Afonso left the surrender terms to his uncle, who showed remarkable leniency.

Henry lived for two years after his final return from Morocco, dying at Sagres on Nov. 13, 1460. The cause of his death is not recorded, but he must have known beforehand that the end was near, having spent the previous months putting his affairs in order. He was buried in the royal vault at Batalha.

The surname Navigator, applied by the English to the prince, though seldom by Portuguese writers, is a misnomer as he himself never embarked on voyages of discovery. His fame rests primarily on his patronage of navigators, for which he is rightly regarded as the initiator of the great age of discovery and the European thrust towards world dominion.

BIBLIOGRAPHY. Although GOMES EANES DE ZURARA is no longer considered an infallible authority on Prince Henry, his narrative of the discoveries, trans. as *Chronicle of the Discovery and Conquest of Guinea* by C.R. BEAZLEY and E. PRESTAGE (1896), is still the best. Lives of Henry in English include: R.H. MAJOR, *The Life of Prince Henry of Portugal, Surnamed the Navigator* (1868, reprinted 1967); E. SANCEAU, *Henry the Navigator: The Life of a Great Prince and his Times* (1947); and E.D.S. BRADFORD, *A Wind from the North: The Life of Henry the Navigator* (1960). Major and Bradford are almost entirely concerned with discovery; Sanceau provides a real biography. E. PRESTAGE, *The Portuguese Pioneers* (1933, reprinted 1966), though dealing with all Portuguese discovery, offers much useful information on Henry's life. F.M. ROGERS, *The Travels of the Infante Dom Pedro of Portugal* (1961), is concerned mostly with the older Prince but stresses strongly his relations with Henry. J.P. OLIVEIRA MARTINS, *The Golden Age of Prince Henry the Navigator* (Eng. trans. 1914), reviews the lives of Henry and his brothers until the death of Pedro. Background material is furnished in abundance by DAMIAO PERES *et al.*, in *História de Portugal: Edição Monumental*, vol. 2 (1931).

(C.E.No.)

Hepatopsida

Liverworts are low, mosslike plants that constitute the class Hepatopsida, about 300 genera and 10,000 species, of the division Bryophyta. Almost totally lacking in direct economic significance, liverworts do have an important role in nature: providing food and cover for minute animals and facilitating the decay of logs and the fragmentation of rocks by virtue of their ability to retain moisture. They also chemically etch rocks and thus promote a slow disintegration. Before 1900 the flattened liverworts were commonly accepted as medicinal herbs since, according to a theory known as the Doctrine of Signatures, the liver-like form of certain species indicated that they should be useful in the treatment of liver ailments. Liverworts usually are among the first plants to colonize denuded soil and often act to reduce erosion. Because of their clear response to changes in environment, they are frequently used as experimental materials in studying the effect of chemical and other factors on the growth process. They also have an aesthetic value, covering what otherwise might be drab components of the landscape.

General features. Liverworts range in size from minute green threads barely visible to the naked eye to pendant ribbonlike forms over 20 centimetres (eight inches) long and almost five centimetres (two inches) wide. The small, filamentous species, although provided with minute leaves, are relatively simple; some of the larger, flattened liverworts, however, are highly differentiated internally, with distinct layers of tissues in the leaflike body, or thallus. Most of the leafy liverworts are intermediate in size, provided with two lateral rows of larger leaves (phyllids) and a third, ventral row of smaller (rarely equal or sometimes lacking) leaves (amphigastria).

Distribution

The liverworts occur from the Arctic to the Antarctic; a few can live in desert habitats, and some are adapted to aquatic environments. The largest number of species are found in the moist tropics, where they may clothe the trunks of shrubs and trees or hang from branches like pendant beards.

Life cycle. The liverworts have a well-marked alternation of generations, in which sexual plants (gametophytes) bearing eggs or sperm (or both) alternate with asexual plants (sporophytes) containing spores capable of producing new gametophytes. Thus, a gametophytic generation alternates with a sporophytic one.

The male and female sex organs (antheridia and archegonia, respectively), depending upon the particular species, are borne either on fleshy stems or on the leafy surface of the plant. Some species have both kinds of sex organs on the same plant, but in species such as *Marchantia polymorpha* these organs are borne on different plants.

The eggs are borne in the basal portion of the flask-shaped archegonium and the sperm in the spherical or sausage-shaped antheridium. A moist habitat is required not only for the growth of the plants but also for the rupturing of both antheridia and the necks of the archegonia. The sperm swim by means of two tiny, whip-like cilia from one plant to another through a thin film of water.

Sperm enter the ruptured tip of the archegonium and move down the canal of the slender neck, and one sperm eventually may reach the egg and unite with it to form a zygote, or initial cell of the embryo. The young sporophyte that results remains within the slowly enlarging archegonium throughout most of its development. At maturity the sporophyte consists mainly of a sporangium, or spore capsule, a foot embedded in the base of the archegonium, and a compressed stalk (seta) between; in

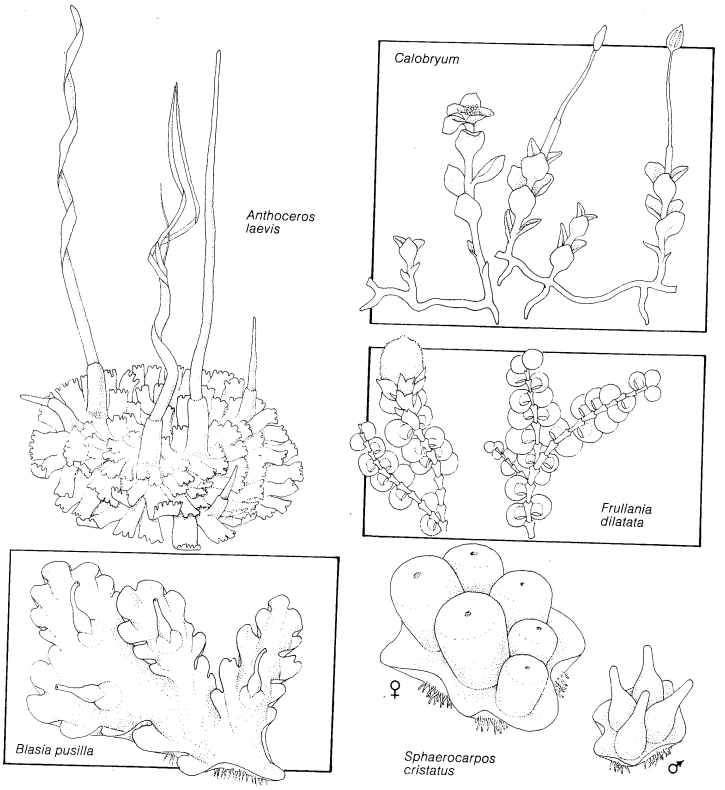


Figure 1: Diversity among liverworts. Symbols: ♀—female, ♂—male.

Drawing by M. Moran based on *An Evolutionary Survey of the Plant Kingdom* by Robert F. Scagel, Robert J. Bandoni, Glenn E. Rouse, W.B. Schofield, Janet R. Stein, and T.M.C. Taylor. © 1965 by Wadsworth Publishing Company, Inc., Belmont, California 94002. Reprinted by permission of the publisher.

most liverworts the seta elongates and carries the capsules and their spores upward. Upon being dispersed by water or wind, a spore may germinate into a very short, threadlike or platelike structure (protonema) that grows into the gametophyte.

While sexual reproduction does occur in liverworts, asexual reproduction probably accounts for most liverwort propagation. Under suitable environmental conditions, almost any living fragment seems capable of growing directly or indirectly into a new gametophyte. Many species of liverworts produce a specialized cell or group of cells that serves as a reproductive structure called a gemma. Several gemmae may be borne on the surface of stems or in specialized flask-shaped or cup-shaped structures. The gemmae may be dispersed by rain as well as by wind.

Liverworts, like higher plants, contain growth-affecting substances and respond to them. This may account for the great variability within a species in response to light, water, and other environmental factors. Plants in environments with diffuse light or excessive moisture may be much elongated and slender, in contrast with the compact nature of those in the more usual habitat. These variations often make it difficult to distinguish between related but valid species if each is collected in a different habitat.

Form and function. *Marchantia polymorpha*, a liverwort commonly discussed in textbooks, has a ribbon-like gametophyte (thallus) with an upper and a lower epidermis. The lower epidermis is provided with two kinds of hairlike cells (rhizoids) that serve to anchor the plant (absorption seems to take place throughout the entire surface of the plant). The upper epidermis has pores—surrounded by rigid, chimney-like tiers of cells—that open into chambers below. These cavities are separated by vertical, single-layered walls and within have cactus-like columns of green cells. With the exception of the rhizoids, all parts of liverworts contain the green pigment bodies, or chloroplasts, and are capable of photosynthesis.

In *Marchantia*, the sex organs are borne on highly modified upright branches of the thallus. The antheridia are borne in cavities on the flattened, upper surface of stalked heads. The archegonia and, consequently, the developing sporophytes are found beneath the arms of a stalked structure that resembles an open umbrella stripped of its fabric. The thalloid forms most related to the leafy liverworts are more simple in structure; some are only one cell in thickness.

The leafy forms of liverworts have little internal differentiation, but occasionally their leaves (phyllids) exhibit some bizarre forms, being composed of branched, hairlike segments in *Trichocolea* and having a pouch-shaped lobule bent up behind the major lobe of the leaf in *Frullania*.

Most liverworts, in addition to chloroplasts and a nucleus in each cell, have oil bodies that are often distinctive and therefore useful in making specific identifications and in determining relationships that exist among species. Oil bodies

Evolution and classification. The variations in the structure of the liverwort gametophyte are extreme and give rise to considerable disagreement as to the origin of and evolution within the liverworts. Moreover, liverworts, because of their delicate tissues, do not preserve well in geological materials, and so known fossils are few and inadequate to suggest ancestry with any degree of security. The complex thalloid forms are thought by some botanists to be more primitive but more increasingly by others to represent the most recent and highly derived type.

Distinguishing taxonomic features. The most important features in arranging the class Hepatopsida into its two subclasses (Hepatidae and Anthocerotidae) are the basic form of the gametophyte (either thalloid or leafy) and the growth pattern of the sporophyte (apical versus basal). Orders are determined primarily by the external appearance of the gametophyte and the position of the sex organs. Microscopic anatomy is a further aid to proper taxonomic ranking at the ordinal level.

Annotated classification. The classification given below is an adaptation of a currently popular scheme.

CLASS HEPATOPSIDA (liverworts proper)
Gametophytes thalloid or "leafy" with a high degree of dorsiventrality; if "leafy," with two lateral rows and sometimes a smaller, ventral row of "leaves." Sporophytes with globular or needle-shaped capsules; capsules with relatively undifferentiated walls, opening by splitting longitudinally or

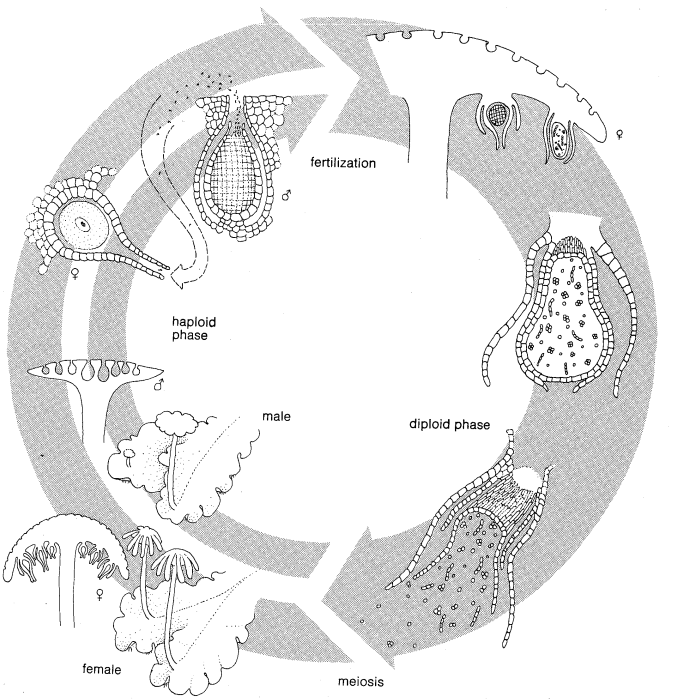


Figure 2: Life cycle of representative liverwort (*Marchantia*).

rarely by irregular erosion, without a central sterile pillar (columella) of cells, except in the Anthocerotidae, often with hygroscopic, filamentous sterile cells (elaters) between the spores. About 300 genera and 10,000 species.

Subclass Hepatidae

Gametophyte, seldom erect, more frequently leafy than thalloid, with several chloroplasts in each cell; the sporophyte with apical growth that produces a globular capsule without columella or stomates.

Order Takakiales

Unusually strange liverworts at first suspected of being algae; in the field they superficially resemble a minute green *Chara*. Archegonia often are borne on a pedestal and always without attending bracts. Antheridia, sporophytes, and rhizoids are unknown. It is represented by a single family and genus, in which 2 species, *Takakia lepidozioides* and *T. ceratophylla*, have been described. The nuclei each have 4 chromosomes, an unusually low number.

Order Calobryales

Gametophytes with radial symmetry superficially resembling mosses. Upright branches arise from buried, rhizome-like stems; rhizoids are totally absent. The archegonia have 4 vertical rows of neck cells, which is less than that in all other groups of the bryophytes. The sporophyte is of the liverwort type but is not basally surrounded by a true perianth. A small order of 1 or 2 genera and about 7 species.

Order Jungermanniales

Gametophytes consisting of 2 types as indicated below. Internal anatomy relatively simple. Capsule usually exserted on an elongated, fragile stalk (seta). The largest order of liverworts, with about 200 genera and over 8,000 species.

Suborder Jungermannineae (leafy liverworts). Gametophytes organized dorsiventrally, with 2 rows of conspicuous but delicate lateral "leaves" and often a row of usually smaller leaves underneath. External form highly variable. The archegonia and sporophytes are borne terminally on stems and branches. About 180 genera and 7,500 species are included.

Suborder Metzgerineae. Gametophytes thalloid, usually 1 cell in thickness except for the midrib (if the wings are more thickened, the midrib is absent or obscure); margins are sometimes incised or lobed, superficially resembling leaves. Archegonia and sporophytes not terminal. About 20 genera and 550 species are included.

Order Sphaerocarpales

Gametophytes small, superficially resembling fern prothallia or forms of members of the Metzgerineae. Sex organs are borne on the surface but in specialized containers. A small order consisting of 2 genera and about 20 species.

Order Marchantiales

Gametophytes dorsiventrally oriented, thalloid, multicellular in thickness with considerable internal differentiation. About 95 genera and 2,000 species.

Subclass Anthocerotidae (horned liverworts, or hornworts)

Gametophytes always thalloid with a single chloroplast per cell; sporophyte with basal growth that produces a needle-shaped capsule with a columella, multicellular elaters, and stomates. A relatively small subclass consisting of not more than 5 genera but of great interest scientifically because of the usually single, large chloroplast in each cell, the basal growth of the sporophyte, and the presence of stomates.

Critical appraisal. The order Takakiales exhibits enough differences from most bryophytes to cause some botanists to question its placement in the division. These differences include the small chromosome number of four per cell (the lowest known among the land plants) and the bractless chlorophyll-containing archegonium, which looks as if it may have been formed from a developing leaf.

The Anthocerotidae have been variously treated taxonomically. Some botanists feel that their slender, needle-shaped sporophytes with basal growth and stomates and the few (often one) large chloroplasts in each cell entitles them to class rank equivalent to Hepatopsida. Their gametophytes, however, are similar to some of the thalloid liverworts, and, until more convincing evidence is forthcoming, its seems unwise to separate them from the Hepatopsida.

BIBLIOGRAPHY. R.M. SCHUSTER, *The Hepaticae and Anthocerotae of North America, East of the Hundredth Meridian*, 2 vol. (1966–69), a scholarly treatise with bibliographies to classic and contemporary works on liverworts; for identification of American species of liverworts, see H.S. CONARD, *How to Know the Mosses and Liverworts* (1956); T. FRYE and L. CLARK, *Hepaticae of North America* (1937); and S.M. MACVICAR, *The Student's Handbook of British Hepatics* (1926), still a useful British field manual despite its age.

(A.J.Sh.)

Heraclius

East Roman (Byzantine) emperor from AD 610 to 641, Heraclius revived the decaying power of the Byzantine state. Most scholars now agree that it was he who initiated the militarization of Anatolia, known as the theme (military district) system, which was further developed and extended by his successors. It placed the Anatolian provinces under military governors and provided land grants not only for frontier soldiers but also for soldiers and peasants in the interior, on condition of hereditary military service. The effects were revolutionary. A flexible defense developed in depth; a free, militarized peasantry arose; agriculture revived; and the state was relieved of much of the burden of soldiers' pay. Unruly mercenaries were largely replaced by native soldiers with a personal interest in protecting the empire. Indeed, during the four centuries that the theme system remained intact, the empire could be defeated but not conquered. Byzantium withstood Islām's fierce onslaughts and sheltered Europe's infant civilization. Yet there is no known record of the inception of the theme system, and it was rather because of his epic struggle against Persia and his recovery of wood believed to be from Christ's cross that Heraclius became a hero of medieval legend.



Heraclius, gold coin, *c.* 610–641. In the Dumbarton Oaks Research Library and Collection, Washington, D.C.
By courtesy of the Dumbarton Oaks Collection, Washington, D.C.

Heraclius was born *c.* 575 in Cappadocia, in eastern Anatolia. His father, probably of Armenian descent, was governor of the Roman province of Africa when an appeal came from Constantinople to save the empire from the terror and incompetence of the emperor Phocas. The governor equipped an expeditionary force and put his devout son, the blond and gray-eyed Heraclius, in command of it.

In October 610 Heraclius dropped anchor off Constantinople, deposed Phocas, and was crowned emperor of a crumbling state, occupied by invaders and wracked with internal dissension. Slavs swarmed over the Balkan Peninsula. The Persians occupied extensive parts of Anatolia. The Turkic Avars, who ruled over the Slavic and other tribes that occupied the region between the Don and the Alps, exacted tribute. With its treasury empty, its economy disrupted, its administration disorganized, its army depleted and demoralized, its factions engaging in civil strife, its peasants enfeebled by excessive exactions, its religious dissenters alienated by persecution, its authority challenged by a powerful aristocracy, the empire lacked the strength necessary to expel the invaders, and possibly even to survive.

Campaign against the Persians

In 614 the Persians conquered Syria and Palestine, taking Jerusalem and the cross, and in 619 occupied Egypt and Libya. In an effort to placate the Avars, Hera-

clius met them at Thracian Heraclea (617 or 619). They sought to capture him, and he rode madly back to Constantinople, hotly pursued. Overlooking their perfidy, he finally made peace with them and was free to take the offensive against the Persians.

In 622, clad as a penitent and bearing a sacred image of the Virgin, he left Constantinople, as prayers rose from its many sanctuaries for victory over the Zoroastrians, the recovery of the cross, and the reconquest of Jerusalem. He was, in effect, leading the first crusade. Indeed, in the ensuing hostilities, a pious poet contrasted the dancing girls in the Persian general's tent with the psalm singers in the Emperor's. In a brilliant campaign, he manoeuvred the Persians out of Anatolia and suggested a truce to the Persian monarch. This offer Khosrow II contemptuously rejected, referring to himself as beloved by the gods and master of the world, to Heraclius as his abject and imbecilic slave, and to Christ as incapable of saving the empire. Mindful of the propagandistic value of Khosrow's response, Heraclius made it public.

The next two years he devoted to campaigns in Armenia, the manpower of which was vital to the empire, and to a devastating invasion of Persia. In 625 Heraclius retired to Anatolia. He had encamped on the west bank of the Sarus River when the Persian forces appeared on the opposite bank. Many of his men rushed impetuously across the bridge and were ambushed and annihilated by the enemy.

Emerging from his tent, Heraclius saw the triumphant Persians crossing the bridge. The fate of the empire hung in the balance. Seizing his sword, he ran to the bridge and struck down the Persian leader. His soldiers closed rank behind him and beat back the foe.

In 626 the Persians advanced to the Bosporus, hoping to join the Avars in an assault on the land walls of Constantinople. But the Romans sank the primitive Avar fleet that was to transport Persian units across the Bosporus and repelled the unsupported Avar assault. Heraclius again invaded Persia and in December 627, after a march across the Armenian highlands into the Tigris plain, met the Persians near the ruins of Nineveh. There, astride his famous war-horse, he killed three Persian generals in single combat, charged into enemy ranks at the head of his troops, killed the Persian commander, and scattered the Persian host.

A month later, Heraclius entered Dastagird with its stupendous treasure. Khosrow was overthrown by his son, with whom Heraclius made peace, demanding only the return of the cross, the captives, and conquered Roman territory. Returning to Constantinople in triumph, he was hailed as a Moses, an Alexander, a Scipio. In 630 he personally restored the cross to the Church of the Holy Sepulchre in Jerusalem.

Since the 4th century, when Roman emperors adopted Christianity, they had endeavoured to preserve uniform theological belief and, notably in Egypt, Syria, and Armenia, had persecuted those with differing Christological views. The animosities thus created had facilitated the Persian conquest, and Heraclius sought to conciliate the dissenters with the doctrine of Christ's single will (monothelitism). He failed.

Rise of Islām

At all events, it was already too late. United by Islām, the Arabs swept out of their arid homeland into Syria (634). Broken in body and spirit by disease, by long years of the cares of state, and by the wounds and emotions of 100 battles, Heraclius did not take personal command of the army, although the sight of him in battle armour would have inspired the troops and silenced the bickering generals. The Byzantines were defeated in a great battle on the Yarmuk (636). Soon, Syria and later Egypt fell to the Arabs. Heraclius returned northward, bearing the "holy wood," once the object of his greatest glory, now the companion of his deepest sorrow. Fearing water, he remained a year on the Asiatic bank of the Bosporus before summoning the courage to cross to Constantinople on a pontoon bridge with foliage hiding the water.

Heraclius' first wife, Eudocia, had died in 612. A year later, he had married his niece Martina, thus offending the religious scruples of many of his subjects, who viewed his second marriage as incestuous and Martina as accursed. It was apparently a happy marriage, Martina accompanying him on his campaigns and bearing him nine children. During his last years, Heraclius seems to have suffered from enlargement of the prostate gland, retention of urine, and a consequent inflammation. After violent spasms, he died on February 11, 641, bequeathing the empire to his two elder sons, the consumptive Constantine III of his first marriage and Heracleonas, his son by Martina.

Assessment

Although Heraclius possessed a deep Christian faith and attributed his successes to God, the once widely accepted view of him as an inspired visionary, who was capable of supreme but spasmodic efforts and wondrous achievements when acting under divine promptings, would appear to be false.

No doubt he was an inspiring military leader who fired his army with religious fervour and whose personal intrepidity, imaginative tactics, and constant concern for his men evoked their love and loyalty. But he was also a cautious and calculating strategist who did not hesitate to employ religion to serve his military ends. Thus, when in 623 his victorious soldiers wanted to penetrate deeper into Persia, contrary to his plan to retire, he referred the matter to God. After his troops had fasted and prayed three days, he opened the Bible in their presence, apparently at random, and read a passage that could be interpreted only as a divine command to withdraw. Moreover, even though he fostered the crusading spirit, he waged war in a less inhumane manner than most of his contemporaries. He did not enslave or massacre the inhabitants of conquered towns and he treated his prisoners of war well, releasing them rather than butchering them when he could not feed them. His mercy contrasted sharply with Khosrow's acerbity and probably hastened his victory in Persia.

As a statesman it is also difficult to think of him as merely a religious fanatic. Certainly he inspired an oppressed and hopeless people with a new spirit of faith, service, and self-sacrifice but the man who restored a state that was sinking under the blows of internecine strife and foreign invasion and gave it the strength to withstand Islām's assaults for four centuries, perhaps even contributed to its survival until 1453, must have had a strong will, great organizing ability, exceptional conciliatory powers, and deep insight into the needs of both state and subjects. With a keen sense of reality, he adjusted the empire to the needs of the 7th century, departmentalizing the great state offices, instituting the theme system, and replacing Latin with Greek as the official language.

BIBLIOGRAPHY. FRANK THEISS, *Die griechischen Kaiser* (1959), has the most recent extensive treatment. ANGELO PERNICE, *L'Imperatore Eraclio* (1905), is scholarly and exhaustive. L. DRAPEYRON, *L'Empereur Héraclius et l'empire byzantin au 7e siècle* (1869); J.B. BURY, *A History of the Later Roman Empire*, vol. 2 (1889), are older works. Brief recent accounts are in GEORGE OSTROGORSKY, *Geschichte des byzantinischen Staates* (1965; Eng. trans., *History of the Byzantine State*, 2nd ed., 1968), with a discussion of sources; and ENNO FRANZIUS, *History of the Byzantine Empire* (1967), a narrative approach.

(E.F.)

Heraldry

Heraldry is the science and art that deals with hereditary symbols employed to distinguish individuals, institutions, and corporations. These symbols, which probably originated as identification devices on shields, are called armorial bearings. Heraldry treats of their use, display, and regulation.

Strictly defined, heraldry denotes that which pertains to the office and duty of a herald, an official who grew to be of considerable importance during the Middle Ages; that part of his work dealing with armorial bearings is properly termed armory. In general usage, however, the term heraldry has come to mean the same as armory.

From the second quarter of the 12th century in western

Europe, heraldic designs are found in general application. Elsewhere, a similar system is to be found only in Japan, in the *mon*, also dating from the 12th century. Other times and places are often said to have produced heraldic systems; for example, ancient Israel in the symbols of the 12 tribes, or the designs used by the Rajput princes in India. These and like instances, however, are more properly considered incipient heraldry, since they did not develop into the complex heraldic practice known in western Europe and Japan.

From 1150 to 1500, the use of heraldry in the West was utilitarian: on armour in warfare, and on seals in peace. In the latter part of that period, it was used in peaceful ways and had much artistic value. Also, because from the beginning the use of arms had been associated with the higher feudal castes, heraldry acquired in later medieval times an identification with the concept of gentility that has persisted. To bear arms was the mark of a gentleman; therefore, to possess the desirable quality of gentility, a man needed to have armorial bearings. The great majority of those who seek to use coats of arms in the late 20th century are actuated by this motive. In the use of corporate arms, the motive of prestige rather than social distinction operates. As long as the possession of arms confers any social distinction, arms will be sought and used. At no previous time has there been so widespread an employment of heraldic devices.

Seals as ancient devices of heraldry

The use of symbols has been universal among civilized communities and also has occurred among many that are semi-civilized. But these symbols have not assumed the character always associated with heraldry. Seals, too, which have a prominent place in heraldic practice, are of an antiquity approaching that of the most ancient civilizations. They were in use in the states that from Sumer onward flourished in Mesopotamia. Their use, for example, in the Babylonian Empire was the same as in medieval western Europe: to authenticate the documents (possibly of baked brick, later papyrus, later still parchment or vellum) on which they appeared or to which they were appended. All persons, literate and illiterate alike, were able to recognize the representation or symbol of a ruler or other potentate. In 12th-century Europe, heraldry first appeared on seals in the representations of persons. There is a clear line of descent from the seals of Assyria and Babylonia to the modern company seal, which is often heraldic.

Although originating in the small half continent of western Europe, heraldry has become universal, often, but not only, by way of western European colonization. Heraldry has spread to a considerable degree in both the Americas, Australia, New Zealand, and South Africa. In the former British India, the hereditary princes adopted the use of heraldry. In the numerous independent states formed in Africa from the former British colonies, official armorial bearings are generally used, and the same is true of the new states that were formerly French colonies. In Russia in the 18th century, the use of armorial bearings was adopted from the West, and state emblems are not unknown in Communist eastern Europe. In the 13th century, the Celtic princes of Wales and Ireland and the chiefs of the Scottish Highland clans took up the use of heraldic symbols from the example of the feudal lords and knights of other parts of Europe.

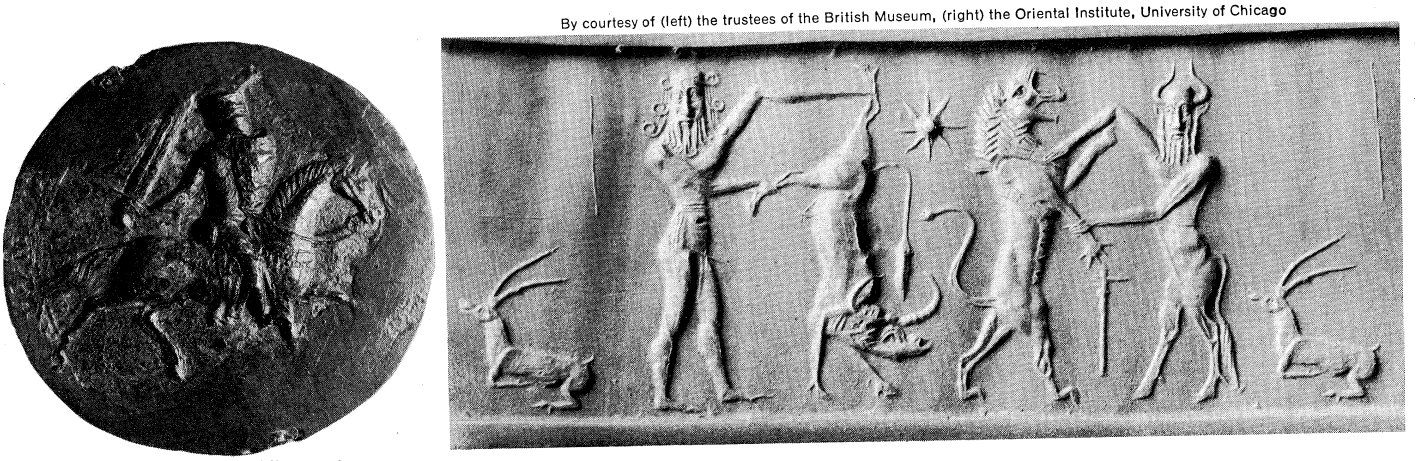
Other kinds of emblematic identification have some similarities with heraldry. An example is the totem system, found among the aborigines of America and Australia, in which an animal, plant, or other object serves as an emblem of family or clan and is often regarded as a reminder of its ancestry. Totemism varies greatly in different countries, as do the theories advanced to explain it. The totem poles used by the Indians of the northwest coast of North America contain a heraldic element in their use of a hereditary symbol for a family or tribe. They therefore come under the heading of approaches to heraldic designs and may be termed semi-heraldic in character.

The Japanese *mon* as a type of armorial bearing

The Japanese *mon* is very definitely a heraldic symbol, having many parallels in its use with the armorial bearings of Europe. It was used on helmets, shields, and breastplates but never, as in Europe, large enough to identify the wearer of the armour at any considerable distance. When identification was desired, the *mon* was displayed on flags. In some European languages, the *mon* has been translated erroneously as "coat of arms." The *mon* most closely resembles the heraldic badge (distinctive mark used by retainers), which in Europe often antedated armorial bearings. Further resemblances to European heraldry in the use of the *mon* include: the decorative use of the symbol on clothes, furniture, and houses; the use on the clothes of retainers of great lords; the legal requirement of registration of the *mon* (dating from the 17th century); and the reservation of the chrysanthemum *mon* to the emperor, with junior members of the imperial family using a different variety of the flower. This last distinction corresponds exactly to the rules of heraldic precedence that apply to the European royal families. That areas so far removed from each other as western Europe and Japan should have developed a system of hereditary symbolism independently of one another is not surprising, for in both areas feudalism was the prevailing medieval political and social system. As in Europe, Japanese heraldry survived the obsolescence of armour and has remained in widespread use in the 20th century.

Despite some uninformed opinion to the contrary, tartan has no connection with heraldry. It is simply a form of weaving cloth that is by no means restricted to the Scottish Highlands. Armorial bearings were adopted by Highland chiefs in imitation of the Lowland chivalry from the 13th and 14th centuries. The badge of the chief

By courtesy of (left) the trustees of the British Museum, (right) the Oriental Institute, University of Chicago



Seals as heraldic prototypes.
(Left) Impression of the great seal of Richard I of England, showing the mounted king bearing arms. In the British Museum. (Right) Cylinder seal impression from the Akkadian Period with a combat scene between a bearded hero and a bull-man, and beasts. In the Oriental Institute, University of Chicago.



Japanese *mon*, or badge.
Drawing by Wm. A. Norman

was adopted and used extensively by the members of his clan.

Flags can be heraldic. That of the United Kingdom is certainly so; it is formed by the amalgamation of the flags of England, Scotland, and Ireland, these showing respectively the crosses of St. George, St. Andrew, and St. Patrick, all of which are displayed heraldically. The United States flag has a quasi-heraldic character and appears to owe its principal ingredients to the armorial bearings of the first president, George Washington. The flag representing the republic of France, by contrast, is not heraldic, being merely an arrangement of the national colours.

In addition to national flags, there are banners, rectangular pieces of cloth showing the armorial bearings of the owner, and standards, strips of cloth that taper gradually to the end and usually bear heraldic devices but not the owner's full coat of arms.

Like all other human creations, heraldic art has reflected the changes of fashion. As heraldry advanced from its utilitarian usages, its artistic quality declined. In the eighteenth-century, for example, heraldry described new arms in an absurdly obtuse manner and rendered them in an overly intricate style. It was not until the 20th century that heraldic art recovered and in many ways improved upon the originals, especially in displaying animals anatomically. There were still far too many drawings, however, of poor quality emanating from official sources.

This article is divided into the following sections:

I. The scope of heraldry
  The chief components of armorial bearings
  The elements and grammar of heraldic design
    The general language
    The field or ground of the shield
    The "charges" on the field
    Minor charges
    The nature and origins of heraldic terminology
  The reading of heraldry
  Manipulation of heraldic design
    Cadency
    Arms of women
    Quarterings and marshalling
    Arms of bastardy
    Nonfamilial heraldry
II. Historical development of heraldry
  Early roots of heraldry
  Growth of heraldry after the 13th century
    The heralds
    Writers on heraldry
    Continental versus British heraldry
  Twentieth-century heraldry
    The survival of the heraldic tradition
    Uses of heraldry for study and verification

I. The scope of heraldry

Institutional arms from medieval times

An early development was the extension of heraldic design from its use by persons or families to its employment by institutions and associations of various kinds, a consequence of the concept that an assembly or body of people can be personified as an individual, much as a limited company or corporation is viewed as a legal "person." Medieval times provided numerous examples of arms borne by municipalities, churches, and colleges. The arms assumed by an individual or granted to him are regarded as being peculiarly his possession, therefore caution must be used in speaking of family arms. This question can be best dealt with in connection with the royal arms of the sovereign of the United Kingdom. These arms are borne in their entirety only by the reigning king or queen. No other member of the royal family can bear the arms without a "difference" mark that will show without doubt that the bearer is not the reigning sovereign. By analogy, the same condition holds for all so-called family arms, which belong to the head of the family; all other members should strictly bear them differenced—that is, with some mark of cadency (a sign indicating the position of the bearer with respect to the head of the family). In Scottish heraldry, this rule is very rigidly enforced, but in England and elsewhere it has been allowed to fall into decay, except in the case of the royal family.

Probably the next development in the scope of heraldry was its use by ecclesiastics. The bishops and the abbots of the monasteries used arms on their seals from the 12th century onward. In this variety of heraldic usage, the arms were not those of individuals but of the body they temporarily represented—as also with arms borne by political units such as nations and cities or by educational establishments, many of which date from the Middle Ages. A great extension of medieval heraldry was connected with what came to be called the livery companies. These were guilds or associations of men in trades whose object was to uphold standards of craftsmanship. Most of them obtained charters from the crown and were granted arms. Among numerous examples in Britain are the Grocers, the Mercers, and the Glaziers companies. Membership in these still-existing companies no longer entails practice of their particular trades, but they possess property and have great charitable interests as well as considerable social esteem. Their armorial bearings are of great antiquity and are much displayed on their halls, letterheads, glass, silver, and so forth. Obviously, armorial bearings were assumed in the Middle Ages by such military bodies as the Knights Templars, the Knights of St. John of Jerusalem, the Teutonic Knights, and the great Spanish orders. Military heraldry has continued to the present: each of the three British armed forces, for example, has badges, or in some cases coats of arms, which are in the care of officers of the English College of Arms. As a matter of fact, one of the greatest modern uses of heraldry occurred during World War II when some 500 emblems were employed as divisional and other signs.

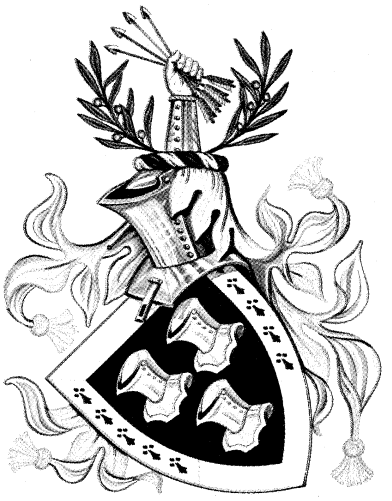
In the 20th century, the development of corporate heraldry has gone far beyond anything known before. Throughout the world, banks, insurance companies, and many other great commercial concerns use arms, together with an ever-increasing number of professional, educational, and trade associations.

Perhaps the event most illuminative of the modern scope of heraldry was the grant, in 1961, by the government of the Republic of Ireland of armorial bearings to the president of the United States of America, John F. Kennedy. Because arms are hereditary and their owners are regarded heraldically as of noble status, the grant amounted to a bestowal of nobility by a state on the head of another state, an occurrence unique in heraldic history.

THE CHIEF COMPONENTS OF ARMORIAL BEARINGS

Heraldry originated when most men were illiterate but could easily recognize a bold, striking, and simple design. The use of heraldry in medieval warfare enabled combatants to distinguish one mail-clad knight from another and thus to know friend from foe. Thus, simplicity was the principal characteristic of medieval heraldry. In the tournament there was a more elaborate form of heraldic design. When heraldry was no longer used in war and heraldic devices had become a part of civilian life, an intricate type of design evolved with an esoteric significance utterly at variance with its original purpose. In modern times, heraldry has often been looked on as mysterious, a matter for experts only. Indeed, over the centuries its language has become intricate and pedantic. Such intricacy appears ridiculous when it is remembered that in the earlier periods swift recognition of a coat of arms or badge could mean the difference between safety

The trend toward complexity and esoteric significance



The arms of U.S. Pres. John F. Kennedy. The gold helmets are a variant on the three silver helmets of an ancient Kennedy coat. The border was added as a further difference. The olive branches and sheaf of arrows are derived from those of the Great Seal of the United States. The lambrequin is blazoned as argent and gules; this is exceptional to the rule that the principal tinctures of the arms (in this case or and sable) be repeated. No motto was included in the grant. The diagonal orientation of the shield is called couché and is optional in all depictions of arms.
Drawing by Wm. A. Norman

and death, and in many medieval instances battles were lost through a mistake over the sameness of two devices of opposing sides.

The shield. The shield is the essential part of the armorial bearings; without it there can be no heraldic device, except for a woman, a distinction that calls for special treatment and is dealt with later. The word shield can be used to describe the coat of arms but in modern times is seldom employed in this way, except in a poetic context. Armorial bearings are generally referred to more briefly as "arms," or as a "coat of arms," a term derived from the surcoat of silk or linen worn over the armour to keep off the rays of the sun and to prevent rust from forming on the armour. The surcoat was a waistcoat-like garment on which were shown the same heraldic insignia as on the shield.

Every other object in heraldic achievement is dependent upon the shield or coat of arms. There can be, and quite often is, a coat of arms consisting solely of a shield without any other object, such as a crest surmounted. The arms of the Churchills of Muston, a branch of the same family to which Sir Winston Churchill belonged, have no crest. The reason is that such families possessed arms before crests became fashionable.

The crest. A crest is the object placed on top of the helmet and bound onto it by what is called the wreath of the colours, which shows the two main colours of the shield (see illustration). Crests were at first made of leather, later of light wood, and, as time went on, of more valuable materials. It is supposed that they were at first borne in tournaments; they became general in families in England from the 16th century when the venal heralds of that time persuaded crestless families to acquire the addition for a payment. Nowadays, a crest is automatically included in any grant of arms made in England, Scotland, or Ireland.

The crest and the age of the automobile

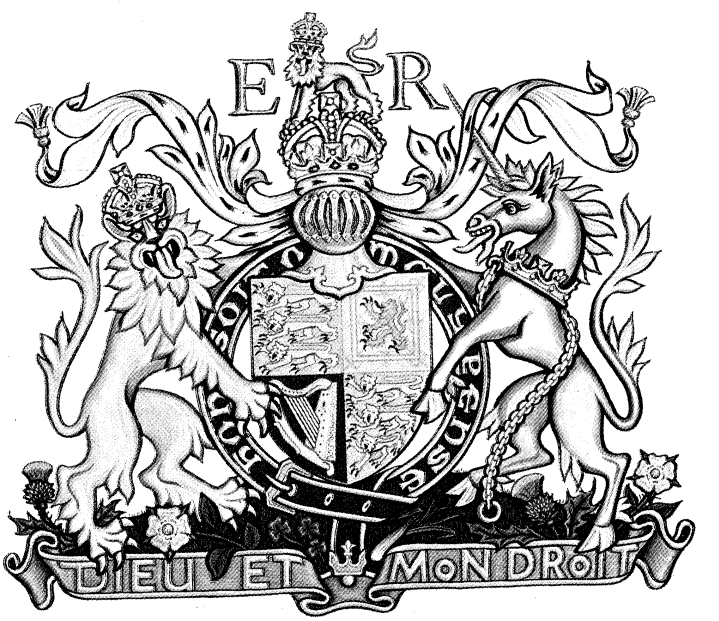
When horsedrawn carriages were in use, it was the rule to show the whole heraldic insignia on the coach or carriage door. With the advent of motorcars and their smaller door space, the arms were usually left off and only the crest and motto shown. This development may be the reason for the mistake frequently encountered in which the whole coat of arms is referred to as a "crest." It should be emphasized that while a coat of arms can exist without a crest, the existence of a crest without a coat of arms is an impossibility.

The helmet. On top of the shield is placed the helmet, upon which the crest is tied by the wreath. Originally, everything in heraldry was strictly utilitarian. As armorial bearings were used with the suit of armour, there had to be a helmet. In later centuries, rules for the depiction of the helmet were elaborated to show the rank of the bearer, with some displayed in profile and some in full face, and with different metals and accoutrements to indicate status. The shape of the helmet has varied greatly in heraldic representation. While the basic features of heraldry remain unchanged, the modes in which the insignia are shown are as subject to change and to fashion as any other human creations. The barrel-shaped helmet was that in use in the 13th century. The tournament helmet, which is often shown in drawings, was of a different type altogether, its shape resembling that of a soup tureen.

Mantling. From the helmet hung the mantling or lambrequin. This was of linen or other material and performed the useful function of shielding the wearer from the sun's rays and also served to catch or deflect sword cuts. The mantling or mantle is painted with the principal colour of the shield, while its lining is of the principal metal. More elaborately styled mantles are used for kings and princes.

Crowns and coronets. These are emblems of the rank of the bearer. With the abolition of most of the great European monarchies, the study of crowns has become mainly of historical and antiquarian interest. The most famous royal crown remaining in the 20th century is that of the United Kingdom; it appears in the sovereign's

Drawing by Wm. A. Norman



The chief components of armorial bearings as indicated on the royal arms of the United Kingdom of Great Britain and Northern Ireland. The royal cipher (ER) is not a part of the arms proper but identifies them as representing Queen Elizabeth II. The roman numeral II is unnecessary here, as the arms of Elizabeth I were different, apart from those of England. The shield shows England (gules three leopards or) quartered with Scotland (or a lion rampant within a double tressure flory counterflory gules) and Ireland (azure a harp or stringed argent). This is the quartering in use since the accession of Queen Victoria in 1837. The shield is encircled by the garter of the Order of the Garter bearing the motto of the order, *Honi soit qui mal y pense.* The dexter supporter, a royally crowned gold lion guardant, and the sinister supporter, a silver unicorn with gold horn, hooves, mane, and tufts and a gold coronet collar and chain, represent England and Scotland, respectively. Atop the full-faced helm of a sovereign with its ermine and gold mantling, or lambrequin, is the royal crown surmounted by the royal crest, a lion statant guardant crowned with the royal crown. The motto *Dieu et mon droit* ("God and my right"), first used by Richard I, appears on the scroll below. The ground beneath the full achievement, called the compartment, is strewn with the floral and plant badges of England (rose), Scotland (thistle), Ireland (shamrock), and Wales (leek).

arms upon the royal helmet and the crest of a golden lion crowned. Coronets (small crowns implying dignity inferior to that of the sovereign) are emblems of rank that are shown, when depicted, between shield and helmet. In Britain there are different coronets specified for the ranks of baron, viscount, earl, marquess, and duke. On the European continent, a much wider use of coronets has prevailed. Among the relics of this usage is the crest coronet, a coronet that supports the crest either instead of the wreath or in addition to it and resting upon it. Another is the chapeau or cap of maintenance, a cap with fur lining that was once worn on the helmet before the development of mantling and that is often used instead of the wreath to support the crest.

The myth of battle cries as sources of mottoes

Mottoes. Many myths have grown up around mottoes. They are often said to have originated as battle cries, but very few actually did. Among those that did are the *Crom a boo*, of the FitzGeralds, the dukes of Leinster, meaning "Crom [one of the family's old castles] forever"; and the "firth fortune and fill the fetters" of some Scottish families, which defies explanation and must refer to some forgotten incident. In succeeding centuries large numbers of mottoes were adopted. They are not part of a coat of arms and can be varied at the user's pleasure, though they are included in a modern grant of arms. More than one motto may be used by the same family. In Scottish arms the motto is usually shown above the crest, in all other countries beneath the arms and always contained in a scroll.

The supporters. These are the figures on either side of the shield of arms and are borne (in English heraldry) by peers, and by other bearers of orders of the highest class, such as Knights of the Garter, the Thistle, St. Patrick, and by Knights Grand Cross. In former times, supporters were used more widely, and a few English families still claim the right. In Scotland their use is much more frequent.

The compartment. The ground or foundation on which the supporters stand is called the compartment. In Scottish arms it is usually a rock or piece of ground strewn with some heraldic object. In England the compartment ought to be shown in the same way, and in the 20th century often is, with the scroll of the motto beneath it; but in the debased heraldic art of the 18th and 19th centuries, the supporters were generally shown as standing on a piece of ironwork or on the scroll.

The achievement. In heraldic writing, the term achievement often carries the same meaning as "arms," but probably its better usage is to describe the whole representation showing shield, helmet, crest, mantling, and supporters. The achievement belongs to only a minority of those who possess arms, since only a few have supporters. In addition, an achievement may include representation of various knightly orders or companionships of knightly orders to which the owner of the arms is entitled. For example, Viscount Montgomery of Alamein, British field marshal and World War II military leader, can show the symbol of the Order of the Garter around his shield; persons with lesser distinctions such as the Distinguished Service Order, Military Cross, and Order of the British Empire may have the decorations shown pendent from their shields. As distinctions of this kind are not hereditary, on the death of the bearer the successor to the arms must not use representations that show these honours.

Badges as symbols of persons or nations

The badge. This is older than the heraldic system. Such symbols expressing a person, a body, or an impersonal idea are found from ancient times. The eagle of Rome was one of the state's symbols and was the special device of the legions. Many such symbols bring to mind the country they represent; *e.g.*, winged bulls with human faces at once recall Assyria. On Trajan's Column in Rome, devices sometimes bear resemblances to later heraldic designs. On Etruscan vases are seen what not unfairly could be called demi-boars or bulls' heads caboshed. Nearer to heraldic times, the planta genista, or broom plant, which gave its name to the Plantaganet dynasty of England between 1154 and 1485, was a badge of the counts of Anjou before that family had armorial bearings. With the growth of heraldry, badges naturally assumed a heraldic character. They could be varied at the will of the holder, who often had more than one. Badges persist to the present and sometimes accompany a grant of arms.



Badges.
English badge: the red rose of Lancaster charged with the white rose of York, surmounted by the royal crown. Italian badge: the knot of the royal house of Savoy. French badge: the porcupine of Orléans, first used by Louis XII; the crown is not always included. Welsh badge: the leek; the daffodil is also a long-established badge of Wales.

Banners and standards. The users of arms in the Middle Ages often displayed them on the fork-tailed pennons of their lances. When the forked ends were cut off, the resulting flag was square, becoming a banner. Particularly valorous conduct was often indicated in this way, and the knight thus distinguished was known as a knight banneret. This last word was sometimes corrupted into baronet, a term that the English antiquary Sir Robert Cotton (1571–1631) used when he suggested to James I of England that he should revive a supposed order of baronets. The resulting Order (1611) was hereditary, however, and had no true connection with medieval knighthood. The banner bears the owner's arms, and in the 20th century anyone who possesses arms is entitled to a heraldic flag in this form. These can sometimes be seen in Great Britain flying over a house. No banner is referred to in the grant of arms made to President Kennedy, but in due course an armorial banner was made for the occasion when his brother, Sen. Robert Kennedy, carried it to the top of a mountain in the Yukon Territory named Mt. Kennedy by the Canadian government in memory of the President. The banner showed the Kennedy arms without the crest. The maker of the banner added a long white streamer on which he put the badge, the latter being part of the Kennedy crest.

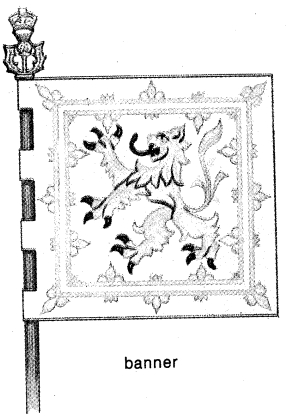
On the standard, the main colours of the arms are shown with the owner's badge.

In England flags are often seen flying above churches. When these show the flag of St. George (England's patron saint), white with a red cross, they carry in the top right-hand corner the arms of the diocese in which the church is situated. Heraldic flags also flew in countries on the Continent in a similar manner. With the disestablishment of heraldic offices in most European countries, contemporary flags are for the most part non-heraldic.

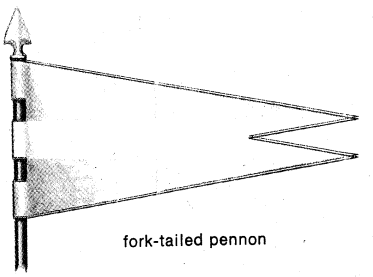
THE ELEMENTS AND GRAMMAR OF HERALDIC DESIGN

Rudiments of heraldry in about 50 terms

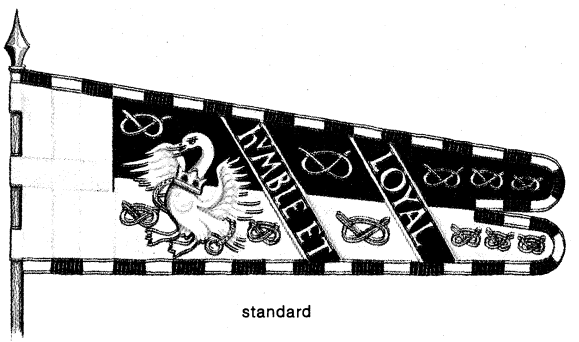
The general language. Provided that a few elementary principles are grasped, enough knowledge of heraldry can be acquired in a relatively short time to enable the student to understand the meaning of coats of arms. The multitude of terms used in heraldry need not worry him: once the rudiments are learned with some 50 of the terms, the meaning of the large remainder can be ascertained as the occasion arises. For example, when Queen Elizabeth II was crowned, some beautifully carved figures were made of the different badges used by her



banner



fork-tailed pennon



standard

Heraldic flags.
Banner: the blazon of the shield is applied to the whole surface of a square or a vertically or horizontally oriented rectangular flag. This is the royal banner of Scotland, which follows the blazon of the second quarter of the royal arms of the United Kingdom. Although it is the banner of the sovereign, it is widely but incorrectly used today as the national symbol. Fork-tailed pennon: shown here is that of the Sovereign Military Order of Malta; gules a cross argent. Standard: the cross of St. George at the hoist identifies this as English. The profusion of badges, the diagonally placed motto, and the border of alternating tinctures are typical. This is the standard of Sir Henry Stafford, *c.* 1475.
Drawing by Wm. A. Norman

ancestors, figures now displayed at Hampton Court Palace. They include one very rare badge—a yale. It is a mythical heraldic creature. Anyone unfamiliar with it could easily ascertain its meaning from the various heraldic glossaries. It is therefore unnecessary to burden the memory with hundreds of terms (a glossary generally contains about 800 terms).

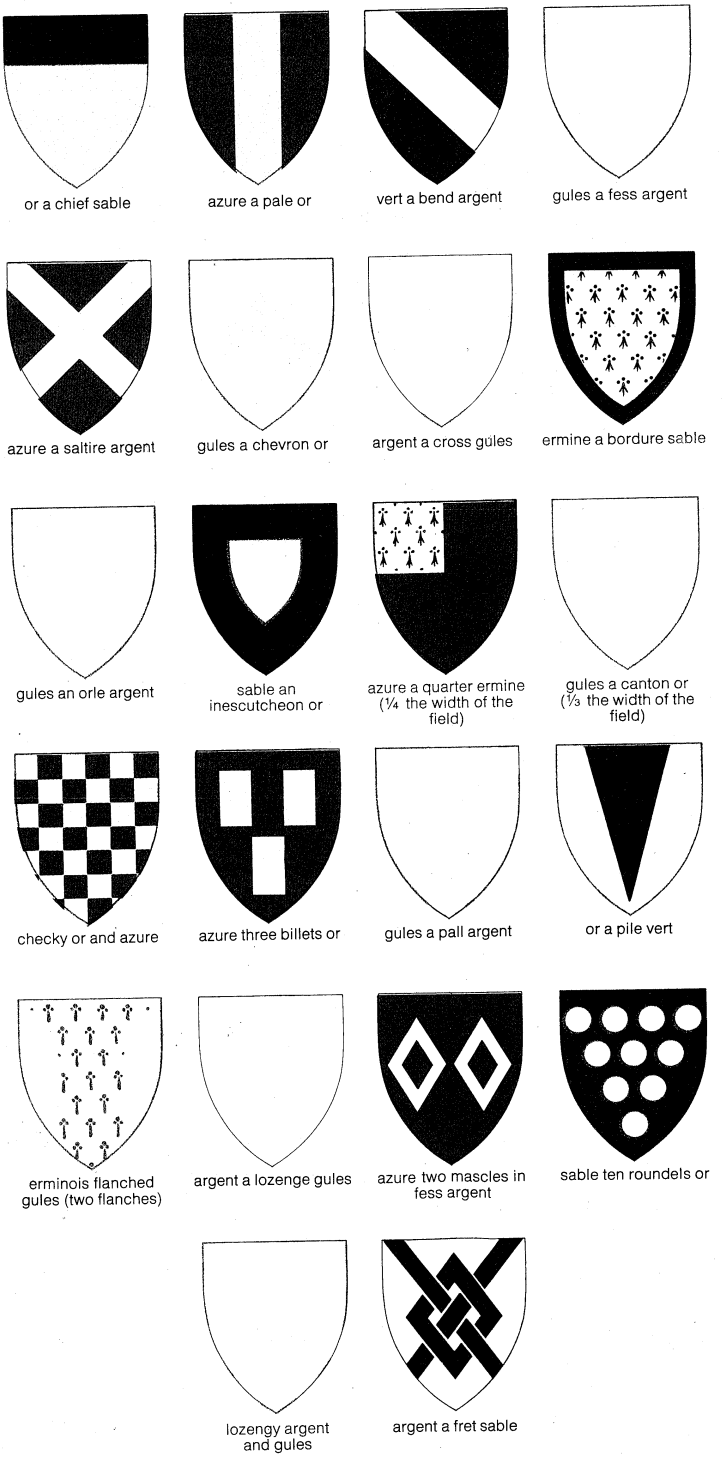
The language of heraldry has a curious look. "Azure three wheat sheaves or" has been known to call forth the question, "Or what?" When it is remembered that *or* is the French for gold, the difficulty diminishes. Much heraldic terminology is a quasi-French, archaic language. In the Middle Ages, the French language was used by the ruling class in much of western Europe, so that it was not unnatural that heraldic terms should be French. In England, by about 1400, English words usually were used in preference. Much modern heraldic terminology, however, is so obscure that it seems purposely designed to puzzle the uninitiated.

The terms dexter and sinister mean merely "right" and "left." A shield is understood to be as if held by a user whom the beholder is facing. Thus the side of the shield facing the beholder's left is the dexter or right hand, and that opposite his right, the sinister or left hand.

The field or ground of the shield. The field or ground of the shield is of one of three, and only three, kinds: a colour, a metal, or a fur. There are five main colours (known as tinctures): azure (blue); gules (red); sable (black); vert (green); and purpure (purple). In English heraldry there are also murrey (sanguine), a tint between gules and purpure, and tenné (an orange-tawny colour). The metals are or (gold) and argent (silver, which is often shown in coloured illustrations as white). The furs consist of ermine (white field with black spots); ermines (black field with white spots); erminois (gold field with black spots); pean (black field with gold spots); and vair (composed originally of pieces of fur from a species of squirrel that was blue gray on the back and white underneath, so that when several of its skins were sewn together, the result was a series of cup-shaped figures, alternately blue and white, which is the way vair is shown).

The furs —from ermine to vair

Drawing by Wm. A. Norman



Ordinaries, basic bearings that may be of any tincture and that may be combined in great variety. A combination of a cross (England) and two saltires (Scotland and Ireland) has resulted in the familiar Union Jack of the United Kingdom. Ermine and certain other textures such as ermines (black with white ermine tails) are regarded as tinctures in their own right and may bear superimposed charges. Discrete charges such as lozenges, mascles, fleurs-de-lis, etc., may be used singly, in pairs, in threes or greater numbers, sometimes in great profusion, as that of ermine tails.

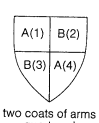
It is considered bad practice to put a colour upon a colour, a metal upon a metal, or a fur upon a fur. Many examples of such bad heraldry are found in old records, but in this as in other instances, the rules now prevalent grew up only very gradually.

The "charges" on the field. The field is said to be "charged" with an object. Heraldic objects are of an immense and growing variety; as more and more arms are devised, more and more new objects appear as charges—telescopes, aircraft, rolls of newsprint, and so on. Charges are divisible into two classes: the ordinaries (or honourable ordinaries as they are often called, deriving their name from the fact that they are so frequently used) and the others.

The ordinaries. The ordinaries comprise some 20 figures. Among them are: the chief, being a third of the shield and the top part; the pale, a third part of the shield, drawn perpendicularly; the bend, a third part of the shield, drawn from the dexter chief to sinister base; the bend, sinister, drawn from the dexter base to sinister chief; the fess, a third part and taking up the centre of the shield; and the chevron, resembling an inverted stripe in the rank badge of a noncommissioned officer. It should be noted that the bar is a diminutive of the fess, of the same shape, and can be placed in any part of the shield. The term bar sinister is often used in fiction as a synonym for bastardy. It has no such significance, bastardy being denoted heraldically in several other different ways. Since the European nations were Christian when heraldry was invented, the cross appears in many forms in heraldry. The cross in a coat of arms does not imply, however, that the original bearers were Crusaders.

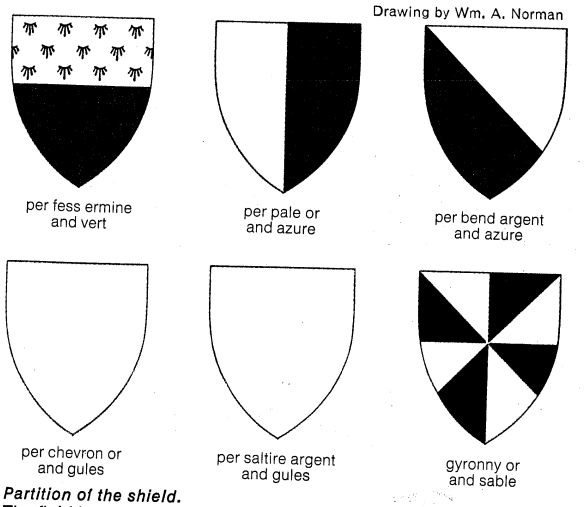
The border, or bordure, is used as a mark of difference sometimes, and in English heraldry over the last 200 years as a sign of bastardy. The orle is an inner border, not touching the ends of the shield, and in which the field is seen within and around it, giving it the appearance of a shield with the middle cut out (voided, in heraldry). The treasure, much used in Scottish heraldry, is a diminutive of the orle. The inescutcheon, a small figure shaped like a shield in the middle of the shield, is used to denote the arms of a heraldic heiress; the quarter occupies one-fourth of the shield; the canton, less than the quarter and one-third of the chief; and chequy or chequé, the field or charge divided by transverse lines horizontally or perpendicularly into equal parts. Billets are oblong figures. Should their number exceed ten, and they are irregularly placed, the field is described as *billetté;* the paile or pall is the upper half of a saltire (St. Andrew's Cross) and half a pale. The pile is in the shape of an inverted pyramid. The flanch, or flanque, is a segment of a circle drawn from a corner of the chief to the base point; the lozenge is a parallelogram having equal sides forming two acute and two obtuse angles; and a mascle is a lozenge voided; the roundel is circular in form. Lozengy is the field divided by diagonal lines transversely and the fret is a mascle interlaced with a saltire.

Minor charges. A field is said to be "powdered" or "semé" when strewn with minor charges; when charged with drops of liquid, it is "guttée." Partition lines divide the shield. The most common ones are straight. "Impalement" or "dimidiation" means the division of the shield into two equal parts by a straight line from the top to bottom. This method is used to show either the arms of husband and wife, the arms of the husband being in the dexter half, or certain types of official arms, as the arms of a bishop's see impaled with his family arms, those of the see being in the dexter half. The shield is divided into four quarters when one coat of arms is quartered with another, as when the children of a heraldic heiress (the eldest daughter of a family of no sons) use their mother's arms with their father's. (Thus in the illustration A = father's arms, B = mother's arms.)



two coats of arms quartered

Other divisions of a shield are: party per pale or simply per pale, division of the field into two equal parts by a perpendicular line (this resembles the impalement just mentioned but does not serve the same purpose of combining arms); party per fess, division into two equal parts by a horizontal line; party per bend; party per chevron; party per saltire; gyronny of eight. When the partition lines are not straight, they can be of several varieties.



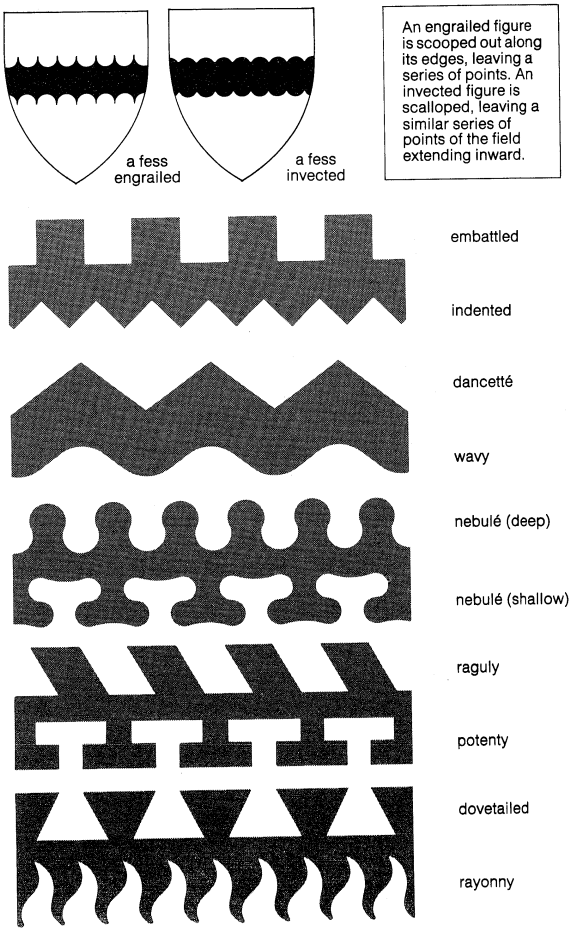
Partition of the shield.
The field is often divided along the lines occupied by ordinaries, just as quartering imitates a cross. "Per fess" means along the line over which a fess would be laid down. The ermine tails illustrated are one type of stylization among many in use. The superior dexter segment on the gyronny shield is called a gyron and is occasionally found singly.

The nature and origins of heraldic terminology. Fanciful explanations have been advanced to account for heraldic charges: for example, argent to denote purity, the bend derived from the military cross belt—the cross a sign of a crusading ancestor—and so on. Since no one wrote about heraldry until it had existed for more than 200 years, these explanations of its symbolism can be discounted. With very few exceptions, the origin of the charges is unknown. The Stourton arms ("sable a bend or between six fountains proper") refer to the six springs in the park of their ancestral estate that are the source of the river Stour. A heraldic fountain does not resemble a real fountain but is put in the form of "roundel wavy argent and azure" (a silver and blue circlet of wavy lines), unless it is expressly stated that the fountain is "proper"; *i.e.*, a natural fountain. The word proper is always used to denote a charge shown in its natural colours or natural form.

Augmentations of honour

The derivation of heraldic charges is more easily discerned in the augmentations of honour, as they are called, when something has been added to a coat of arms by the (British) crown in recognition of services rendered. The arms of the British naval hero Admiral Horatio Nelson show fresh heraldic charges added to his ancestral arms as his victories were gained. Within the past 300 years, augmentations have generally been recorded. An example is the augmentation granted by Queen Victoria to commemorate the discovery by the English explorer John Hanning Speke of the sources of the Nile. The honour, granted posthumously, consisted of the addition to the existing arms of a chief azure upon which appeared a representation of flowing water proper superinscribed with the word Nile in gold lettering. Usually, however, the origins of the various objects used in heraldry are not known. Numerous historical instances of augmentations of honour occurred in continental Europe, especially in connection with the Holy Roman emperors. Frederick II, for example, granted to Conrad Malaspina an augmentation of a chief of the empire, thereby adding an eagle displayed sable to the Malaspina arms of per fess gules and or overall a thorn branch vert with five flowers argent in pale.

Heraldic descriptions are called blazons. The term is derived from the French *blason*, the etymology of which is uncertain. Originally, it denoted the shield of arms



Types of divisions between tinctures.
A line described as flory, or flory counterflory, employs a series of small fleurs-de-lis that have substance of their own beyond the two areas being divided. Rayonny (or *rayonné*) may alternate straight points with curved points.
Drawing by Wm. A. Norman

itself and still retains this meaning, but it is now generally used in a derivative sense as meaning the description of the arms. Blazon is thus a noun, and there is also the verb to blazon; *i.e.*, to describe a coat of arms.

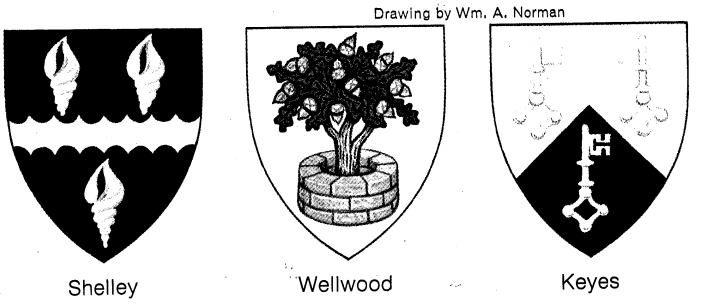
There are four generalizations that are useful in the deciphering of blazons. First early coats of arms are simple because they were original and there were so few of them that elaborate differentiation was not required. As time brought many more coats of arms into being, simple coats became much rarer, and the passing of warlike usage made arms much more complicated. Second, punning or canting arms are very common as, for example, trumpets for Trumpington, a spear for Shakespeare. It is notable, however, that many armorial allusions which were formerly obvious now require research for elucidation. Other allusions have been lost entirely. Third, in grants of arms to people bearing the same name but having no relationship with each other, difference marks have had to be put in. Again, in consequence, blazons have become much more complicated. Finally, in the course of centuries and frequent intermarriages among arms bearers, many quarterly and grandquarterly coats have come about. Quarterly and grandquarterly coats are much more difficult to describe than the simple coats.

Apart from the ordinaries and those other charges that have been mentioned incidentally, there are some peculiarities of heraldic charges that need to be noted. Mythical birds and animals are much used, the product of ancient and medieval natural history—or the lack of it. Such are the dragon, griffin, wyvern, harpy, phoenix, and martlet. In addition, there are some creatures bearing the names of real animals but not resembling them in all respects. The heraldic tiger is more like a lion or a wolf in some features. When the real tiger became known to heraldry, it was described as a Bengal tiger

Heraldic descriptions of animals and their postures

The heraldic description of animals is very important. Rampant means on the hindlegs, while rampant guardant is the same posture but full faced. Reguardant means looking back; passant, walking. Combattant signifies two animals fighting on hindlegs. Couchant is lying down; dormant, sleeping; and sejant, sitting. A beast of the hunt is called at gaze when looking full face, trippant when at trot, with one foot raised, and statant when standing. Parts of an animal may be a charge—a demi-lion or demiwolf, a lion's or bear's gamb or foreleg. Heads are described as erased when cut off by a jagged line, couped when cut by a straight line, and caboshed when the severed head looks straight forward. A bird shown with wings expanded is said to be displayed. Creatures placed back to back are addorsed. A fabulous bird, the phoenix, is known to heraldry, as was the legendary pelican that fed her young on her own blood and was called "in her piety," then being considered an emblem of Jesus Christ, who fed or redeemed his flock with his own blood. The martlet is another fabulous bird widely known outside heraldry, owing to John Milton's reference to the herald's martlet, which has no legs. It is a frequent charge, resembling a swallow, and is used in cadency to denote the fourth son. Other terms have special heraldic significance. Armed is used of the horns, teeth, or claws of a beast, or the beak or talons of a bird, and of the human being when in armour. The term slipped applies to flowers and fruit when the stalk is seen. Counterchanged refers to the field when it has two tinctures, a metal and a colour, and when one is the background for the other on one side of the shield and then their relationship is reversed on the other side. An example is the Warner arms: "Per bend argent and gules two bendlets between six roses all counterchanged"; where the three roses on argent will be gules and three on the gules will be argent. The field is separated by a partition line and the charge or charges of the arms are said to be countercharged when the charge or portion of a charge that lies on the metal is of the colour and vice versa.

Drawing by Wm. A. Norman



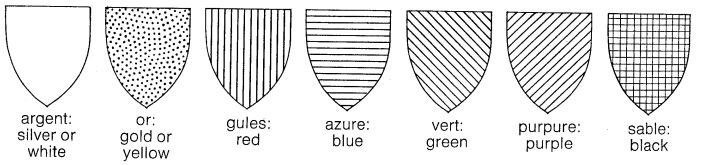
Canting, or punning, arms, derived from the literal meaning or from the sound of a name.
(Left) Shelley: sable a fess engrailed between three whelk shells or. (Centre) Wellwood: argent an oak tree growing out of a well all proper. (Right) Keyes: per chevron gules and sable, three keys or.

THE READING OF HERALDRY

The Petra-Sancta system of denoting colour by black-and-white patterns

A method has been devised to indicate heraldic colours in black-and-white illustrations. Known as the system of Sylvester Petra-Sancta, an Italian herald, it makes use of the following equivalents: or is denoted by dots or points; azure by horizontal lines; vert by lines from dexter chief to sinister base; purpure by lines from sinister chief to dexter base; argent by a plain field; gules by perpendicular lines; sable by cross lines horizontal and perpendicular. Furs are depicted with black or white spots on the appropriate ground; vair and countervair are shown by alternate lines and plain surfaces.

Describing, or blazoning, of arms must always begin with an identification of the ground of the shield, such as argent or gules or ermine. For a woman who is not married, the arms normally appear on a lozenge, not a shield, but the field or ground in this instance, too, must



Conventional representations of tinctures used by engravers and others when actual colours are not practicable.
Drawing by Wm. A. Norman

be the start of the blazon. Then come the charges. A typical blazon is thus: "sable [ground of shield] a chevron ermine between three lions rampant argent crowned or" (arms ascribed to a family of Hinstoke). The chevron is a fur; the lions are silver, appear on the sides of the chevron and its base, and have gold crowns. One important feature in heraldic writing is economy of words; technically, it should be possible to avoid punctuation marks, thus: "azure a fess between three stags trippant or" (Hird). Here both fess and stags are in gold. When three beasts are depicted, they are shown in the most convenient way around the main charge; that is to say, two in the upper part of the shield and one below. A straightforward coat with only one charge on the field is that of the Italian Segni family of Agnani, which gave to the church three popes, Innocent III, Gregory IX, and Alexander IV: gules an eagle displayed chequy sable and or. Economy, however, can be carried too far as in the following: "azure a lion rampant double queued barry of ten argent and gules armed and langued of the last crowned or, within a bordure of the second and third" (Mountbatten). Here is an example of a usage that grew up in past centuries designed to avoid repetition of the name of a tincture but in reality serving to create confusion. "Of the last" means that the lion's claws and tongue are in red or gules. "Of the second and third" means simply argent and gules. There is no real economy since more words actually are used and reference has been made to the earlier parts of the blazon. This type of blazon is used when a large amount of heraldic insignia has to be described, but it makes such long blazons unnecessarily complicated. Anyone who is writing a blazon should not use this jargon of last, first, second, and so on.

The helmet is the next item to be characterized, although in blazons it is usually taken for granted and left undescribed. When it is mentioned, it is said to be "befitting his degree." Although the helmet need not appear in written descriptions, it always should be depicted in illustrations. It is a bad feature of many drawings that the helmet is absent, showing the crest as if it were airborne above the shield and thus unsupported.

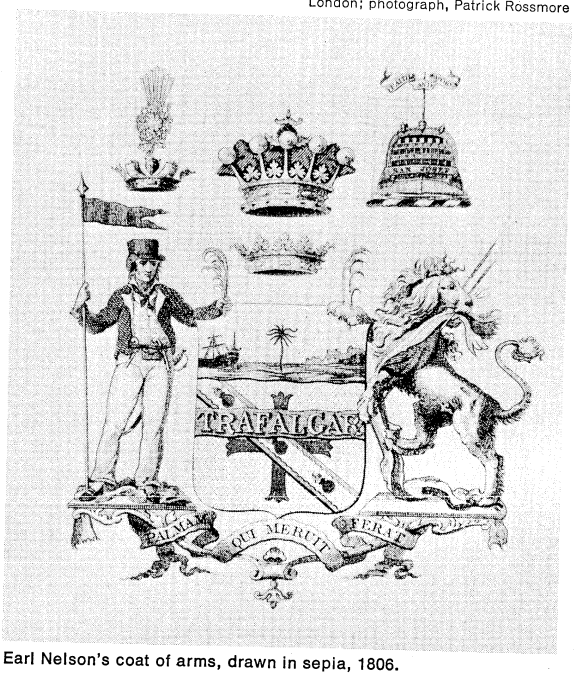
Blazoning the crest

The crest must always be mentioned when, as in the vast majority of cases, one exists. In formal blazons, the wreath (also called the torse) is given as well; thus, crest —"on a wreath of the colours, a wolf passant proper" (Trelawny). The wreath is not usually mentioned, however, because like the helmet it is always assumed to be there. The term colours refers to the two chief colours of the ground of the shield. As with the shield, the older the crest the simpler it will be. Most people can envisage on a knight's helmet the figure of a wolf walking, but it is difficult to picture someone in armour wearing as his head crest "the stern of a Spanish man-of-war on waves of the sea all proper thereon inscribed 'San Josef' with the motto above, 'Faith and Works' " (Nelson). This latter example belongs to the period of decadent heraldry in the late 18th century and 19th century in England.

The mantling, or lambrequin, is mentioned in formal descriptions but not in general usage. The supporters and compartment pertain only to a few classes of arms bearers, and in descriptions the supporters are blazoned after the crest (or crests). The compartment is not usually described but sometimes has to be, as in the arms of the earl of Perth: supporters (two savages = two ancient Caledonians) stand on "a compartment strewn with caltraps" (from "caltrops," iron instruments designed to maim horses' feet and used by the Scots with great effect at the Battle of Bannockburn, 1314).

The motto comes at the end of the description, not being part of the arms. The badge is rarely found, except among very historic families (and by a strange inversion in some 20th-century grants), but when it occurs it, too, comes at the end of the blazon. It can be very simple, as with that of Lord Mowbray, Segrave, and Stourton—"a sledge or." It may be very elaborate, as with Constantine—"a hurt [*i.e.*, a roundel in azure] charged with a leopard's face and surmounted upon the edge with two fleurs-de-lis in pale or, and as many roses in fess, argent, barbed and seeded proper." In this example the roses are silver but the leaves are proper. Coronets of rank are not usually mentioned in English or Scottish heraldry, but caps of maintenance and crest coronets must be blazoned with the crest. Banners and standards are not as a rule mentioned in blazons, though they may be when they occur in a modern grant.

By courtesy of the National Maritime Museum London; photograph, Patrick Rossmore



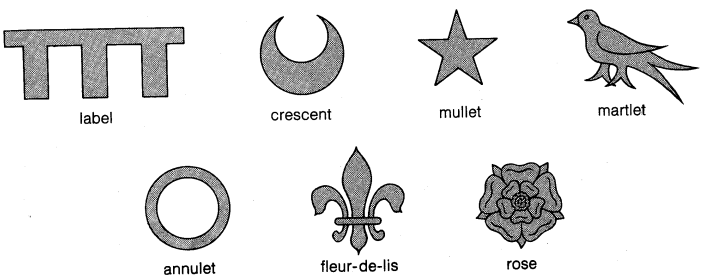
Earl Nelson's coat of arms, drawn in sepia, 1806.

MANIPULATION OF HERALDIC DESIGN

It is clear that the vast majority of heraldic charges are without the foundation of legend often assigned to them, though in many instances the real origin of the charge is lost. A Jacobean dramatist could write of a family as old as the first virtue that merited an escutcheon—a nice poetic flourish that should remain in poetic realms. In modern grants the heralds try to give some allusion to the grantee's work and achievements and his place of origin, and canting arms still appear. Many arms are recondite, however, in their significance, and much has to be learned before the significance of the charges is known.

Cadency. The rules evolved over the centuries to denote particular distinctions in heraldry are fairly straightforward. Cadency is the use of various devices designed to show a man's position in a family, with the aforementioned basic aim of reserving the entire arms to the head of the family and to differentiate the arms of the rest, who are the cadets, or younger members. Heraldic works in the 16th century refer to cadency marks as: a label for the eldest son during his father's lifetime; a crescent for the second son; a mullet (five-pointed star) for the third; a martlet (a mythical bird), the fourth; an annulet (a small ring), the fifth; a fleur-de-lis, the sixth; a rose, the seventh; and so forth. These marks were not always used in the Middle Ages. Differences might be shown instead by a change of tincture, by adding small charges to the field, and the like. Both on the Continent and in England, rules of cadency have long ceased to be used. It is customary for

The use of devices to show position in family



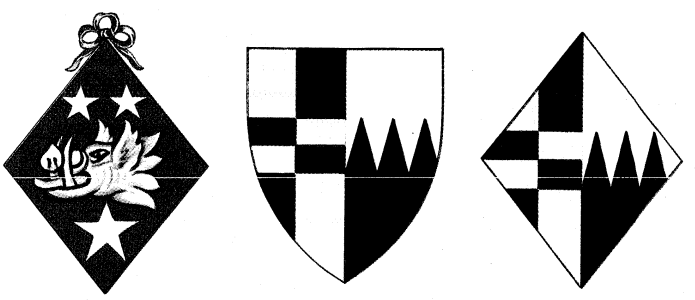
Marks of cadency, used to difference the arms of cadets of the same family. The label, the mark of the eldest male heir, is a notable feature of the arms of the prince of Wales, the heir to the throne. The second cadet displays a crescent, the third a mullet, and so on. These symbols may be of any tincture and may be used otherwise than as signs of cadency.

Drawing by Wm. A. Norman

all members of a family to use the entire arms of the head. There are, however, two exceptions. Very occasionally, a crescent is used for difference by a noble family showing descent from a second son. The other exception occurs in the arms of the British royal family, in which the cadency system exists in rigour. The reason is that the royal arms are arms of sovereignty and cannot be shared. The sovereign alone can have the whole undifferenced arms. Nor does any member of the royal family—not even the prince of Wales—have any right to the use of arms until they have been granted to him by the sovereign. A label of difference with marks is placed on the arms; a three-pronged label for the children of the sovereign, a five-pronged label for grandchildren. The Duke of Windsor after his abdication as Edward VIII in 1936 was granted arms with a label.

In Scotland the position on cadency is very different. Since heraldry is regulated in Scotland by acts of the Scots Parliament before the Union in 1707 with England, and confirmed by the British Parliament, the regulation of arms is very precise. The strict observance of cadency is probably because the Celtic clans formed the original social system in Scotland before the advent of feudalism. Thus only the chief of the name can have the entire arms. He matriculates, or enters his arms, in the registers of the Lord Lyon King of Arms (whose court has jurisdiction over armorial bearings in Scotland). This registration also applies to his eldest son (subject to suitable differencing of the arms) who inherits them in due course. The younger sons must petition for a matriculation of the paternal arms with a suitable difference indicating the position of each in the family. As families from the descendants of the original grantee continue to be established, so there is matriculation and rematriculation, in a carefully prescribed manner.

Drawing by Wm. A. Norman



Arms of women.
A woman adopts the undifferenced arms of her father. An unmarried woman: the arms of her father are displayed on a lozenge; a true lover's knot signifies unmarried status. A married woman: the wife's arms, to the sinister, impale those of the husband, to the dexter; the husband displays the combined arms as head of the family, and the wife shares his escutcheon. A widow: the woman's arms revert to the lozenge, retaining the deceased husband's impaled arms.

Arms of women. Arms of women are shown during spinsterhood or widowhood on a lozenge, not on a shield, without a crest (except in Scotland, where a woman who is chief of a clan and head of the name, such as MacLeod of MacLeod, is allowed a crest). A woman divorced and not remarried also uses a lozenge. The arms of a married woman are shown in conjunction with her husband's by impalement, the division of the shield into two equal portions, the husband's arms on the dexter and the wife's on the sinister. Should she be a heraldic heiress, the arms of her family are placed upon an inescutcheon (originally inner escutcheon) or "escutcheon of pretence" (a small shield whose position in the fess point of the huband's shield gives it precedence over all of the other parts of the shield). The only exception to these rules for women is that of a queen regnant like Elizabeth II, who, being sovereign and thus considered in heraldic terms to be herself the source of honour for all her subjects, possesses the full arms of sovereignty of her royal house and kingdom.

In the Middle Ages, arms for women were often shown on shields, and those like Joan of Arc who bore arms in battle may have used crests.

Quarterings and marshalling. In the quarterings and the marshalling arrangement of more than one coat of arms in the same shield, the position of heiresses must be considered first. The children of a heraldic heiress are entitled on her death to quarter her arms with their father's (the arrangement is to show the shield divided into four quarters so that quarters 1 and 4 are the father's arms, 2 and 3 the mother's). This positioning of the quarterings is also used in England when an additional

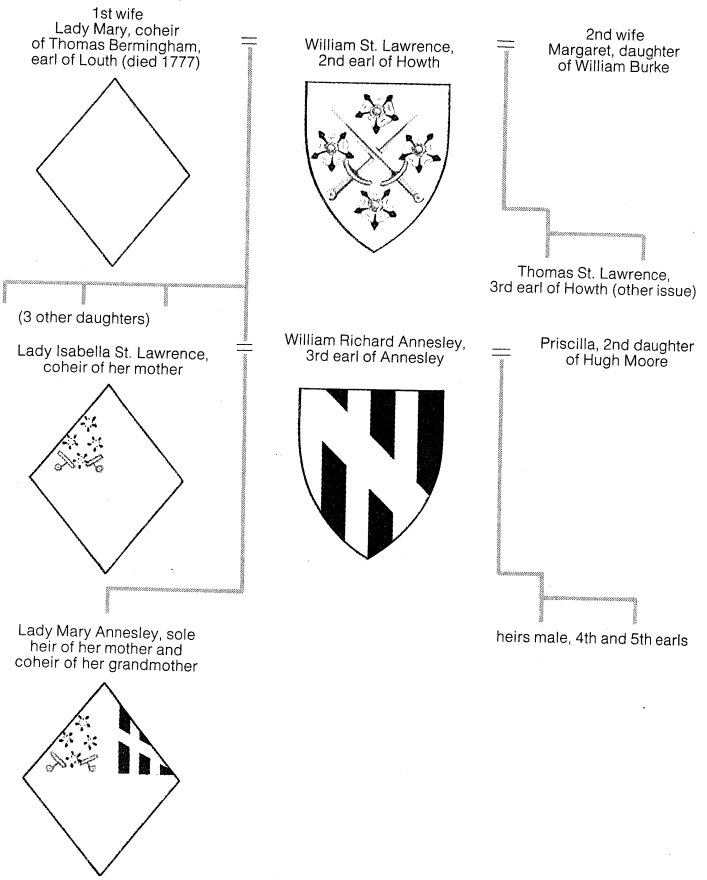
Picturepoin



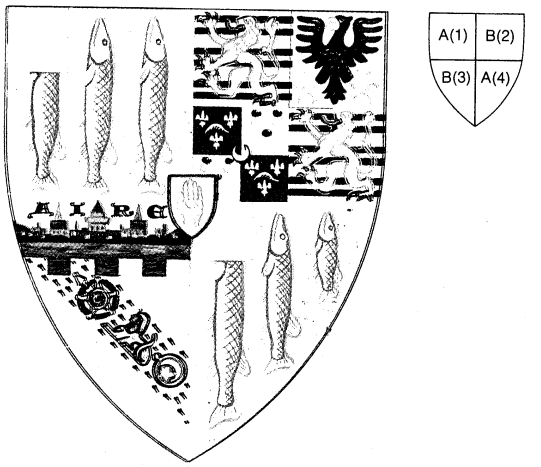
Royal arms of England in the form used between 1340 and 1603: quarterly France and England, as claimed by Edward III. Roof boss from St. George's Chapel, Windsor Castle, 1483–1528.

surname and arms are taken, almost always in obedience to a will. This is the "name and arms" clause peculiar to English law. Its operation over the last 200 years is responsible for the double- or treble-barrelled surname common in England and also found in Scotland. Thus in Salusbury-Trelawny, the original Trelawny arms appear in quarters 1 and 4 and the assumed additional arms for Salusbury in 2 and 3. A famous historical case is that of King Edward III of England, who in 1340 claimed the throne of France in right of his mother, a French princess. He then quartered the lilies of France (the fleurs-de-lis) with the lions (leopards) of England. England should have been placed in 1 and 4, but Edward gave this position to France, probably because of the greater size and resources of France at that time. In this form, the royal arms continued until 1800, when the empty title of king of France was dropped and the lilies went out with it.

When quarterings are inherited from a woman, no crest



Heirship through the female lines.
Drawing by Wm. A. Norman



Marshalling of several coats of arms.
(Left) The arms of the Cameron-Ramsay-Fairfax-Lucy family, blazoned: arms-quarterly, 1st and 4th gules semé of cross-crosslets, three lucies hauriant argent, a canton of the last (Lucy); 2nd, grand quarter counterquartered, 1st and 4th argent. Three bars gemel sable surmounted of a lion rampant gules, armed and langed azure (Fairfax); 2nd parted per pale argent and or, an eagle displayed sable, armed beaked and membered gules (Ramsay); 3rd counterquartered, 1st and 4th azure a branch of palm between three fleurs-de-lis or; 2nd and 3rd gules three annulets or stoned azure. In the centre of these quarters a crescent or (Montgomerie); 3rd gules, three bars or, on a bend ermine, a sphinx between the badge of the royal [Portuguese Order of the Tower and Sword] and the gold medal presented to Col. John Cameron of Fassifern by command of the Grand Signior, in testimony of that sovereign's high sense of his services in Egypt, and on a chiel embattled a representation of the town of Aire in France, all proper (Cameron of Fassifern). (Right) representation of two coats of arms quartered.
Drawing by Wm. A. Norman

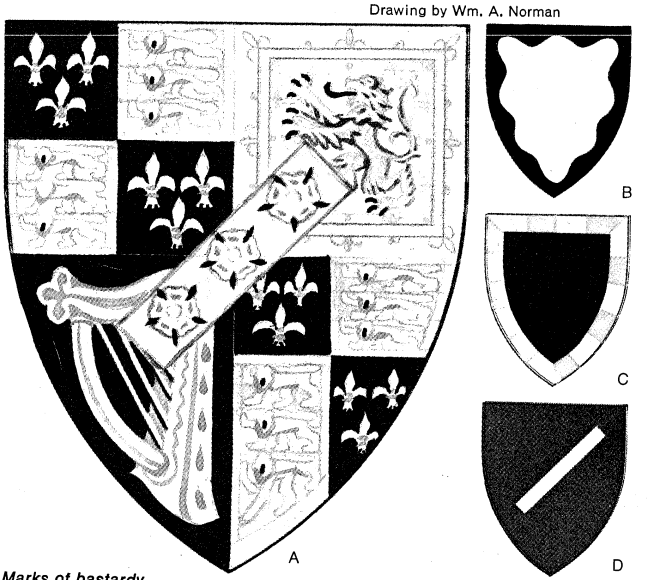
is transmitted with them, because a woman cannot pass on a crest. The matter alters, however, when additional arms are taken in obedience to a will; then a double crest is likely. There is no reason why a further assumption may not occur, so that triple or quadruple hyphenated names are found: for example, the English county family Sawbridge-Erle-Drax has quarterly arms, 1 and 4 Drax, 2 Erle, and 3 Sawbridge. This type of quartering is not difficult to follow, but real problem in marshalling several coats in one shield arises when more than one heraldic heiress occurs in the same family. Some families of long descent—like the dukes of Westminster in England—have often married heraldic heiresses and thus acquired many quarterings. Sometimes several hundred quarterings are attributed to the head of a great family. A splendid instance of quartering occurred in the achievement of the empress Maria Theresa. Before her accession to the imperial throne she was queen of Hungary and Bohemia and by marriage grand duchess of Tuscany. As a sovereign in her own right she bore a shield on which there were 29 quarterings. The dukes of Rohan-Chabot in France bore a quartered shield with Navarre, Scotland, Brittany, and Flanders; overall was an escutcheon of Rohan quartering Chabot.

Problem of marshalling several coats in one shield

Even without such large numbers of arms to deal with, the marshalling of quarterings is still a problem. Various methods have been tried, often with results as difficult to decipher. An interesting illustration of the marshalling of several coats of arms is that of the baronet Cameron-Ramsay-Fairfax-Lucy. The arms are said to be quarterly with the arms of Lucy in 1 and 4. Then in 2, the description runs, grand quarter counterquartered. This means that quarter 2 is itself a quarterly coat, 1 and 4 of which are for Fairfax, 2 for Ramsay, with the third quarter counterquartered; *i.e.*, itself quartered, showing two coats of arms. The third quarter is that for Cameron.

Arms of bastardy. The heraldic illustration of bastardy was achieved in several ways in the older days of chivalry. Little social or moral obloquy attended the status of bastardy, possibly because of the late marriages of the upper classes and their arranged unions. Thus the arms of a bastard were merely differentiated as perhaps those of a distant cadet line would be. From early times the bend or bendlet sinister was used. The erroneous use of bar sinister might have come about because the French for bend sinister was *une barre*. In the arms of British royal bastards like the dukes of St. Albans, of

Drawing by Wm. A. Norman



Marks of bastardy.
These are common marks of illegitimacy but do not invariably have that meaning. (Left) The arms of the duke of St. Albans debruised by a baton sinister, in this case charged with three roses. (Right, top) The bordure wavy (or a bordure wavy sable). (Centre) The bordure compony (vert a bordure compony argent and gules). (Bottom) The baton sinister (purpure a baton sinister argent).

Buccleuch, of Grafton, and others, a baton sinister is used to denote bastardy. The royal arms of these illegitimate scions are said to be "debruised" (crossed or partly covered) by the baton sinister. English heralds in the last 150 years have often signified bastardy by the use of the bordure.

Nonfamilial heraldry. The arms of other than families or individuals will often be encountered, especially in modern times when the granting of arms to private persons has ceased in some countries but where grants of corporate arms are frequent. Such is the case in Sweden, where grant of neither title nor arms occurs, but where grant of arms to public bodies, such as local administrative units, is frequent. Historically, it was an easy passage from the arms of individuals to those of personifications. This is particularly evident in the military sphere, where the great crusading orders led to the many important orders of chivalry in the principal European countries. The Elephant of Denmark, the Golden Fleece of Spain and of Austria, the Holy Spirit of France, the Garter of England, and the Thistle of Scotland, were all preceded by the orders of military monks, and all have insignia that contain heraldic features and that occur in many arms illustrations. Most of the older bishops' sees have official arms; in the Anglican Church, the diocesan bishops as well as the missionary bishops have arms. In the Roman Catholic Church, the episcopal sees all have arms; and new arms are granted by the pope, who, as head of the Vatican state, is a temporal sovereign as well as spiritual head of the Church. The arms of the popes often contain charges that are added to their individual arms after their election to the papacy or to earlier ecclesiastical office.

The military orders of chivalry

Dominion and colonial arms are necessarily interconnected with royal arms, since the British crown in each country is or was the source of honour and must have granted arms to its various territories. Because of the vast extent of the former British Empire, the richest collection of arms of dominion is to be found in the numerous members of the Commonwealth. Canada, for instance, has arms for the sovereign of Canada as authorized by George V in 1921, and the 12 provinces of Canada have similar arms approved by the sovereign. Much the same is true of the former French colonies, though there was no sovereign to grant the arms. The arms of political units are used throughout the Western world. The cities and boroughs of the United Kingdom, for example, have their heraldry, as do the states of America.

The blazoning of these nonfamilial arms is conducted on the same principles as for family arms, except that the explanation of the charges is usually forthcoming.

Drawing by Wm. A. Norman



The arms of Canada, derived from the royal arms. The three maple leaves in the base, originally green, were altered to red to conform to the new national flag adopted in 1964. French traditions and ties are represented by the gold fleurs-de-lis and the white lilies in the compartment.

II. Historical development of heraldry

The exact date or place of origin of the heraldic system in western Europe in the 12th century is not known; neither are the precise reasons for its first employment. But limits can be drawn to indicate generally when heraldry began, and probable reasons for its emergence can be deduced. In the Bayeux Tapestry (in Bayeux, France), produced in the last quarter of the 11th century, is to be found a pictorial record that shows conditions in everyday life and in warfare at the time of the Norman Conquest of England. The English and Norman soldiers are armed alike. None of the Englishmen has a design of any kind on his shield or armour. In a few scenes, Normans or Frenchmen have designs on their shields that have a rough heraldic resemblance. In scene VIII of the tapestry, four of the followers of Guy, count of Ponthieu, have shields with these devices: some kind of creature holding what appears to be a fish in its mouth, a rough design emerging from the left side of the shield, a cross, and an animal rather like a sheep. In scene XII the messenger of William the Conqueror bears a winged creature on his shield, and this reappears in scenes XIII and XV. In scene XVIII a cross or a variant of it is seen on a shield, but probably as a boss (metal protuberance) to strengthen it. Scene LXXV has a Norman knight with a design of a birdlike creature on the shield, but generally the Normans shields have only bosses in the middle.

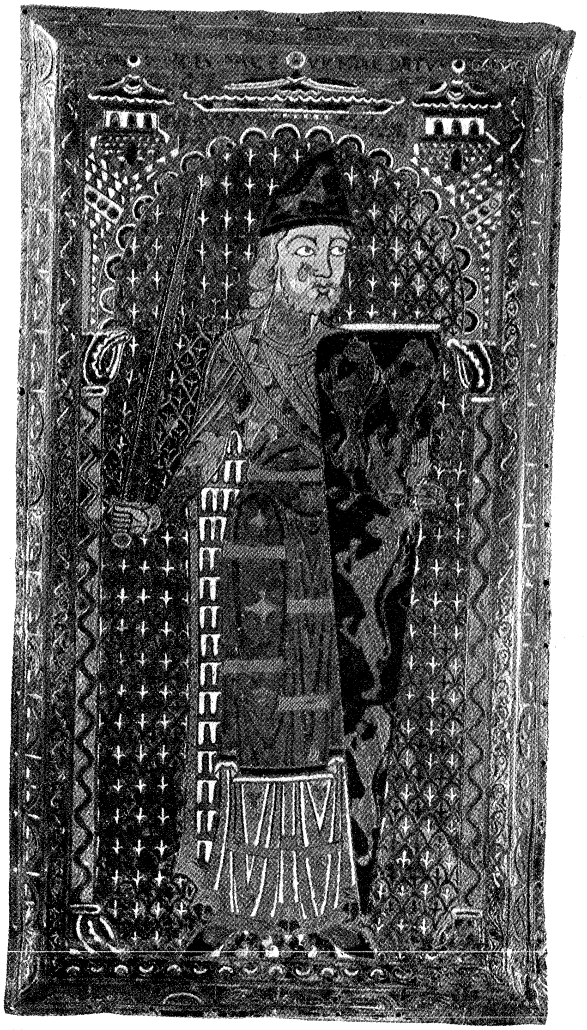
Bayeux Tapestry and its record of shields

In 1066, at the time of the Conquest, heraldry clearly did not exist. The most that can be said is that possibly some of the rudiments out of which it emerged were then present. Even a generation later, in 1095–99 at the time of the First Crusade, there is evidence that heraldry was not yet in use. *The Alexiad*, the history of the Emperor Alexius (reigned 1081–1118) written by his daughter, the princess Anna Comnena, contains a vivid description of the Frankish barbarians—as the Crusaders appeared in the eyes of the civilized Byzantines. The Princess gave a very careful account of the Crusaders' armour. She said that Alexius exhorted his archers to shoot at the Franks' horses rather than their riders because their armour rendered them practically invulnerable. "For the Frankish weapon of defence is this coat of mail, ring plaited into ring, and the iron fabric is such excellent iron that it repels arrows and keeps the wearer's skin unhurt. An additional weapon of defence is a shield which is not round, but a long shield, very broad at the top and running out to a point, hollowed out slightly inside, but externally smooth and gleaming with a brilliant boss of molten brass." In her description, therefore, are the same shields as depicted in the Bayeux Tapestry. Thus, all the pictures from later times of King Alfred the Great, William the Conqueror, or Charlemagne bearing coat armour can be disregarded as anachronisms.

EARLY ROOTS OF HERALDRY

It is not until a generation after the First Crusade that unmistakable evidence of heraldic designs appears. The earliest evidence is in an enamel (Musée de Tessé, Le Mans, France) made not later than 1151 showing Geoffrey, count of Anjou, bearing a shield azure with possibly four rampant golden lions (the exact number is not discernible because of the position in which the shield was depicted). The count was the son-in-law of King Henry I of England (reigned 1100–35), and, according to a chronicle, Henry in knighting Geoffrey bestowed upon him a shield that bore painted lions. In addition, from 1136 heraldic devices appear on seals. It is possible that the insignia were first used on seals and then spread to the shields of the warriors. Simultaneously, body armour was becoming all enveloping, and some means of distinguishing men in full armour became necessary. The heavy barrel-type helmet closed in the face of the wearer except for the opening of his visor, and a mix-

The first unmistakable evidence of heraldic designs



Plaque from the tomb of Geoffrey Plantagenet, count of Anjou, enamel, Limoges school, *c.* 1151–60. The stylized pattern of blue and white lining the figure's cloak represents a series of squirrel skins, called vair, frequently mentioned in blazons. In the Musée de Tessé, Le Mans, France.
Giraudon

ture of plate and ring mail enclosed the whole body. Yet another factor was the influence of the Crusades, in which men from a dozen different lands had to be distinguished from one another.

Whatever the reasons for its origin, heraldry within a few years was found throughout all of western Christendom. The first English king to bear arms was the crusader, Richard I the Lion-Heart (1157–99). The three gold leopards or lions of England have been used by every dynasty since his time.

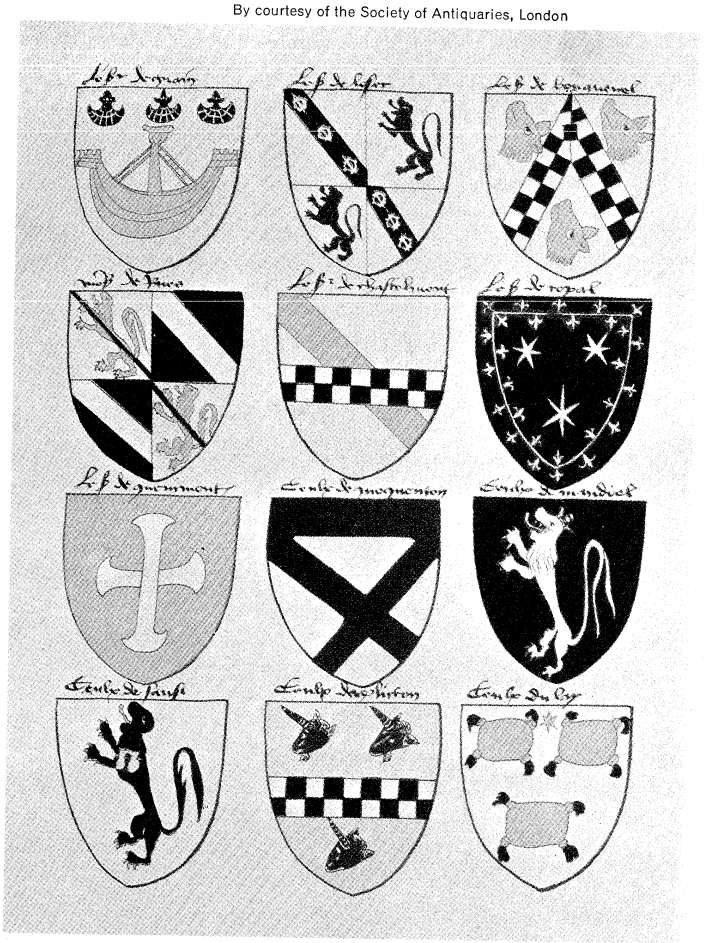
Seals. The earliest body of evidence of heraldic insignia is found in seals, large numbers of which have been preserved in England, France, and Germany, with fewer surviving in Spain and Italy. For the first century of heraldry, indeed, seals supply the bulk of information. It is, for instance, from seals that the rise and development of the English royal arms can be traced. In the first years of his reign, Richard I's seals show the design of a lion rampant to the left side. Some think that two lions were used since only half of the shield can be seen. At the end of his life, Richard's seals indicate clearly the rise of the three lions or leopards used by all sovereigns since. The adoption of the same coat of arms by subsequent dynasties is also found in the royal arms of Sweden and of Denmark; but unlike the English, those royal families place their family arms on an inescutcheon in the middle of the shield.

Roll of arms. Next to seals as evidence of heraldic usage come the rolls of arms, which in England date from about 1250. These are lists of arms, often with pictures drawn ("tricked") on the rolls, of persons who were present on a particular occasion, such as at a tournament or on a military expedition. England and Belgium (Flanders) are rich in the rolls of arms. France, Spain, and Scotland have fewer surviving examples. In place of the rolls, collections of painted books of arms have been preserved in Germany. A notable roll is the *Armorial de Berry*, dating about 1445, the work of a French herald, Gilles le Bouvier, who travelled widely and recorded arms borne in France, England, Scotland, Germany, Italy, and other European countries.

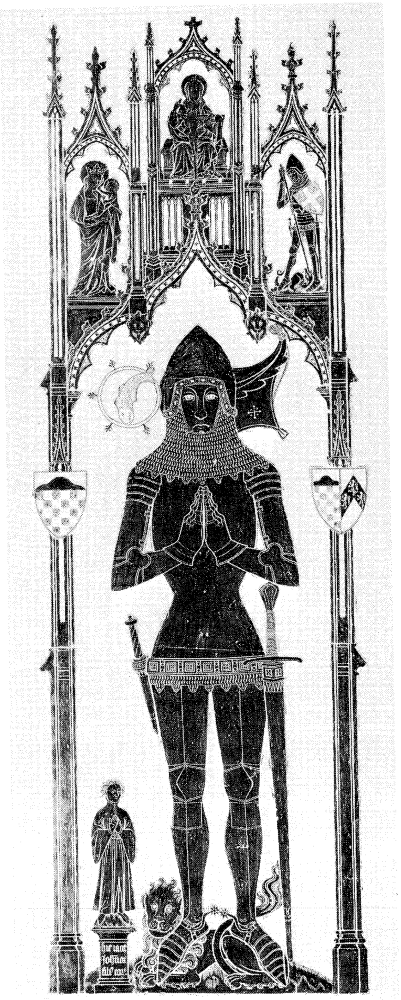
The *Armorial*

Records in stone and glass. Another very important source of information is to be found in representations on stone, wood, glass, and in books and engravings. Over the gateway of Bodiam Castle in Sussex can be seen the arms cut in stone of three owners of the castle, the families of Bodiam (who took their name from the place), Wardedieux, and Dalyngrygge. Of such arms nothing would be known without these centuries-old memorials. In Rome, many examples occur of the arms of various popes in their palaces and other buildings, for instance the bees of the Barbarini pope Urban VIII in the Palazzo Venezia. Heraldic glass is usually much more recent in origin but of immense value in supplying information as it is always in colour, while other memorials often are not. Very few churches of any great age in western Europe are without armorial illustration. Switzerland in particular has splendid memorials in stained glass; for example, the Dom, or main Protestant church in Berne, has windows that are aflame with glorious heraldic colours. Sweden has a fine collection of coloured plaques of arms in the House of Nobles in Stockholm; Denmark in the Castle of Fredericksburg houses the shields of the Knights of the Order of the Elephant, in which can be read the history of heraldry over several centuries.

By courtesy of the Society of Antiquaries, London



Page from the *Armorial de Berry*, by Gilles le Bouvier *c.* 1445, showing the simplicity of the early coats. In the collection of the Society of Antiquaries, London.



Sir Nicholas Hawberk, rubbing from his tomb brass, Cobham Church, Kent, 1407. Hawberk's arms appear alone on the dexter and impaled with the arms of his wife on the sinister.
By courtesy of Robert Lind

Church brasses. Brasses in churches are an important source of heraldic information. It was formerly the custom to put a brass tablet over the grave slab, and on this would be shown a figure of the deceased with his armorial bearings. Many fine examples of this are found in old English churches. A very fine collection of floor brasses is in the small church of Stopham in Sussex, which has been the memorial place of the local Barttellot family for many centuries. Also found in churches are hatchments, heraldic paintings on wood that were made for deceased persons and hung over their house doors, being later set up in the local church where they have often been preserved.

As for written material—such as official enactments, grants of arms, and books—nothing is dated earlier than the 14th century.

GROWTH OF HERALDRY AFTER THE 13TH CENTURY

The heralds. The initial meaning of the term herald is uncertain. Some authorities derive the word from two German words—*Heer*, "a host," and *Held*, "a champion" —not a very obvious etymology. It is clear that heralds were in existence from the 13th century, if not earlier. In their beginnings they were more or less menials, ranked with the jugglers and the minstrels, dependent on the great lords. First used as messengers who wore livery, they later were used in the jousts and tourneys to announce the contenders. To identify the knights, they had to know the arms on the shields, and from this grew their knowledge and skill in the art of heraldry.

Messengers and aides at jousts and tourneys

Heraldic colleges and offices. From this lowly origin have come the colourful figures of the English College of Arms, who now alone, save for the Scottish heralds, possess a high position in the modern world. The Lord Lyon, the head of the Scottish heralds, derives his office from a much higher source than do the heralds in other parts of Europe. The Sennachie, or official bard of the Scots' king, was the record keeper of the old Celtic kingdom of Scotland, and from the Sennachie is derived the Lord Lyon, a great officer of state in Scotland.

The older statements found in many books that the medieval heralds were either identical, or in some way connected, with the old Greek *kēryx* or Latin *fetialis* need only be stated to be dismissed. Since ancient times men have been found who, because their persons were accepted as sacred, were able to carry messages and other communications between nations either hostile or strange to one another. These ambassadors bore several names before the development of a diplomatic corps. In the earlier Middle Ages, for instance, churchmen, monks, or priests were used for this type of service. When William I the Conqueror sent a messenger to Harold II of England, it was a monk who carried William's denunciation of Harold. Heralds were not then in existence.

As they ascended the social scale, heralds began to serve as ambassadors between the different courts, a function that was still theirs in the first half of the 17th century. In 1627, for example, Sir Henry St. George was joined in a commission with Lord Spencer and Peter Young to present the insignia of the Order of the Garter to Gustavus II Adolphus, king of Sweden, who then knighted Sir Henry and granted him an augmentation to his arms showing the royal arms of Sweden.

At first every great noble had his herald, and the royal heralds were distinguished from the others by the greater importance of their masters. Gradually, it came about

By courtesy of the Court of the Worshipful Company of Tallow Chandlers, London



Detail of the "Grant of arms to the Worshipful Company of Tallow Chandlers," London, 1456. The herald in the illuminated initial wears a tabard that bears the king's arms front, back, and repeated on the sleeves, and a coronet as a mark of royal authority. In the collection of the Worshipful Company of Tallow Chandlers, London.

that a king would form his heralds into a college or corporation. The King of France did so in 1407; it was not until 1484 that the King of England followed by establishing the College of Arms, which has been housed for 300 years in the same building in London. The English College, sometimes called the Heralds' College, has outlived all similar elaborate establishments in Europe, except that in Scotland. Outside Great Britain, heraldic offices are found in the 20th century in Sweden, Denmark, the Republic of Ireland, and Spain.

The English College

The English College is under the control of the earl marshal, an office that has been hereditary for the past 300 years in the family of the duke of Norfolk. The holder of the dukedom is always earl marshal. Under him are 13 officers of arms; three kings of arms (Garter, Norroy and Ulster, and Clarenceux); six heralds (Windsor, Richmond, York, Lancaster, Chester, and Somerset); and four pursuivants (Rouge Dragon, Rouge Croix, Bluemantle, and Portcullis). These medieval names are derived from sources connected with royalty, titles, badges, or orders of knighthood. Pursuivants are "followers," or junior heralds. In Scotland the Lord Lyon is the head of the heraldic officers, of whom there are three heralds (Albany, Marchmont, and Rothesay) and four pursuivants (Carrick, Kintyre, Unicorn, and Ormond). In England and Scotland, the officers are not civil servants but members of the queen's household. In both countries there are also heralds extraordinary, who are appointed at times for special reasons or functions.

Heraldic visitations from 1530 to 1686

Records and grants. In England an important development came with the Heraldic Visitations. From 1530 in the reign of Henry VIII to 1686 in the reign of James II, commissions were issued by the sovereign to the heralds directing them to proceed to a county in England or Wales and to inspect the arms in use there. The records of the Visitations have been preserved and constitute a valuable body of genealogy as well as of heraldry. From the period of the Visitations the heralds built up huge collections of family history and began to record pedigrees in their registers.

From about a half century before the foundation of the College of Arms, the English heralds are found to be issuing grants of arms on behalf of the sovereign. This is some 300 years after the first appearance of heraldry, which obviously much antedated not only royal colleges or corporations of heralds but even the existence of heralds themselves. From this evidence, it seems clear that in the early days of heraldry men assumed arms to suit themselves without reference to any authority. A very simple coat of arms would not be difficult to invent. That three unrelated persons from three different counties could bear these same arms is not only not surprising but proof that the arms were self-chosen. When disputes over ownership among the three came up, the matter was referred to the king. His judgment was final, but it is noteworthy that one of the defeated, compelled to give up his arms, then consoled himself with a new coat that was also self-derived and -assumed. Unquestionably, the great majority of ancient coats of arms, borne before 1500, were never granted but were taken by the owners.

Writers on heraldry. The earliest writing on heraldry extant is a short treatise by Bartolo da Sassoferrato, whose *Tractatus de insigniis et armis* ("Tract on Insignia and Arms") was published about 1356. In his small book Bartolo describes the various categories of arms bearing and how they have been assumed. He refers specifically to arms granted by a prince and gives reasons for their value but asks why one man may not bear arms identical with those of another.

In 1355 Bartolo had been sent to Pisa from Perugia as an envoy to the Holy Roman emperor, Charles IV, from whom he received many privileges, including a grant of arms, which were the same as those of the Emperor as king of Bohemia but with changed tincture: "or a lion rampant with two tails gules." An American scholar, L.M. Mladen, remarked of this grant and others made by Charles IV at the same time: "Charles was in all probability the first ruler ever to grant arms. To my knowledge, no earlier occurrence has been found."

The first English heraldic writer was John of Guildford, or Johannes de Bado Aureo, whose *Tractatus de Armis* ("Tract on Arms") was produced about 1394. Then came a Welsh treatise, the *Llyfr Arfau* ("Book of Arms"). Nicholas Upton, a canon of Salisbury Cathedral, about 1440 wrote *De Studio Militari* ("On Military Studies"). John of Guildford's treatise was printed in 1654 with Upton's work and the *Aspilogia* of Sir Henry Spelman by Sir Edward Bysshe, Garter king of arms, who edited and annotated all three works. The whole was in Latin; no complete English version of Upton's book has been published.

These books are by authorities who were concerned with the realities of heraldry in their own day. A tendency away from actuality and toward the fanciful and absurd manifested itself from the end of the 15th century. Some of these farfetched conceits showed themselves in *The Boke of St. Albans* (1486). Yet by comparison with the vast mass of nonsense contained in the folios of the 16th century, those conceits were reasonable. The works of Sir John Ferne, *Blazon of Gentrie* (1586), Gerard Legh, *The Accedens of Armorie* (1562), and John Guillim, *A Display of Heraldrie* (1610) not only perpetuate the nonsensical natural history of olden days but are largely responsible for the belief that heraldic charges have a definite symbolic meaning and that they were granted as the reward of valorous deeds.

The trend to the nonsensical after the 15th century

Continental versus British heraldry. Much greater significance was attached in former times to heraldic insignia, though the attitude varied from country to country. Heraldry has become more widespread than at any other time, but as a sign of rank it has hardly any remaining value.

A distinction can be made between the Continent and Great Britain regarding medieval and later heraldry. The doctrine of the Seize Quartiers (16 quarterings) prevailed over most of the Continent but not in Britain. This theory required that all of a person's 16 ancestors (16 great-great-grandparents) should have been entitled to bear arms. This is known as the "Proof of the Seize Quartiers" and was the reason why Frederick II the Great of Prussia, though professing the views of the Enlightenment in the Age of Reason, diligently scrutinized his courtiers' quarterings. The theory is based on the rigidity of a noble caste that married only with its own kind. On the Continent, every member of a noble family is noble; hence the enormous numbers of titles. Similarly, the continental royalty tended to marry only with other royal families. As a result both royal and noble families formed a class apart from the bulk of the people.

Seize Quartiers

Continental heraldic insignia, therefore, from their origins until the late 18th century, provided symbols to indicate a higher caste and, in fact, were signs of nobility. Yet, strangely, in several countries heraldry was in wide general use as a means of identification, serving in the same way as a surname. In France, for example, it is abundantly clear that from the 13th century, not only the bourgeoisie of the towns but also the peasants bore heraldic arms. The usage had percolated down from the noble class. The earliest example of the use of arms by a peasant is that of Jaquier le Brebiet in 1369, whose arms show a punning allusion to his name (*brebis*, "sheep"—three sheep held by a girl). In other European lands—Hungary and the Low Countries—burgher or peasant arms were also found, but neither in these lands nor in France were the possessors regarded as noble.

Peasant and bourgeois arms

In France the regulations of arms followed a quite different course from those in England. Although King Charles VI had in 1407 led the way in creating a college of arms, his heralds lost influence over the next two centuries. They had no power, unlike their Scottish and English counterparts, to grant arms and they gradually faded into insignificance. To overcome the loss, Louis XIII in 1615 appointed a *juge géneral d'armes* ("general judge of arms"), an official whose powers resembled those of the Lord Lyon. The French royal government showed itself very broad-minded on the possession of arms. A decree of Louis XIV in 1696, designed to raise money, ordered all persons who bore arms to register

them. Even those who were not part of the arms-bearing population were forced to buy arms. Later, in 1760, an ordinance was framed by which the lesser townsfolk, artisans, and peasants were to be excluded from the use of arms—after these classes had used them for 400 years. But the Parlement of Paris refused to allow the ordinance to be implemented. Later, during the French Revolution (1789), arms were suppressed as signs of feudalism.

The view of arms held in England was and is quite different. No such thing as a noble caste has ever existed in England. Only the reigning peer and his wife are regarded as noble; the rest of the family are commoners, and only a few of them bear what are called courtesy titles. Moreover, except for the Hanoverian (1714–1837) and Victorian (1837–1901) epochs, the royal house has not necessarily married royalty but much more often the nobility, a practice to which it has returned in the present century. As a result, the Continental doctrine of Seize Quartiers does not apply in England. It cannot, since noble and non-noble are so mixed. Nor are there any such things as non-noble arms. All arms are on the same basis; all are signs of gentility—nobility, in fact. Arms then have long had a high social significance in England; those who possess them have social prestige. The situation is somewhat different in Scotland where arms are bound up with clanship and family solidarity. To the Scotsman arms are not so much a matter of social status as of family or clan, and he proves his right to them by process of law in Lyon Court.

A matter of clan and family in Scotland

TWENTIETH-CENTURY HERALDRY

The survival of the heraldic tradition. *France and Italy.* Without a monarchy, heraldry can still flourish, but it does not normally do so. The French Revolution abolished arms, which returned with the monarchy, and now, although France is a republic, a person assuming arms to which he is not entitled may be prosecuted. In the same way, it is not permissible to assume a name of a great family. In England, however, the assumption of the name Windsor (the name of the British royal family) is possible for anyone. There is, however, no recognized heraldic authority in France, nor for that matter in other European countries that have abolished their monarchies. Most such republican states have associations that seek to maintain heraldic standards. Thus in Italy the Collegio Araldico (Heraldic College) consists of experts whose main object is to promote heraldic and genealogical studies. An association of nobles was formed as the National Heraldic Council of the Italian nobility, under the authority of former king Umberto II; it tries to regulate the use of arms and titles.

Many changes in heraldry have taken place since 1945, as it is not a static subject.

Communist countries. In Communist European countries, the study of genealogy and heraldry has been generally suppressed. Since about 1956 it has been possible to obtain much more statistical information from the U.S.S.R. than formerly, but no heraldic data are supplied. Much the same is true of the other Communist countries in Europe, except Czechoslovakia. In Communist lands, heraldry is viewed as it was by the French revolutionaries, as part of the feudal past, although there is reason to believe that in these countries heraldic archives are carefully preserved.

Ireland. A different development occurred in two of the republics that have emerged from the British Empire —Ireland and South Africa—both of which have set up their own heraldic offices. As early as 1382, there was an Ireland King of Arms who was responsible for all matters armorial in that country. The last holder of the office died in 1487, and in 1553 Edward VI created a new armorial king under the title of Ulster, to control bearings throughout Ireland. His place of business was in Dublin Castle. When the Irish Free State, or Eire (now the Republic of Ireland), was established in 1921–22, the Ulster Office was reserved as an appointment of the British crown with the then-current Ulster to hold office for life. After his death in 1940, an arrangement was made between the British and Irish governments by which the heraldic office in Dublin Castle with its records was taken over by the Irish authorities. Photostat copies were made of the records and sent to the College of Arms, London. The Irish government appointed a Chief Herald of Ireland, and the Ulster Office became known as the Genealogical Office. A civil servant was then appointed as Chief Herald of Ireland. The office of Ulster King has now been united with that of Norroy King in the College of Arms, London. The Irish Herald carries out the duties formerly performed by Ulster in the 26 counties of the Republic of Ireland; Norroy and Ulster has jurisdiction over the six counties of Northern Ireland (Ulster).

Chief Herald of Ireland

South Africa. In South Africa an act was passed in 1962 under which was established a Bureau of Heraldry and a Heraldry Council for the grants, registration, and protection of coats of arms, badges, and other emblems. A state herald is appointed as head of the Bureau of Heraldry. The Heraldry Council consists of the state herald and at least seven other members appointed by the government minister responsible.

The United States. There has been a remarkable evolution of heraldry in the United States. Ever since the American Revolution, the use of arms, especially of arms of English families with whom the users were related or whose surname they bore, has continued. The English College of Arms claims heraldic jurisdiction over persons of English and Welsh descent (Wales has been reckoned with England in this and all other administrative matters since the union of England and Wales, 1542). The Lord Lyon of Scotland claims jurisdiction likewise over persons of Scottish descent throughout the world. In addition, the English College at one time claimed a worldwide imperial jurisdiction over anyone who could be brought within the definition of British subject. Under this jurisdiction even the Indian princes were occasionally granted arms by the College of Arms, although they were not British subjects but independent rulers who had entered into treaty relations with the British crown. Many Americans have been granted arms by the college by virtue of their descent from English or Welsh forebears, or by the Lord Lyon if they are of Scottish descent. Irish Americans often were granted arms from Dublin, either from the Ulster King or his successor. Americans of Northern Irish descent have been granted arms by Norroy and Ulster. In addition, there are several states of the United States that were formerly Spanish territory, and the Spanish Kings of Arms, the equivalent of the English and Scottish heralds, exercised a heraldic authority over persons of Spanish descent in the old Spanish Empire. By extension, they have recently granted arms to Americans who are resident in those formerly Spanish states but who are not of Spanish descent.

Heraldry in states from the old Spanish Empire

To these classes of arms obtained by Americans from overseas must be added such instances as the award of arms to Pres. Dwight D. Eisenhower in Denmark. Also, Americans of French, German, Italian, Polish, and other European origins have inherited from their immigrant ancestors arms once granted or recorded by heraldic authorities no longer in existence. All these classes of arms share one feature whatever their origin: they are hereditary honours granted to American citizens by other countries. As they do not carry titles, they do not contravene the principle of the American Constitution on this subject. An American who receives a knighthood of some foreign state possesses only an honorary knighthood; he is not "Sir." But for the citizen of an independent sovereign power to approach and receive from another power a hereditary honour has seemed to many Americans an undesirable procedure. There have been and still are thousands of assumptions of arms by Americans on consideration of mere mail-order salesmanship; arms in these transactions are, at the best, supposed to be those "of the name" (that is, belonging to a name rather than a person), a view for which no justification exists.

Endeavours have been made to set up American authorities who would not only record but also grant arms. The New England Historic Genealogical Society of Boston appointed a Committee on Heraldry that since 1928 has issued rolls of arms, in which have been entered the

names and arms of those who have submitted their claims to its judgment. The use of this method of issuing or publicizing arms recalls the usage referred to above, which has not been practiced in Europe for the past 400 years. In the introduction to the second roll (1932) it is stated: "There is certainly no legal reason, perhaps no reason at all, why an American gentleman should not assume *in more majorum* any new coat that pleases his fancy, but he should not assume an old coat, for if he does, he is very likely denying his own forefathers and he surely is affirming what he has no sufficient reason to be true." Not only have British-derived arms been registered but also continental European arms. In addition, the committee has assisted inquirers in devising new coats of arms, not only for schools, colleges, and other bodies, but also for individuals. In the introduction to the first roll a very reasonable view toward heraldry was expressed: "Taking into consideration the early history of coat armour there seems to be no reason in this country at least, why anyone provided he observes the simple rules of blazon and does not appropriate the arms of another, may not assume and use any coat he desires." The American College of Heraldry and Arms, Inc., was established in the state of Maryland in 1966. It has two divisions: The American College of Arms, which is concerned with the arms of individuals, their registration, and more important, the granting of arms; and the College of Arms of the United States, which deals with arms, crests, standards, devices, and the like for all manner of corporate concerns. Arms were granted by the college to Pres. Lyndon B. Johnson during his term of office, to Pres. Richard M. Nixon, and to Vice Pres. Spiro T. Agnew.

Revival of England's ancient Court of Chivalry

England. Heraldic development has also occurred in England, where in 1954 the ancient Court of Chivalry was revived. This was once the court of the lord high constable and of the earl marshal, and it dealt with matters relating to knights and gentlemen. Although it was concerned also with matters of military discipline, it was not the forerunner of the modern court-martial in the armed forces. The court gradually declined in the 17th and 18th centuries and had not sat from 1735 until its revival.

The office of lord high constable has long ceased to be hereditary or of permanent status in England. During coronations, however, a constable is appointed for the occasion. Therefore, in the revived court that sat in 1954 to deal with a test case, the *Manchester Corporation* v. the *Manchester Palace of Varieties*, the earl marshal (the duke of Norfolk) presided with a surrogate, who was the lord chief justice of England appearing in his capacity of a doctor of civil law. As a result of the sitting, the jurisdiction of the court was confirmed, and the City of London recorded its crest and supporters.

The unorganized condition of heraldry in many European states has spurred private attempts to bring some order into the field. The movement known as the International Congresses of Heraldry and Genealogy began in 1928 with a meeting in Barcelona, Spain. A second Congress was held in Rome and Naples in 1953, and from that time regular meetings occurred at two- or three-year intervals. Out of these was established an international organization, El Instituto Internacional de Genealogia y Heraldica (The International Institute of Genealogy and Heraldry), with its office in Madrid.

Uses of heraldry for study and verification. The incidental uses of heraldry, apart from its general significance in providing distinguished symbols, are considerable. Heraldry illustrates much in history and literature that is otherwise obscure. Heraldry on buildings, in manuscripts, and in paintings is of immense value for purposes of identification. It serves to link one person with another, to connect families, and to disclose origins of states and of institutions. With the use of heraldry in connection with genealogy, from which it cannot easily be separated, much that is difficult to follow becomes easy to interpret. In every building that contains armorial engravings, or other pictures of arms, there is in fact a concise history of the place, which enables the onlooker who understands the subject to have a useful clue to the background and details of the building.

In law, the place of heraldry is sometimes, as in modern Scotland, very precise; in England, on the other hand, it is most involved. To understand the latter is to gain an insight into the development of English law. Similarity of arms does not always indicate identity of family. Identity of arms may denote merely assumption of devices. Because of the English laws of inheritance, not only an estate but a surname and arms can pass by eventual succession to persons unconnected by blood with the original owner. Thus in the families of Lytton, Carew, and branches of Trelawny, instances occur in which possession of the name and arms is no proof, but the contrary, of blood relationship. Sometimes, however, identity of a name with a puzzling discrepancy in the arms can only be explained by a study of the family's heraldry.

BIBLIOGRAPHY. There are large numbers of guides to heraldry, the majority of a sameness that makes them apparently derived from an old original. Among the most useful for the beginner are: A.C. FOX-DAVIES, *Heraldry Explained* (1906, reprinted 1971), a small, well-illustrated book, written by the clearest expositor of heraldry as practiced today; CHARLES MACKINNON, *The Observer's Book of Heraldry* (1966), a very useful and clearly written work; and SIR ANTHONY R. WAGNER, *Heraldry in England* (1946, reprinted many times), a very brief account but with 15 plates in colour. L.G. PINE, *Teach Yourself Heraldry*, new ed. (1972), is designed for the beginner and has a glossary of terms. Also very good for the beginner is J.R. PLANCHE, *The Pursuivant of Arms or Heraldry Founded upon Facts* (1852).

More specialized works include an English translation of the *Armorial de Berry*, originally written *c.* 1445, which can be found in the *Proceedings of the Society of Antiquaries of Scotland*, 72:84–114 (1938); A.C. FOX-DAVIES, *A Complete Guide to Heraldry* (1909, revised and reprinted many times), a very helpful guide to anyone concerned with blazons either drawn or written (with nearly 800 illustrations); and *Armorial Families: A Directory of Gentlemen of Coat Armour*, 2 vol. (1929–30, reprinted 1970), a rich store of coloured and black-and-white illustrations, which gives the student the opportunity to practice and to observe blazons. The writings of OSWALD BARRON should be read as a correction to some of Fox-Davies' theories on heraldic history. Barron's works contain much fascinating heraldic lore, delightfully set out, learned yet not abstruse. Some of his best work may be found in the 12 volumes of *The Ancestor* which he edited in 1902–05. This series also includes many informative articles by writers of the calibre of J.H. ROUND and SIR HENRY MAXWELL-LYTE. Two volumes now reprinted in one constitute *A Treatise on Heraldry: British and Foreign* by JOHN WOODWARD and GEORGE BURNETT (1892, reprinted 1969), which contains some inaccuracies because the authors were not able to check all instances of heraldic descriptions sent to them. It is, however, most rewarding to the student, beautifully illustrated, with examples of practices drawn from all over Europe. SIR THOMAS INNES, *Scots Heraldry*, 2nd ed. rev. (1956), not only introduces the reader to Scots heraldry but explains much in heraldic practice that may not otherwise be clear. A work by another Scottish writer who gives a good view of heraldry in the international sense is *The Nature of Arms* by G.R. GAYRE (1959). For information on Irish heraldry, see SIR CHRISTOPHER and ADRIAN LYNCH-ROBINSON, *Intelligible Heraldry* (1948). *Boutell's Heraldry* has been edited and re-edited again and again since its first appearance in the last century. The editions of V. WHEELER-HOLOHAN, C.W. SCOTT-GILES, and J.B. BROOKE-LITTLE are all very useful. A large book, *Shield and Crest* by JULIAN FRANKLYN (1960), gives much otherwise not easily obtained information.

Very little continental work on heraldry has been translated into English. This applies not only to the lesser known European languages, as in Scandinavia, but also to French, Spanish, Italian, and German. Perhaps the best course is to consult periodical publications like *The Augustan*, which contains articles from a wide variety of sources, more than are likely to be found in any one European country. Several non-English-language encyclopaedias, such as the *Enciclopedia Italiana* (1929–37), have excellent articles on heraldry. Many non-English writers do transcend their national boundaries in writing on the subject. REMI MATHIEU in *Le Système héraldique français* (1946), not only writes illuminatingly of French heraldry, but helps to clarify the whole subject. On Spanish heraldry, see JOSE ASENSIO Y TORRES, *Tratado de heraldica y blason*, 3rd ed. rev. (1854, reprinted 1929); and LUCAS DE PALACIO, *De Genealogia y heraldica* (1946), the latter author being especially interested in the connection between totem-

ism and heraldry; the Japanese *mon* is exhaustively dealt with by CARROLL PARISH in *The Augustan*, vol. 11, no. 1 (1968).

Other works that will help the student after he has acquired a sound knowledge of this subject are: SIR ANTHONY R. WAGNER, *Heralds and Heraldry in the Middle Ages* (1939) and *Historic Heraldry of Britain* (1939); and L.G. PINE, *The Story of Heraldry* (1952, reprinted many times), an account of much controversial matter, including the present English heraldic position in law; G.D. SQUIBB, *The High Court of Chivalry: A Study of the Civil Law in England* (1959); C. PAMA, *Lions and Virgins* (1965), which discusses the history of arms in South Africa from 1487 to the present; and *International Heraldry* (1970), concerned with heraldry throughout the world.

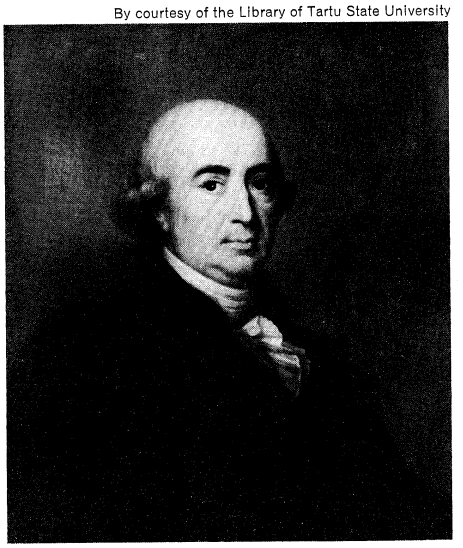
Works of reference in English are easily available. SIR J.B. BURKE, *General Armory* (1883, reprinted 1969), gives the description under alphabetical order of surnames of scores of thousands of coats of arms. JOHN WOODY PAPWORTH, *Ordinary of Arms* (1858–74, reprinted 1961), is the equivalent of the *General Armory* turned inside out, the idea being to enable the seeker to trace a coat of arms when he does not know the owner's name. Both books contain inaccuracies, small blame to the compilers who had to deal with much uncheckable material, at least at the time when they wrote. JAMES PARKER, *A Glossary of Terms Used in Heraldry*, new ed. (1970), with 1,000 illustrations, is a helpful book. Editions of *Debrett's Peerage*, of *Burke's Peerage*, and of *Burke's Landed Gentry* abound in illustrations and descriptions of arms. Recently published sources on the European continent, such as the *Grand Armorial* of France and the *Libro d'Oro* of Italy, have useful information on the heraldry of these countries.

(L.G.P.)

Herder, Johann Gottfried von

Herder, a German critic, philosopher, and Lutheran theologian of the 18th century, exerted a great influence in the history of German thought through his broad scholarship and prolific and fertile writings. His significance lies in his capacity to transcend the perspectives of his day, which were limited by the Enlightenment's extolling of reason, and to interpret human culture and history afresh. His influence, augmented through his contacts with the young Goethe, made him a harbinger of the Romantic movement.

By courtesy of the Library of Tartu State University



Herder, oil painting by Gerhard von Kügelgen, 1808. In the Library of Tartu State University.

Herder was born on August 25, 1744, at Mohrungen, in East Prussia, the son of poor parents. Beginning in the summer of 1762, he studied theology, philosophy, and literature at Königsberg, coming into close contact with Immanuel Kant, the founder of critical philosophy, as well as with Johann Georg Hamann, one of the Enlightenment's prominent critics.

Early travels and first works. In November 1764 Herder went to teach and preach in Riga. There he published his first works, which included two collections of fragments, entitled *Über die neuere deutsche Literatur: Fragmente* (1766–67; "On Recent German Literature: Fragments") and *Kritische Wälder, oder Betrachtungen die Wissenschaft und Kunst des Schönen betreffend* (1769 and 1846; "Critical Forests, or Reflections on the Science and Art of the Beautiful").

Self-discovery on a sea voyage

In the summer of 1769, he set out on a sea voyage from Riga to Nantes, which brought him a deeper understanding of his destiny. The *Journal meiner Reise im Jahr 1769* (1769; "Journal of My Voyage in the Year 1769"), which he concluded in Paris in December, bears witness to the change that it effected in him. Herder saw himself as a groundless being who had left the safe shore and was journeying into an unknown future. It would be his vocation to unveil that future through insights gained from the past, so that its character might be felt by his contemporaries. In his fruitful criticisms of his own time, Herder anticipated the possibilities of the developments generations ahead, among them, the ideas of Goethe, the brothers August Wilhelm and Friedrich von Schlegel, and Jacob and Wilhelm Grimm in poetical and aesthetic theory; Wilhelm von Humboldt in the philosophy of language; G.W.F. Hegel in the philosophy of history; Wilhelm Dilthey and his followers in epistemology; Arnold Gehlen in anthropology; and the Slav nationalists in political thought.

During a visit to Strasbourg, where he arrived in September 1770, as the companion of Prince Peter Frederick William of Holstein, Herder experienced a momentous meeting with the young Goethe, who was stirred to recognize his own artistic faculties through Herder's observations on Homer, Pindar, Shakespeare, and on literature and folk songs.

Career at Bückeburg. In April 1771 Herder went to Bückeburg as court preacher. The works that he produced there were fundamental to the *Sturm und Drang*, a literary movement with Promethean and irrationalist motifs, without which German Romantic literature could not have arisen. In the Romanticism Herder espoused, the medium of thought is feeling (*Gefühl*), which he compared to the sense of touch. Whereas sight apprehends things at a distance, feeling enjoys an immediate experience of reality, which it apprehends as a power reacting against man's own vital energy. Yet man at the same time experiences his own body, in which his vital power asserts itself against the world. At the moment when man recognizes the limits imposed by his environment without becoming dependent on it, a balance of forces is achieved between the two in which the individual body is converted into the aesthetic gestalt (or integral structure) and the identification of the individual with reality is consummated. (See his book *Plastik*, 1778.)

Among the works of this period is his *Abhandlung über den Ursprung der Sprache* (1772; "Essay on the Origin of Language"), which finds its origin in human nature. For Herder, knowledge is possible only through the medium of language. Although man and the world are united in feeling, they separate themselves in consciousness in order to link themselves anew in the "intentional," or object-directed, act in which the objective meaning of a *word* is rooted. Thus, what earlier had been apprehended dimly but not specifically recognized in feeling is expressly designated. Feeling and reflection thus interpenetrate each other; and the word, being at once sound and significance, is the cause of this union. Every signification of something therefore includes a certain emotional attitude toward it that reflects the particularity and the outlook of its users. Thus, the structure of language is a true image of human nature.

Whereas the psychologists of the time were carefully distinguishing man's various faculties (conation, feeling, knowledge), Herder stressed the unity and indivisible wholeness of man's nature. Consciousness and *Besonnenheit* ("reflective discernment") are not simply "higher" faculties added to an animal foundation; instead, they designate the structure of man as a whole with qualitatively unique human desire and human sensitivity. But, because in man instinct and sensitivity are subject to reflection, or "broken off" (*gebrochen*), man is "the first liberated member of creation."

Beginning of Herder's philosophy of history

Herder's philosophy of history also began to take form at this time, growing up from his attempt to use the past in order to assess the present situation and future probabilities. He had already outlined in the *Fragmente* the scheme of a typical historical development on the analogy of the ages of a man's life. By this means he tried to determine the situation of German poetry that was then current. The essay on Shakespeare and *Auch eine Philosophie der Geschichte zur Bildung der Menschheit* (1774; "Another Philosophy of History Concerning the Development of Mankind"), opposing Rationalism in historiography, were the first writings to show a deeper understanding of historical existence as the product of the contradiction between individuation and the whole of history; this contradiction itself forms the logical basis of historical development. If two forces are in conflict, one can be seen as striving to persevere and to emerge from the whole as an individual structure. Yet the whole is not satisfied with any single form: in historical catastrophes it frees itself to shape a new form of things, which is shattered again in turn when its time is past. The individual is not only an end but also a blind, unfree instrument taken or rejected by God. Even the philosopher can see the future only by tracing its conditions from patterns of past development, in order to counteract it.

Further works prepared during this period were his *Älteste Urkunde des Menschengeschlechts* (1774–76; "Oldest Records of the Human Race") on Hebrew antiquities, and his *An Prediger: Funfzehn Provinzialblätter* (1744; "To Preachers: Fifteen Provincial Papers"). Two especially important works were his essay on Shakespeare and "Auszug aus einem Briefwechsel über Ossian und die Lieder alter Völker" (1773; "Extract from a Correspondence About Ossian and the Songs of Ancient Peoples"), published in a manifesto to which Goethe and Justus Möser, a forerunner of *Sturm und Drang*, also contributed. As Herder showed in his exposition of Shakespeare and Homer, in the genuine poetic utterance, hitherto-hidden aspects of man's life are revealed by virtue of the creative function of language. "A poet is the creator of the nation around him," he wrote, "he gives them a world to see and has their souls in his hand to lead them to that world." Poetic ability is no special preserve of the educated; as the true "mother tongue of mankind" (Hamann), it appears in its greatest purity and power in the uncivilized periods of every nation; this for Herder was proved by the Old Testament, the *Edda*, and Homer: hence Herder's concern to retrieve ancient German folk songs and his attention to Norse poetry and mythology, to the work of the minnesinger, and to the language of Luther.

First years at Weimar. Thanks to Goethe's influence, Herder was appointed general superintendent and consistory councillor at Weimar in 1776. There, anticipating Goethe, he developed the foundations of a general morphology, which enabled him to understand how a Shakespearean play, for instance, or the Gospel According to John, in the historical context of each, was bound to assume the individual form that it did and no other. Herder's method achieves its results by recognizing contradictions and by resorting to a higher unity—a method by which Herder earns a place in the history of dialectical logic: compare his treatise "Über die dem Menschen angeborene Lüge" (1777; "On the Innate Lie in Man").

It was at this time also that Herder completed his transition to Classicism. Among the works of this period are *Vom Erkennen und Empfinden der menschlichen Seele* (1778; "Of the Knowing and Sensing of the Human Soul"), *Briefe, das Studium der Theologie betreffend* (1780–81; "Letters Concerning the Study of Theology"), *Vom Geist der ebräischen Poesie* (1782–83; Eng. trans., *The Spirit of Hebrew Poetry*, 1833), and his collection of *Volkslieder* (1778–79; "Folk songs"). Herder regarded poetry as a mode of coming to terms with reality. Whereas most of his contemporaries saw it either as a product of learning or as a means of amusement, he considered poetry to spring from the natural and historical environment experienced by feeling, rather as an involuntary reaction to the stimulus of events than as a deliberate act. Such feeling is the organ of a dynamic relationship between man and the world, which is expressed far more readily in the sounds, stresses, and rhythms of speech than in an image. This "voice of feeling" achieves the status of art only when it is detached from the man and from the historical environment that created it and becomes rounded off to constitute a world by itself.

Summit and later years of his career. Herder's work at Weimar reached its peak in *Zerstreute Blätter* (1785–97; "Sporadic Papers") and in the unfinished *Ideen zur Philosophie der Geschichte der Menschheit* (1784–91; Eng. trans., *Outlines of a Philosophy of the History of Man*, 1800). In the latter work, the result of his intercourse with Goethe, Herder attempted to demonstrate that nature and history obey a uniform system of laws. Already in the development from earth to mankind, a striving of forces was at work, aiming to balance one another by generating determinate forms or individual existents. This same phenomenon could be observed as a law of "humanity" in man's communal life, in which contending forces are reconciled. At any passing moment the measure is individual, but the principle of the development toward form is general. But too often man, in his freedom, works against nature, for his sense of the measure of things and his reason are immature. One must trust, however, that growing insight and goodwill will lead men to act according to the truth that they recognize and, through the conflict of nations, will reach the equilibrium of a structure embracing all mankind.

The attempt to unite nature and history

The basic premises underlying the *Ideen* are resumed in the dialogues *Gott: einige Gespräche* (1787; 2nd ed., *Einige Gespräche über Spinozas System*, 1800; "Several Discourses on Spinoza's System"), in which Herder combines the views of the Rationalists Gottfried Wilhelm Leibniz, Benedict de Spinoza, and Anthony, earl of Shaftesbury.

Financial difficulties, differences of opinion over the French Revolution, and, above all, his self-assertive nature, which could not bear the proximity of a greater man, led to an estrangement of Herder from Goethe. On Herder's side this resulted in a bitter enmity toward the whole Classical movement in German poetry and philosophy. His *Briefe zu Beförderung der Humanität* (1793–97; "Letters for the Advancement of Humanity") and his *Adrastea* (1801–03; treatises on history, philosophy, and aesthetics) emphasized the didactic purpose of all poetry, thus contradicting that very theory of the autonomy of the work of art that he himself had helped to establish. With the *Christliche Schriften* (1794–98; "Christian Writings"), the *Metakritik zur Kritik der reinen Vernunft* (1799; "Metacritique of the Critique of Pure Reason"), and the *Kalligone* (1800; metacritique of Kant's *Critique of Judgment*), Herder began his attack on Kant, whose philosophy he saw as a threat to his own historical view of the world. In this attack he had the support of Christoph Wieland, an influential poet and novelist.

Herder died at Weimar on December 18, 1803. The first collected edition of Herder's works was produced by his widow (45 vol., 1805–20). There is also a critical edition by B. Suphan (33 vol., 1877–1913; reprint, 1967–68).

BIBLIOGRAPHY. The authoritative biography is still R. HAYM, *Herder nach seinem Leben und seinen Werken*, 2 vol. (1880–85; 2nd ed., 1954). Important supplements are provided by A. GILLIES, *Herder* (1945); and by R.T. CLARK, *Herder: His Life and Thought* (1955, reprinted 1969), which has a detailed bibliography. The leading monograph written according to a comparative method is T. LITT, *Kant und Herder als Deuter der geistigen Welt* (1930; 2nd ed., 1949). For special aspects, see R. STADELMANN, *Der historische Sinn bei Herder* (1928); F. MEINECKE, *Die Enstehung des Historismus*, 2 vol. (1936). F.M. BARNARD, *Herder's Social and Political Thought* (1965); J.K. FUGATE, *The Psychological Basis of Herder's Aesthetics* (1966); and H.B. NISBET, *Herder and the Philosophy and History of Science* (1970). For further bibliography, see K. GOEDEKE, *Grundriss zur Geschichte der deutschen Dichtung*, 3rd ed., vol. 4 (1916); E. KEYSER (ed.), *Im Geiste Herders* (1953); W. WIORA (ed.), *Herder-Studien* (1960); and F.W. KANTZENBACH, *Johann Gottfried Herder in Selbstzeugnissen und Bilddokumenten* (1970).

(H.D.I.)

Heredity

Heredity is a complex of biological processes that result in progeny resembling their parents in many characteristics; progeny are not exact duplicates of their parents but usually differ in many traits. This difference is called variation. Heredity and variation, two sides of the same coin, are the subject matter of the science of genetics (*q.v.*).

The following sections are included in this article:

I. Early conceptions of heredity
  Pre-Mendelian discoveries and speculations
  Biological and legal inheritance
II. Mendel's experiments and their significance
  Discovery and rediscovery of Mendel's laws
  Universality of Mendel's laws
  Apparent exceptions to Mendelian inheritance
III. Physical basis of heredity
  Chromosomes and genes
  Linkage of traits
  Heredity and nucleic acids
  Heredity and development
IV. Heredity and environment
  Preformism and epigenesis
  Heredity in health and disease
  Heritability
V. Changing heredity
  Heredity as a source of constancy and change
  Useful and harmful changes
VI. Heredity and evolution
  Natural selection and Darwinian fitness
  Varieties of natural selection
  Natural selection in operation
  Genetics of race and species differences
  The sudden origin of new species
VII. Heredity and applied science
  Medico-legal applications and genetic counselling
  Outbreeding and inbreeding
  Genetic load and hybrid vigour
  Genetic improvements of animals and plants
  Improving mankind

I. Early conceptions of heredity

PRE-MENDELIAN DISCOVERIES AND SPECULATIONS

Heredity was for a long time one of the most puzzling and mysterious phenomena of nature. This was so because the sex cells, which form the bridge across which heredity must pass between the generations, are usually invisible to the naked eye. Only after the invention of microscopes early in the 17th century, and the discovery of the sex cells, could the essentials of heredity be grasped. Before that time, Aristotle (4th century BC) speculated that the relative contributions of the female and the male parents were very unequal—the female was thought to supply what he called the "matter" and the male the "motion." The *Institutes of Manu*, composed in India between AD 100 and 300, consider the role of the female like that of the field and of the male like that of the seed; new bodies are formed "by the united operation of the seed and the field." In reality both parents transmit the heredity pattern equally, and on the average, children resemble their mothers as much as they do their fathers. Nevertheless, the female and male sex cells may be very different in size and in structure; the mass of an egg cell is sometimes millions of times greater than that of a spermatozoon.

The ancient Babylonians knew that pollen from a male date palm tree must be applied to the pistils of a female tree to produce fruits. R.J. Camerarius showed in 1694 that the same is true in corn (maize). C. Linnaeus in 1760 and J.G. Kölreuter, in a series of works published from 1761 to 1798, described crosses of varieties and species of plants. They found that the hybrids were on the whole intermediate between the parents, although in some characteristics they may be closer to one and in others closer to the other parent. Kölreuter compared the offspring of reciprocal crosses; *i.e.*, of crosses of variety *A* functioning as a female to variety *B* as a male and the reverse, variety *B* as a female to *A* as a male. The hybrid progenies of these reciprocal crosses were usually alike, indicating that contrary to the belief of Aristotle, the hereditary endowment of the progeny was derived equally from the female and the male parents. Many experiments on plant hybrids were made in the 1800s. These investigations revealed that hybrids were usually intermediate between the parents. They more or less incidentally recorded most of the facts that later led Gregor Mendel (see below) to formulate his celebrated rules and to found the theory of the gene. Apparently, none of Mendel's predecessors saw the significance of the data that were being accumulated. The general intermediacy of hybrids seemed to agree best with the belief that heredity was transmitted from the parents to offspring by "blood," and this belief was accepted by most 19th-century biologists, the evolutionist Charles Darwin being included among these.

Blood theory of heredity

The blood theory of heredity, if this notion can be dignified with such a name, is really a part of the folklore antedating scientific biology. It is implicit in such popular phrases as "half blood," "new blood," "blue blood," etc. It does not mean that heredity is actually transmitted through the red liquid in blood vessels; the essential point is the belief that a parent transmits to each child all its characteristics and that the hereditary endowment of a child is an alloy, a blend of the endowments of its parents, grandparents, and even more remote ancestors. This idea appeals to those who pride themselves on having a noble or otherwise remarkable "blood" line. It strikes a snag, however, when one observes that a child has some characteristics that are not present in either parent but are present in some other relatives or were present in more remote ancestors. Even more commonly one sees that brothers and sisters, though showing a family resemblance in some traits, are clearly different in others; how could the same parents transmit different "bloods" to each of their children?

Mendel disproved the blood theory. He showed (1) that heredity is transmitted through factors (now called genes) that do not blend but segregate; (2) that parents transmit only one-half of the genes they have to each child, and they transmit different sets of genes to different children; and (3) that although brothers and sisters receive their heredities from the same parents, they do not receive the same heredities (an exception is identical births). If the eminence of some ancestor was really the reflection of his genes, it is quite likely that some of his descendants, especially the more remote ones, would not inherit these "good" genes at all. In sexually reproducing organisms, man included, every individual has a hereditary endowment, or genotype, uniquely his own.

BIOLOGICAL AND LEGAL INHERITANCE

In many languages the same words are used for both the inheritance of biological traits and the inheritance of property. Biological and legal inheritances are, however, very different processes. Inherited objects are actually transferred from one owner to another; inherited traits are not. In 1580 the French writer Montaigne wondered how he could have inherited gallstones from his father when his father did not have them until 25 years after the son was born, and the son himself did not have them until he was 45 years old. This seemed sheer mystery. Of course nobody inherits a gallstone (or even a gallbladder for that matter); what parents transfer to their offspring in the sex cells are the genes, and some genes, when they interact with some environments, allow the formation of gallstones.

The hereditary endowment received in the sex cells comes to be realized in the processes of the individual development through a sequence of stages from a single cell, to an embryo, an infant, a child, a youth, an adult, a senescent, a cadaver. In all of these stages heredity acts under a particular set of environmental conditions: what develops, then, depends both upon the heredity of the organism and upon the environment in which the organism develops. W. Johannsen in 1909 distinguished two important concepts—the phenotype and the genotype of the organism. The phenotype is what the organism actually is: its outward appearance, its bodily structures,

physiological processes, behaviour, etc. The phenotype changes continuously; the same individual shows different phenotypes in childhood, in adulthood, and in old age. The genotype, on the other hand, is the hereditary endowment, the sum total of the genes that the individual has received from its parents. It can be said that the genotype does not change during an individual's life. This statement must be carefully qualified; it does not mean that the genotype is isolated from its surroundings; far from that, the gene is among the most physiologically active constituents of cells and organisms. The genes, in fact, engender the synthesis of their own copies (see GENE). An adult has not only the genes that were present in the sex cells from which he came but also millions of true copies of these genes (unless some of them have changed by mutation; see below).

Distinction between phenotype and genotype

In man, skin coloration is certainly inherited, and yet exposure to the sun may make the skin darker and protection from the sun may make it paler; in insects, individuals that develop from starving larvae may be dwarfs compared to those coming from well-fed larvae; in some plants, leaves that are formed when the plant grows on land are quite different in shape from those formed under water. Is, then, the pale or the tanned skin, the dwarf or the giant size, the entire or the lacy leaf inherited? Not really, what is inherited is the norm of reaction of the developing organism to its environments. The genotype determines which phenotypes can arise in any given sequence of environments in which the individual who carries that genotype lives and develops. Except in albinos, the norm of reaction of most human genotypes permits a certain latitude of skin pigmentations to develop depending on exposure to or protection from sunlight; the norm of reaction of some insect species allows a great variety of body sizes to develop depending upon ample or scarce diets; the norm of reaction of some plant species allows a variety of leaf shapes to develop depending on local environment. Given a certain genotype, what actually develops is decided by the environments: given an environment, what develops is decided by the genotype.

The genotype–phenotype distinction is still not comprehended by many people, including some scientists. Lamarckism is a school of thought that assumes that characters acquired during an individual's life are inherited by his progeny, or, to put it in modern terms, that the modifications wrought by the environment in the phenotype are reflected in similar changes in the genotype. If this were so, the results of training or exercise of a person's abilities or of his musculature would make the training and exercise much easier or even dispensable in his offspring. Not only the pioneer French evolutionist Jean Lamarck but also other 19th-century biologists, including Charles Darwin, accepted the inheritance of acquired traits without hesitation. It was questioned by August Weismann, whose famous experiments in the late 1890s on amputation of tails in generations of mice showed that such somatic modification resulted neither in disappearance nor even in shortening of the tails in the descendants. Weismann concluded that the hereditary endowment of the organism, which he called the germ plasm, is wholly separate and is protected against the influences emanating from the rest of the body, the somatoplasm or soma. The germ plasm–somatoplasm are related to the genotype–phenotype concepts, but they are not identical and should not be confused with them.

Superstitions regarding heredity

The noninheritance of acquired traits does not mean that the genes cannot be changed by environmental influences: X-rays and other mutagens (see *Changing heredity*, below) certainly do change them, and the genotype of a population can be altered and molded by selection. It simply means that what is acquired by parents in their physique and their intellect is not inherited by their children. Related to these misconceptions are the beliefs in "prepotency"—that some individuals impress their heredities on their progenies more effectively than others—and in "prenatal influences" or "maternal impressions"—that the mental states and the sights experienced by a pregnant female are reflected in the constitution of the child to be born. How ancient these beliefs are is suggested in the Book of Genesis, in which Laban produced spotted or striped progeny in sheep by showing the pregnant ewes striped hazel rods. Another such belief is "telegony," which goes back to Aristotle; it alleged that the heredity of an individual is influenced not only by his father but also by males with whom the female may have mated and who have caused previous pregnancies. (An ancient English law holds a man who seduces the wet nurse of the heir to the throne guilty of polluting the "blood" of the royal family.) Even Darwin, as late as 1868, seriously discussed an alleged case of telegony: that of a mare that was mated to a zebra and subsequently to an Arabian stallion by whom the mare produced a foal with faint stripes on his legs. The simple explanation for these results is that such stripes occur naturally in some breeds of horses.

All these beliefs, from inheritance of acquired traits to telegony, must now be classed as superstitions. They do not stand up under experimental investigation and are incompatible with what is known about the mechanisms of heredity and about the remarkable and predictable properties of the genetic materials. Nevertheless, some people still cling to these beliefs. Some animal breeders take telegony seriously and do not regard as "pure bred" the individuals whose parents are admittedly "pure" but whose mothers had mated with males of other breeds. Weirdest of all were the Lysenkoists. The agronomist T.D. Lysenko (*q.v.*) was able for close to a quarter of a century, roughly between 1938 and 1963, to make his special brand of Lamarckism the official creed in the Soviet Union and to suppress most of the teaching and research in orthodox genetics. He and his partisans published hundreds of articles and books allegedly proving their contentions, which effectively abrogate the achievements of biology for at least a century. They were discredited in 1964.

II. Mendel's experiments and their significance

DISCOVERY AND REDISCOVERY OF MENDEL'S LAWS

Gregor Mendel (*q.v.*) published his work in the proceedings of the local society of naturalists in Brünn, Austria (now Brno, Czech.), in 1866, but none of his contemporaries appreciated its significance. It was not until 1900, 16 years after Mendel's death, that his work was rediscovered independently by H. de Vries in Holland, C.E. Correns in Germany, and E. Tschermak von Seysenegg in Austria. Like several investigators before him, Mendel experimented on hybrids of different varieties of a plant. Mendel investigated the common pea plant (*Pisum sativum*). His methods differed in two essential respects from those of his predecessors. First, instead of trying to describe the appearance of whole plants with all their characteristics, Mendel followed the inheritance of single, easily visible and distinguishable traits, such as round versus wrinkled seed, yellow versus green seed, purple versus white flowers, etc. Second, he made exact counts of the numbers of plants bearing this or that trait; it was from such quantitative data that he deduced the rules governing inheritance.

Experiments with pea plants

Since pea plants reproduce usually by self-pollination of their flowers, the varieties Mendel obtained from seedsmen were "pure"—*i.e.*, descended for several to many generations from plants with similar traits. Mendel crossed them by deliberately transferring the pollen of one variety to the pistils of another; the resulting first-generation hybrids, denoted by the symbol F_1, usually showed the traits of only one parent. For example, the

Table 1: Pea Plants with Dominant and Recessive Characters Obtained by Mendel in the Second Generation of Hybrids

number dominant		number recessive		ratio
Round seed	5,474	Wrinkled seed	1,350	2.96 : 1
Yellow seed	6,022	Green seed	2,001	3.01 : 1
Purple flowers	705	White flowers	224	3.15 : 1
Tall plants	787	Short plants	277	2.84 : 1

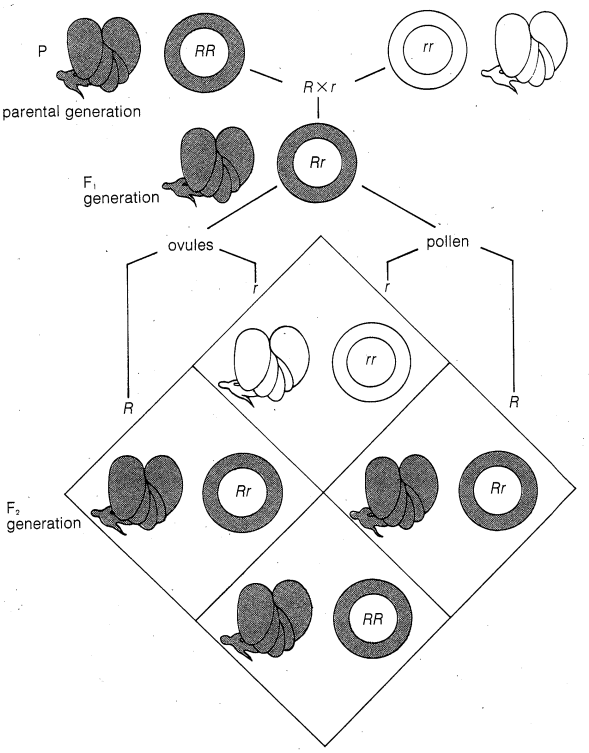


Figure 1: *Mendel's law of segregation.*
Cross of a purple-flowered and a white-flowered strain of peas. *R* stands for the gene for purple flowers and *r* for the gene for white flowers. Dark rings indicate the presence of a dominant gene for purple flowers.
From T. Dobzhansky, *Evolution, Genetics and Man*, (1955); John Wiley & Sons, Inc.

crossing of yellow-seeded plants with green-seeded ones gave yellow seeds; the crossing of purple-flowered plants with white-flowered ones gave purple-flowered plants, etc. Traits such as the yellow-seed colour and the purple-flower colour Mendel called dominant; the green-seed colour and the white-flower colour he called recessive. It looked as if the yellow and purple "bloods" overcame or consumed the green and white "bloods." That this was not so became evident when Mendel allowed the F_1 hybrid plants to self-pollinate and produce the second hybrid generation, F_2. Here both the dominant and the recessive traits reappeared, as pure and uncontaminated as they were in the original parents (generation P). Moreover, these traits appeared in constant proportions, namely about ¾ of the plants in the second generation showed the dominant trait and ¼ showed the recessive, a 3 to 1 ratio. It can be seen in Table 1 that Mendel's actual counts were as close to the ideal ratio as one could expect allowing for the sampling deviations present in all statistical data.

Mendel concluded that the sex cells, the gametes, of the purple-flowered plants carried some factor that caused the progeny to develop purple flowers, and the gametes of the white-flowered variety had a variant factor that induced the development of white flowers. W. Johannsen, the Danish geneticist, proposed in 1909 to call these factors genes.

An example of one of Mendel's experiments will illustrate how the genes are transmitted and in what particular ratios. Let *R* stand for the gene for purple flowers and *r* for the gene for white flowers (dominant genes are conventionally symbolized by capital letters and recessive genes by small letters). Since each pea plant contains a gene endowment half of whose set is derived from the mother and half from the father, each plant has two genes for flower colour. If the two genes are alike, for instance, both having come from white-flowered parents (*rr*), the plant is termed a homozygote (Figure 1). The union of gametes with different genes give a hybrid plant termed a heterozygote (*Rr*). Since the gene *R*, for purple, is dominant over *r*, for white, the F_1 generation hybrids will show purple flowers. They are phenotypically purple, but their genotype contains both *R* and *r* genes, and these alternative (allelic or allelomorphic) genes do not blend or contaminate each other. Mendel inferred that when a heterozygote forms its sex cells, the genes segregate and pass to different gametes. This is expressed in the first law of Mendel, the law of segregation of unit genes. Equal numbers of gametes, ovules, or pollen grains are formed that contain the genes *R* and *r*. Now, if the gametes unite at random, then the F_2 generation should contain about ¼ white-flowered and ¾ purple-flowered plants. The white-flowered plants, which must be recessive homozygotes, bear the genotype *rr*. About ⅓ of the plants exhibiting the dominant trait of purple flowers must be homozygotes, *RR*, and ⅔ heterozygotes, *Rr*. The prediction is tested by obtaining a third generation, F_3, from the purple-flowered plants; though phenotypically all purple-flowered, ⅔ of this group of plants reveal the presence of the recessive gene allele, *r*, in their genotype by producing about ¼ of white-flowered plants in the F_3 generation.

First and second laws of Mendel

Mendel also crossbred varieties of peas that differed in two or more easily distinguishable traits. When a variety with yellow round seed was crossed to a green wrinkled seed variety (fig. 2), the F_1 generation hybrids produced yellow round seed. Evidently yellow (*A*) and round (*B*) are dominant traits, green (*a*) and wrinkled (*b*) are recessive. In the F_2 generation, Mendel found 315 yellow round, 101 yellow wrinkled, 108 green round, and 32 green wrinkled seeds, a ratio approximately 9 : 3 : 3 : 1. The important point here is that the segregation of the colour (*A–a*) is independent of the segregation of the trait of seed surface (*B–b*). This is expected if four kinds of gametes are produced in equal numbers, carrying the four possible combinations of the parental genes: *AB*, *Ab*,

From T. Dobzhansky, *Evolution, Genetics and Man*, (1955); John Wiley & Sons, Inc.

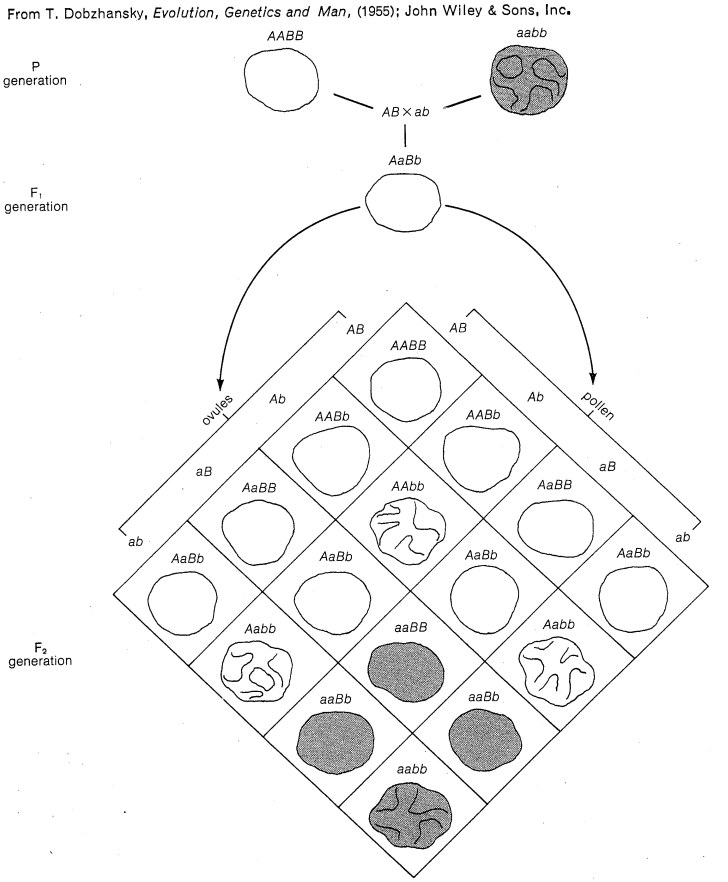


Figure 2: *Mendel's law of independent assortment.*
Cross of peas having yellow and smooth seeds with peas having green and wrinkled seeds. *A* stands for the gene for yellow and *a* for the gene for green; *B* stands for the gene for a smooth surface and *b* for the gene for a wrinkled surface.

aB, and *ab*. Random union of these gametes gives, then, the four phenotypes in a ratio 9 dominant–dominant : 3 recessive–dominant : 3 dominant–recessive : 1 recessive–recessive. Among these four phenotypic classes there must be nine different genotypes, a supposition that can be tested experimentally by raising a third hybrid generation. The predicted genotypes are actually found. Another test is by means of a backcross (or testcross)—the F_1 hybrid (phenotype yellow round seed; genotype *AaBb*) is crossed to a double recessive plant (phenotype green wrinkled seed; genotype *aabb*). If the hybrid gives four kinds of gametes in equal numbers and if all the gametes of the double recessive are alike (*ab*), the predicted progeny of the backcross are yellow round, yellow wrinkled, green round, and green wrinkled seed in a ratio 1 : 1 : 1 : 1. This prediction is realized in experiments. When the varieties crossed differ in three genes, the F_1 hybrid forms 2^3 or eight kinds of gametes (2^n = kinds of gametes, n being the number of genes). The second generation of hybrids, the F_2, has 27 (3^3) genotypically distinct kinds of individuals but only eight different phenotypes. From these results and others Mendel derived his second law, the law of recombination or independent assortment of genes.

UNIVERSALITY OF MENDEL'S LAWS

Although Mendel experimented with varieties of peas, his laws have been shown to apply to the inheritance of many kinds of characters in almost all organisms. In 1902 Mendelian inheritance was demonstrated in poultry and in mice. Albinism was the first trait in man shown, in 1903, to be a Mendelian recessive, with pigmented skin the corresponding dominant.

In 1902 and 1909 A.E. Garrod initiated the analysis of inborn errors of metabolism in man in terms of biochemical genetics. Alkaptonuria, inherited as a recessive, is characterized by excretion in the urine of large amounts of the substance called alkapton or homogentisic acid, which renders the urine black on exposure to air. In normal (*i.e.*, nonalkaptonuric) persons the homogentisic acid is changed to acetoacetic acid, the reaction being facilitated by a certain enzyme. Garrod advanced the hypothesis that this enzyme is absent or inactive in homozygous carriers of the defective recessive alkaptonuria gene; hence the homogentisic acid accumulates and is excreted in the urine. Mendelian inheritance of numerous traits in man has been studied since then (see GENETICS, HUMAN).

Forms of Mendelian inheritance

Mendelian inheritance takes a variety of forms. In the first place, clear-cut dominance and recessiveness are by no means always found. When red- and white-flowered varieties of four-o'clock plants or snapdragons are crossed, the first generation of hybrids have flowers of intermediate pink or rose colour. The F_2 generation shows a segregation in a ratio 1 red : 2 pink : 1 white. Suppose, then, that a gene allele R_1 is responsible for red and R_2 for white flowers; the homozygotes R_1R_1 and R_2R_2 are red and white respectively, and the heterozygotes R_1R_2 have pink flowers. A similar pattern of lack of dominance is found in shorthorn cattle. In diverse organisms, dominance ranges from complete (a heterozygote indistinguishable from one of the homozygotes) through incomplete (heterozygotes exactly intermediate) to excessive or over-dominance (a heterozygote more extreme than either homozygote).

An organism should not be thought of as an aggregate of independent traits, each determined by one gene. A "trait" is really an abstraction, a term of convenience in description. One gene may affect many traits (a condition termed pleiotropic), and many traits are affected by many genes. The gene white in *Drosophila* flies is pleiotropic; it affects the colour of the eyes and of the testicular envelope in the males, the fecundity and the shape of the spermatheca in the females, and the longevity of both sexes. In man many hereditary diseases are pleiotropic in origin, syndromes or complexes of diversified characters, inherited all together and determined by a single gene. On the other hand, the coat colour in many mammals is determined by numerous genes interacting to produce the result. Thus, in domestic cats there is a gene with three variants (alleles): one is dominant and allows for pigmentation, the other is recessive and in homozygous condition causes albinism, or complete lack of pigmentation; the third is intermediate, producing the so-called Siamese colour pattern. Moreover, the difference between yellow and black cats is due to a gene that in heterozygous females gives the tortoiseshell pattern; a dominant gene adds the tabby pattern; a recessive gene is responsible for the Maltese dilution ("smoky"); a dominant gene produces irregular white spotting. The great variety of colour patterns in cats, dogs, and other domesticated animals is the result of different combinations of complexly interacting genes. The gradual unravelling of their modes of inheritance was one of the active fields of research in the early years of genetics.

Two or more genes may produce similar and cumulative effects on the same trait. The skin colour difference between Negroes and so-called whites is due to several (probably four or more) interacting pairs of genes, each of which increases or decreases the skin pigmentation by a relatively small amount. The gene for albinism (lack of pigment) is not a part of this system; its dominant allele is necessary for the development of any skin pigment, and its recessive homozygous state results in the albino condition regardless of how many other pigment genes may be present. Albinism occurs in some individuals among people who belong to the dark- or intermediate-pigmented races, such as Negroes and American Indians, as well as among the Caucasians. The gene that determines whether albinism or pigmentation will express itself is called epistatic, and the genes whose expressions depend on the epistatic gene are called hypostatic.

APPARENT EXCEPTIONS TO MENDELIAN INHERITANCE

Polygenic inheritance. The greatest difficulties of analysis and interpretation are presented by the inheritance of many quantitative or continuously varying traits. The yield of grain in different varieties of corn (maize) and of milk in different breeds of cattle; egg-laying capacity in poultry; body size and proportions in breeds of sheep or dogs; and also stature, shape of the head, intelligence, and temperament in man range in continuous series from one extreme to the other and are in addition dependent on environmental conditions. Crosses of two varieties differing in such characters usually give F_1 hybrids intermediate between the parents, and the F_2 hybrids continue to be more or less intermediate on the average. At first sight, this situation suggests a blending inheritance through "blood" rather than Mendelian inheritance; in fact, it was probably observations of this kind of inheritance that suggested the folk idea of "blood theory." It has, however, been shown that these characters are polygenic—*i.e.*, determined by several or many genes, each taken separately producing only a slight effect on the phenotype, as small as or smaller than that caused by environmental fluctuations in the same characters. That Mendelian segregation does take place with polygenes, as with the genes having major effects (sometimes called oligogenes), is shown by the variation among F_2 and further generation hybrids being usually much greater than that in the F_1 generation. By selecting among the segregating progenies the desired variants, for example, individuals or families with greatest yield, best size, or a desirable behaviour, one can produce new breeds or varieties sometimes exceeding the parental forms. Hybridization and selection are consequently potent methods used for improvement of agricultural plants and animals (see ANIMAL BREEDING; PLANT BREEDING).

Mendelian segregation in polygenes

Higher animals and plants are mostly diploid; *i.e.*, they have two sets of genes (and chromosomes); one set received from the female parent and the other from the male parent. Some simpler forms of life, including most micro-organisms, are haploid, having a single set of genes and chromosomes during most of their life cycle. The sexual process in such organisms involves fertilization and emergence of a diploid state, which reverts more or

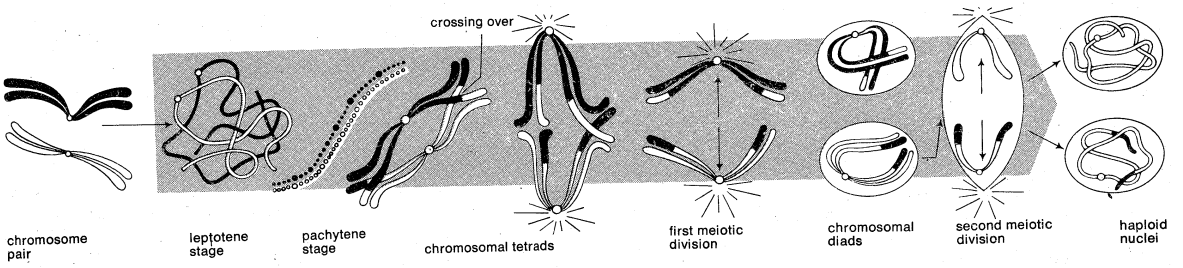


Figure 3: Behaviour of chromosomes at meiosis (see text).
From T. Dobzhansky, *Evolution, Genetics and Man*, (1955); John Wiley & Sons, Inc.

less rapidly to the haploid by means of meiosis, a special form of cell division. Haploid organisms display a much simpler Mendelian inheritance than that observed by Mendel in his pea crosses. When two strains of a haploid unicellular alga or of a bread mold, which differ in n genes, are crossed, sexual union results in a diploid heterozygote, which goes through meiosis; Mendelian segregation and recombination give rise to 2^n kinds of cells or spores with different combinations of the parental genes. Gene inheritance can be observed even in bacteria, where at least some strains undergo a sexual process and Mendelian segregation.

Cytoplasmic inheritance. It appears that all inheritance transmitted through the chromosomes of the nuclei of the sex cells is basically Mendelian inheritance. There are, however, cases of non-Mendelian, or cytoplasmic, inheritance. The classical example is the four-o'clock plant (*Mirabilis jalapa*) studied by C.E. Correns in 1909. Some garden varieties of this normally green plant have leaves streaked with white. Such variegated plants have some branches with only white, others with only green, and still others with mixtures of white, green, and variegated leaves. The seeds coming from the flowers borne on white branches give rise only to white-leaved seedlings, while those from variegated branches give white-, green-, and variegated-leaved plants. The colour of the progeny is clearly independent of the source of the pollen used to pollinate the flowers. The explanation is that the chlorophyll, the green pigment of the leaves, is carried in the cytoplasm of plant cells in special structures called chloroplasts and that the primordia from which the chloroplasts develop are transmitted usually through the female line (ovules) and not through the male line (pollen).

Inheritance in the four-o'clock plant

The cytoplasmic inheritance of the so-called killer trait in certain strains of the protozoan *Paramecium aurelia* is a curious example. These strains produce a substance that poisons and kills individuals of normal "sensitive" strains of the same species. The killer trait depends upon particles called "Kappa" in the cytoplasm of the killer strains. Kappa particles, the size of small bacteria or large viruses, reproduce themselves by division and are transmitted only through the cytoplasm. Yet their maintenance requires also the presence of a Mendelian (nuclear) gene called *K;* experiments can be contrived to obtain individuals with Kappa cytoplasm but without the dominant gene *K;* in the progeny of such individuals the Kappa particles gradually disappear and the killer character is lost.

III. Physical basis of heredity

CHROMOSOMES AND GENES

Chromosomes as carriers of heredity. The human body starts with the union of two sex cells. The body of a newborn infant consists of some 200,000,000,000 (2×10^{11}) cells. Cells arise only through division of other cells; when cells divide, their nuclei also divide by means of a remarkable manoeuvre called mitosis (see CELL AND CELL DIVISION). As early as 1848 cell nuclei were observed to resolve themselves during mitosis into smaller structures; later these structures were found to absorb certain dyes and so came to be called chromosomes (coloured bodies). The number of the chromosomes in a nucleus is usually constant in all individuals of a species but often different in different species—for example, 46 in man, 40 in the house mouse, 8 in the vinegar fly, or fruit fly (*Drosophila*), 20 in corn (maize), 24 in the tomato, 48 in the potato. During a mitotic division each chromosome duplicates itself and splits lengthwise, each duplicant passing to one of the two daughter cells formed during division. The mechanism of mitosis insures that all body cells have true copies of all chromosomes that were present in the fertilized egg from which the body developed.

Fertilization involves fusion of the nuclei of two sex cells or gametes. These nuclei contain haploid sets of chromosomes—23 in man, 4 in *Drosophila*, 10 in corn, etc. A fertilized egg cell, or zygote, contains, then, the diploid chromosome number, given above, $23 + 23 = 46$ in man, $4 + 4 = 8$ in *Drosophila*, etc. Sooner or later, however, the diploid zygotes must produce another crop of haploid gametes. This occurs by means of a process called meiosis; its essential features are shown in Figure 3. For the sake of simplicity, the diploid cell is shown to contain a single pair of chromosomes, one member of which is represented black (from the father) and the other white (from the mother). At the leptotene stage the chromosomes appear as long thin threads. At pachytene they pair, the corresponding portions of the two chromosomes lying side by side. The chromosomes then duplicate and contract; at this stage the chromosomes are usually held together by chiasmata, which are believed to be caused by crossing-over (exchanges of sections of the chromosomes). The first meiotic division separates the chromosomal tetrads into diads. The second meiotic division then gives haploid nuclei.

Behaviour of chromosomes during division

Female gametes, egg cells, are often much larger and quite different in structure from male gametes. Yet, as pointed out above, the mother and the father transmit heredity equally. Female and male gametes are, however, similar in one respect; they contain similar sets of chromosomes. This fact and the orderly precision of the processes of mitosis and meiosis led biologists to infer that the chromosomes must be the carriers of the hereditary materials.

Chromosomes as carriers of genes. When Mendel's work was rediscovered, the question of the relationship between the genes and chromosomes inevitably arose. It was felt that the chromosomes must be the carriers of the genes. Indeed, suppose that the white chromosomes shown in Figure 3 carries the gene for albinism and the black chromosome carries the gene for dark pigmentation. It is evident that the two gene alleles will undergo segregation at meiosis and that one-half of the gametes formed will contain the albino gene and the other half the pigmentation gene. Following the scheme in Figure 1, random combination of the gametes with the albino and the pigmentation gene will give two kinds of homozygotes and one kind of heterozygote in a ratio 1 : 1 : 2. Mendel's law of segregation is, thus, the outcome of chromosome behaviour at meiosis. The same is true of the second law, that of the independent assortment.

Consider the inheritance of two pairs of genes, such as Mendel's factors for seed coloration and seed surface in peas; these genes are located on different pairs of chromosomes. Since maternal and paternal members of different chromosome pairs are assorted independently, so are the genes they contain. This explains, in part, the genetic variety seen among the progeny of the same pair of parents. As stated above, man has 46 chromosomes in his body cells and in the cells (oogonia and spermatogonia) from which the sex cells arise. At meiosis these 46

chromosomes form 23 pairs, one of the chromosomes of each pair being of maternal and the other of paternal origin. Independent assortment is, then, capable of producing 2^{23}, or 8,388,608, kinds of sex cells with different combinations of the grandmaternal and grandpaternal chromosomes. Since each parent has the potentiality of producing 2^{23} kinds of sex cells, the total number of possible combinations of the grand-parental chromosomes is $2^{23} \times 2^{23} = 2^{46}$. The population of the world is now between 3,000,000,000 and 4,000,000,000 persons, or between 2^{31} and 2^{32} persons. It is, therefore, certain that only a tiny fraction of the potentially possible chromosome and gene combinations can ever be realized. Yet even 2^{46} is an underestimate of the variety potentially possible. The grandmaternal and grandpaternal members of the chromosome pairs are not indivisible units. Each chromosome carries many genes, and the chromosome pairs exchange segments at meiosis. This is evidence that the genes rather than the chromosomes are the units of Mendelian segregation.

LINKAGE OF TRAITS

Simple linkage. As pointed out above, the random assortment of the maternal and paternal chromosomes at meiosis is the physical basis of the independent assortment of genes and of the traits they control. This is the basis of the second law of Mendel. The number of the genes in a sex cell is, however, much greater than that of the chromosomes. When two or more genes are borne on the same chromosome these genes may not be assorted independently (some purists consequently prefer to speak of a second "rule" rather than "law" of Mendel). When the parental combinations of traits are more frequent than new combinations, the genes are said to be linked.

Linked genes in Drosophila

When a *Drosophila* fly homozygous (pure breeding) for a normal grey body and long wings is crossed with one having a black body and vestigial wings the F_1 consists of hybrid gray, long-winged flies (Figure 4). Gray body (*B*) is evidently dominant over black body (*b*), and long wing (*V*) is dominant over vestigial wing (*v*). Now consider a backcross of the heterozygous F_1 males to double recessive black-vestigial females (*bbvv*). Independent assortment would be expected to give in the progeny of the backcross the following: 1 gray-long : 1 gray-vestigial : 1 black-long : 1 black-vestigial. In reality only gray-long and black-vestigial flies are produced, in approximately equal numbers; the genes remain linked in the same combinations in which they were found in the parents. The backcross of the heterozygous F_1 females to double recessive males gives a somewhat different result: 42 percent each of gray-long and black-vestigial flies and about 8 percent each of black-long and grey-vestigial classes. In sum, 84 percent of the progeny have the parental combinations of traits, and 16 percent have the traits recombined. The interpretation of these results given in 1911 by the U.S. geneticist T.H. Morgan laid the foundations of the theory of linear arrangement of genes in the chromosomes.

Traits that exhibit linkage in experimental crosses (such as black body and vestigial wings) are determined by genes located in the same chromosome. As more and more genes became known in *Drosophila melanogaster*, they fell neatly in four linkage groups corresponding to the four pairs of the chromosomes this species possesses. One linkage group consists of sex-linked genes, located in the X chromosome (see below); of the three remaining linkage groups, two have many more genes than the remaining one; this corresponds to the presence of two pairs of large chromosomes and one pair of tiny dotlike chromosomes. The numbers of linkage groups in other organisms are equal to or smaller than the numbers of the chromosomes in the sex cells; *e.g.*, 10 linkage groups and 10 chromosomes in corn (maize), 19 linkage groups and 20 chromosomes in the house mouse, at least 4 linkage groups and 23 chromosomes in man.

We have seen above that the linkage of the genes black and vestigial in *Drosophila* is complete in heterozygous males, while in the progeny of females there appear

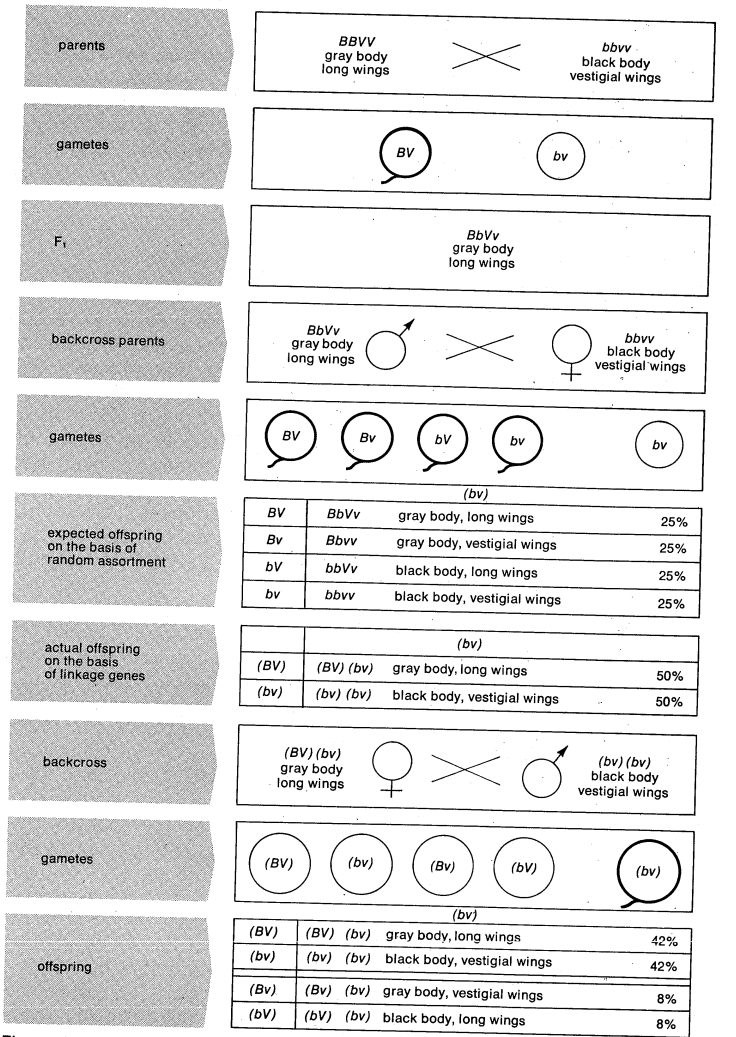


Figure 4: Linkage of genes as illustrated by body colour and wing length in *Drosophila* flies (see text).

Reprinted by permission of the publisher, from L.H. Snyder and P.R. David, *The Principles of Heredity* (Lexington, Mass.: D.C. Heath and Company, 1957)

about 17 percent of recombination classes. With very rare exceptions, the linkage of all genes belonging to the same linkage group is complete in *Drosophila* males, while in the females different pairs of genes exhibit all degrees of linkage from complete (no recombination) to 50 percent (random assortment). Morgan's inference was that the degree of linkage depends on physical distance between the genes in the chromosome: the closer the genes the tighter the linkage, and vice versa. Furthermore, Morgan perceived that the chiasmata, crosses that occur in meiotic chromosomes, indicate the mechanism underlying the phenomenon of linkage. As shown schematically in Figure 3, the maternal and paternal chromosomes (represented black and white) cross over and exchange segments, so that a chromosome emerging from the process of meiosis may consist of some maternal (grandmaternal) and some paternal (grandpaternal) sections. If the probability of crossing-over taking place is uniform along the length of a chromosome (which was later shown to be not quite true), then genes close together will be recombined less frequently than those far apart. This realization opened a hitherto scarcely imagined opportunity to prepare "maps" showing the arrangement of the genes and the estimated distances between them in the chromosome by studying the frequencies of recombination of various traits in the progenies of hybrids. In other words, the linkage maps of the chromosomes are really summaries of many statistical observations on the outcomes of hybridization experiments. In principle at least, such maps could be prepared even if the chromosomes, not to speak of the chiasmata at meiosis, were unknown. But here is

Chromosome maps

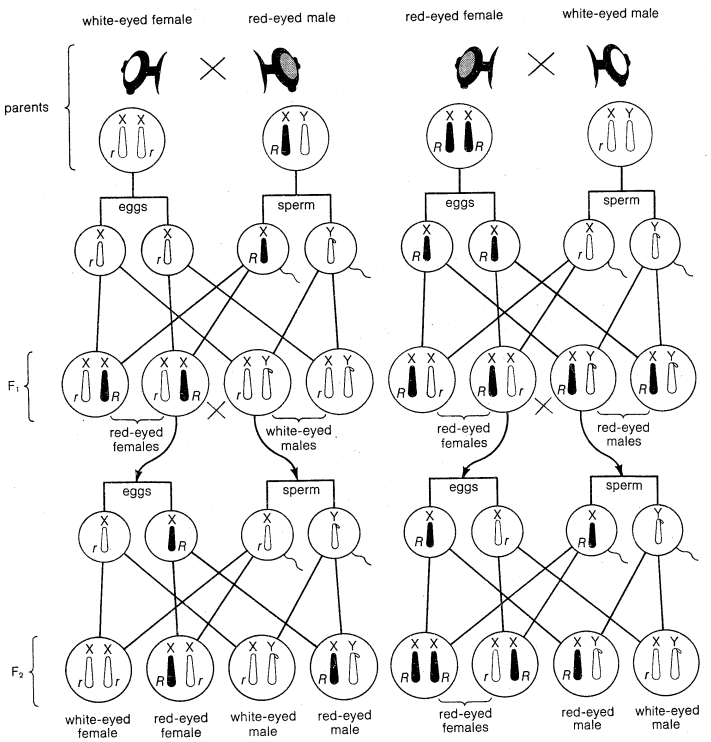


Figure 5: Sex-linked inheritance of white eyes in *Drosophila* flies (see text).

From *Life: An Introduction to Biology* by Simpson, Pittendrigh and Tiffany, Copyright © 1957 by Harcourt Brace Jovanovich, Inc., and reproduced with their permission

an interesting and relevant fact: in *Drosophila* males the linkage of the genes in the same chromosome is complete, and observations under the microscope show that no chiasmata are formed in the chromosomes at meiosis. In most organisms, including man, chiasmata are seen in the meiotic chromosomes in both sexes, and observations on hybrid progenies show that recombination of linked genes occurs also in both sexes.

The most detailed chromosome maps have been constructed for Morgan's classical material—*Drosophila melanogaster*. Less detailed chromosome maps exist for some other species of *Drosophila* flies, for corn, the house mouse, the bread mold *Neurospora crassa*, and for some bacteria and bacteriophages. The mapping of human chromosomes is a most arduous task, because no experimental crosses can be arranged and because the investigator has to rely on observation in families in which linked traits happen to be segregating.

Sex linkage. The male of many animals has one chromosome pair, the sex chromosomes, consisting of unequal members called X and Y. At meiosis the X and Y chromosomes first pair, then disjoin and pass to different cells. One-half of the gametes (spermatozoa) formed contain the X and the other the Y chromosome. The female has two X chromosomes, and all egg cells normally carry a single X. The eggs fertilized by X-bearing spermatozoa give females (XX), and those fertilized by Y-bearing spermatozoa give males (XY).

Sex-linked genes

The genes located in the X chromosomes exhibit what is known as sex-linkage or crisscross inheritance. The classic case, described by T.H. Morgan in 1910, is that of the white eyes in *Drosophila* (Figure 5). White-eyed females crossed to males with the normal red eye colour produce red-eyed daughters and white-eyed sons in the F_1 generation, and equal numbers of white-eyed and red-eyed females and males in the F_2 generation. The cross of red-eyed females to white-eyed males gives a different result: both sexes are red eyed in F_1, the females in the F_2 generation are red eyed, half of the males are red eyed and the other half white eyed. As interpreted by Morgan, the gene that determines the red or white eyes is borne on the X chromosome, and the allele for red eye is dominant over that for white eye. Since a male receives its single X chromosome from his mother, all sons of white-eyed females also have white eyes. A female inherits one X chromosome from her mother and the other X from her father. Red-eyed females may have genes for red eyes in both of their X chromosomes (homozygotes) or may have one X with the gene for red and the other for white (heterozygotes). In the progeny of heterozygous females one-half of the sons will receive the X chromosome with the gene for white and will have white eyes, and the other half will receive the X with the gene for red eyes. The daughters of the heterozygous females crossed with white-eyed males will have either two X chromosomes with the gene for white and hence white eyes or will have one X with white and the other X with the gene for red eyes and will be red-eyed heterozygotes.

In man, the red-green colour blindness and hemophilia are among traits showing sex-linked inheritance and consequently due to genes borne in the X chromosome.

In some animals—birds, butterflies and moths, some fish, and at least some amphibians and reptiles—the chromosomal mechanism of sex determination is, as it were, a mirror image of that described above. The male has two X chromosomes and the female an X and Y chromosome. Here the spermatozoa are all alike in having an X chromosome; the eggs are of two kinds, some with X and others with Y chromosomes, usually in equal numbers. The sex of the offspring is, then, determined by the egg rather than by the spermatozoon. Sex-linked inheritance is altered correspondingly. A male homozygous for a sex-linked recessive trait, crossed to a female with the dominant one, gives in the F_1 generation daughters with the recessive trait and heterozygous sons with the corresponding dominant trait. The F_2 generation has recessive and dominant females and males in equal numbers. A male with a dominant trait crossed to a female with a recessive trait gives uniformly dominant F_1, and a segregation in a ratio of 2 dominant males : 1 dominant female : 1 recessive female.

Observations on pedigrees or experimental crosses show that certain traits exhibit sex-linked inheritance; the behaviour of the X chromosomes at meiosis is such that the genes they carry may be expected to exhibit sex-linkage. This evidence still failed to convince some skeptics that the genes for the sex-linked traits were in fact borne in certain chromosomes seen under the microscope. An elegant experimental proof was furnished in 1916 by the U.S. geneticist C.B. Bridges. As stated above, white-eyed *Drosophila* females crossed to red-eyed males usually produce red-eyed female and white-eyed male progeny. Among thousands of such "regular" offspring there are occasionally found exceptional white-eyed females and red-eyed males. Bridges constructed the following working hypothesis. Suppose that during meiosis in the female, gametogenesis occasionally goes wrong, and the two X chromosomes fail to disjoin (Figure 6). Exceptional eggs

From *Life: An Introduction to Biology* by Simpson, Pittendrigh and Tiffany, Copyright © 1957 by Harcourt Brace Jovanovich, Inc., and reproduced with their permission

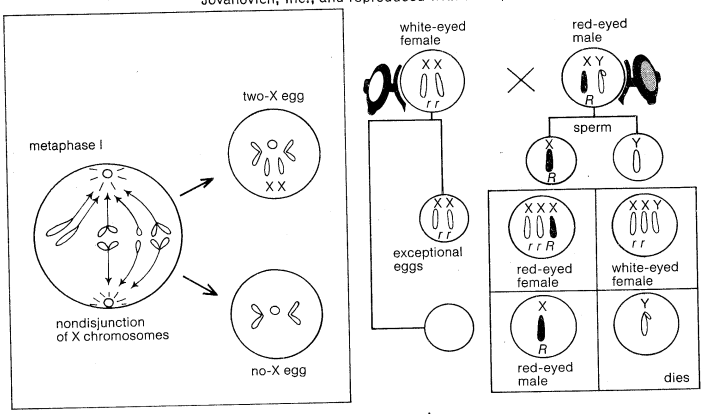


Figure 6: Nondisjunction of the X chromosomes in Drosophila flies, an explanation for the appearance of unexpected types of offspring. Abnormal meiosis in the female (left) is responsible for the exceptional eggs (right).

Verification of sex-linked inheritance

will then be produced carrying two X chromosomes and eggs carrying none. An egg with two X chromosomes coming from a white-eyed female fertilized by a spermatozoon with a Y chromosome will give an exceptional white-eyed female. An egg with no X chromosome fertilized by a spermatozoon with an X chromosome derived from a red-eyed father will yield an exceptional red-eyed male. This hypothesis can be rigorously tested. The exceptional white-eyed females should have not only the two X chromosomes but also a Y chromosome, which normal females do not have. The exceptional males should, on the other hand, lack a Y chromosome, which normal males do have. Both predictions were verified by examination under a microscope of the chromosomes of exceptional females and males. The hypothesis also predicts that exceptional eggs with two X chromosomes fertilized by X-bearing spermatozoa must give individuals with three X chromosomes; such individuals were later identified by Bridges as poorly viable "superfemales." Exceptional eggs with no Xs, fertilized by Y-bearing spermatozoa, will gives zygotes without X chromosomes; such zygotes die in early stages of development.

HEREDITY AND NUCLEIC ACIDS

In 1869 a substance containing nitrogen and phosphorus, was extracted from cell nuclei. It was originally called nuclein, but is now known as deoxyribonucleic acid (DNA). DNA is the chemical component of the chromosomes that is chiefly responsible for their staining properties in microscopic preparations. In addition to DNA, chromosomes contain a variety of proteins, mostly histones but sometimes protamines. The question naturally arises whether the nucleic acids or the proteins, or both together, are the carriers of the genetic information, which makes the genes of the same organism and of different organisms specifically different. Until the early 1950s most biologists were inclined to believe that the proteins were the chief carriers of heredity; at present the weight of the evidence is overwhelmingly in favour of ascribing this role to the nucleic acids. Nucleic acids are not only chemically simpler substances than the proteins but they do not differ greatly in composition in quite different organisms. By contrast, proteins are large, some of them gigantic molecules, and the possible variety of chemically different proteins is virtually unlimited. The diversity of the genes, therefore, seemed likely to rest on the diversity of the proteins.

The weight of evidence for DNA and against protein as the genetic material directed the attention of biologists to the nucleic acids. For details of the determining experiments see GENE.

HEREDITY AND DEVELOPMENT

It has been shown that the sex cells contain in the chromosomes of their nuclei material particles called genes. Variations in the quality or the dosage of the genes produce visible or phenotypic changes in the organism. When parents carry different genes, their hybrid offspring show a distribution of the phenotype traits as described by Mendel's laws. Between the genes in the sex cells and the traits, such as seed colours and shapes in peas or eye and skin colours in man, there intervene, however, very complex processes of the individual's development. The genes determine an organism's potentialities, some of which are gradually realized in a body growing and developing in a certain sequence of environments. The question naturally arises: how does the realization of the genetic potentialities take place? A fertilized egg cell carries a certain complement of genes; by repeated divisions these cells give rise to numerous cells of different kinds (*e.g.*, epidermal, glandular, muscle, skeletal, nerve cells, etc.). What makes these cells different if they carry the same genes? Developmental genetics (also called phenogenetics) attempts to find answers to this question.

Problems of developmental genetics may be approached, generally, in two ways. First, one may attempt to retrace the pathways leading from a gene to a phenotypic trait, starting from the latter and going backward toward the beginning of the development. Second, the action of the genes in the nuclei and the cells that carry them may be studied. Both approaches have in fact been utilized. The first approach usually discloses that the visible expressions of gene differences are brought about by divergence of metabolic processes controlled by the genes. Garden varieties of sweet peas, primroses, dahlias, and verbenas present a great diversity of flower colours. Many investigators made crosses between colour varieties of several species and analyzed the results in terms of the Mendelian genes responsible for the often striking differences observed. Biochemical studies correlated with the genetic ones can be illustrated in a group of water-soluble pigments: anthoxanthins (yellow colours) and anthocyanins (blues, pinks, purples, reds, etc.). These pigments are all chemically related, having in their molecules three rings derived from the condensation of sugars. The differences are due to substitution of hydrogen atoms by hydroxyl (OH) groups at specified positions in the molecules, attachment of extra sugars, and changes in the acidity (pH) of the cell sap. Each of these substitutions and changes is due to a particular gene that modifies or blocks a certain chemical reaction involving usually the colourless precursor of the pigments.

Approaches to developmental genetics

Using the bread mold *Neurospora crassa* as their experimental object, G.W. Beadle and E.L. Tatum initiated in the late 1930s their classical studies on gene control of individual steps in the chains of biochemical reactions. This body of work earned for them a Nobel Prize in 1958. Most strains of *N. crassa* can grow on a "minimal" medium, having a suitable carbohydrate (sugar), certain inorganic salts, and the vitamin biotin. From these sources *N. crassa* can synthesize all the amino acids, vitamins, and other substances necessary for all its vital activities. Now suppose that a gene controlling one of the synthetic reactions is so altered by a mutation that it no longer yields the enzyme necessary for that reaction. The strain with the altered gene cannot, therefore, grow on the minimal medium unless the necessary substance is added to the medium. Let a metabolic pathway consist of a chain of reactions $A \rightarrow B \rightarrow C$ (for example, ornithine $\rightarrow$ citrulline $\rightarrow$ arginine). Each step is mediated by a different enzyme controlled usually by a single gene. A change in the gene that mediates, for example, the step $A \rightarrow B$ may eliminate this reaction. The products B and C then fail to be formed; the compound A, which is normally transformed into B and C, accumulates in abnormally high quantities. However, since the enzyme that normally facilitates the reaction $B \rightarrow C$ is available, the organism is able to form D if C is added to the "minimal" medium.

Gene control in bread mold

The control of the successive chemical steps in the metabolic pathways by a single gene for each step appears to be a general phenomenon in the living world. This is what is involved in many inborn errors of metabolism in man. Much important work published in recent years is concerned with analysis of metabolic pathways in a variety of organisms. One of the striking results of this work has been the demonstration that these pathways are frequently similar in quite diverse organisms, which suggests the presence in them of similar genes or at least of genes with similar functions.

The properties that distinguish one gene from another are specified by the sequence of the "letters" of the DNA "alphabet"—A (adenine), C (cytosine), G (guanine), and T (thymine). The protein "alphabet" has 20 "letters," the 20 amino acids, and the properties of the proteins are specified by the sequences of these letters. The "messages" inscribed in the genes by means of the 4-letter alphabet are translated into the 20-letter alphabet of the proteins through the mediation of ribonucleic acids (RNA). How the "letters" in RNA specify the sequence in which amino acids will follow each other is discussed in the article GENE.

IV. Heredity and environment

PREFORMISM AND EPIGENESIS

A notion that was widespread among pioneer biologists in the 18th century was that the fetus, and hence the adult

The homunculus of early microscopists

organism that develops from it, is preformed in the sex cells. Some early microscopists even imagined that they saw a tiny "homunculus," a diminutive human figure, encased in the human spermatozoon. The development of the individual from the sex cells appeared deceptively simple—it was merely increase in size and growth of what was already present in the sex cells. The antithesis of the early preformation theories were theories of epigenesis, which claimed that the sex cells were structureless jelly and contained nothing at all in the way of rudiments of future organisms. The naïve early versions of preformation and epigenesis had to be given up when embryologists showed that the embryo develops by a series of complex but orderly and gradual transformations (see DEVELOPMENT, ANIMAL). Darwin's "Provisional Hypothesis of Pangenesis" was distinctly preformistic; A. Weismann's theory of determinants in the germ plasm as well as the early ideas about the relations between genes and traits also tended toward preformism.

Heredity was defined at the beginning of this article as a process that results in the progeny resembling their parents. This definition is valid as far as it goes, but what about characteristics subject to environmental modification? It would be convenient if some characters were hereditary and others environmental, but in reality such a dichotomy does not exist. As mentioned above, the genotype determines not the characters but the norm of reaction of the developing organism to its environments. The development results in a sequence of phenotypes; "characters" or "traits" are really abstractions used in talking or writing about the phenotype, the appearance the organism presents to us. The "norm" of reaction is perhaps not a felicitous phrase, although it seems to have established itself as a term in genetics; the "norm" does not imply that certain manifestations of the genotype are normal and others abnormal but subsumes the entire repertory of the phenotypes that a carrier of a given genotype can develop in all environments in which it can live.

A trait that is hereditary (*e.g.*, skin colour in man) may be modified by environmental influences (suntanning). And conversely, a trait sensitive to environmental modifications (*e.g.*, weight in man) may well be genetically conditioned. In a sense then, all traits have both a genetic and an environmental component. Organic development is preformistic insofar as a fertilized egg cell contains a genotype that conditions the events that may occur and epigenetic insofar as a given genotype allows a variety of possible outcomes. These considerations should dispel the reluctance felt by many people to accept the fact that mental as well as physiological and physical traits in man are genetically conditioned. Genetic conditioning does not mean that heredity is the "dice of destiny." At least in principle, but not invariably in practice, the development of a trait may be manipulated by environmental changes.

Hygiene, medicine, education, even social and political systems may from this point of view be regarded as environmental engineering devices by means of which organic development may be controlled. Whether man will be the slave or the master of his inherited genes will depend on the degree of understanding of human nature that may be achieved.

HEREDITY IN HEALTH AND DISEASE

Extent of genetic diseases

That the difference between health and disease is sometimes due to heredity has been known for a long time. The authors of the Talmud were aware of the fact that hemophilia was transmitted from mothers to some of their sons; the male progeny of mothers who lost two boy infants because of excessive bleeding were exempt from the rite of circumcision. How great is the sum total of ill health due to defective heredity is nevertheless still very imperfectly known. In the 1950s A.C. Stevenson estimated for the population of Northern Ireland that about 26 percent of the hospital beds were occupied by genetically handicapped patients; some 6 percent of the consultations with medical practitioners and 8 percent of those with specialists involved such persons. At least 4 percent of the infants born had more or less seriously incapacitating genetic defects. This last figure was an underestimate because it did not include some quite serious conditions, such as schizophrenia, in which genetic factors play an important role, nor did it include approximately 14 percent of the recordable pregnancies that resulted in abortions or stillbirths, some of which are also genetically conditioned. There is no reason to believe that the population of Northern Ireland during those years was genetically handicapped either especially severely or exceptionally lightly. It is representative of what may prevail widely.

Diseases due to simple recessive genes are generally more difficult to study genetically because the affected children are usually born to parents, neither of whom is affected. For example, juvenile amaurotic idiocy, a grave neurological disorder that causes death of the afflicted children long before sexual maturity, cannot for that reason be present in a parent. Both parents may, however, be heterozygous and carry one dose of the defective gene; although they themselves enjoy good health, one out of four of their children on the average may develop the disease.

The penetrance level of genes

In some cases the dichotomy of hereditary versus environmental defect or disease cannot always be sustained. The genes for brachydactyly, albinism, juvenile amaurotic idiocy, hemophilia, and other conditions that are classic examples of Mendelian inheritance in man have what is called complete penetrance; that is, individuals carrying them exhibit certain traits regardless, so far as is known, of the environments in which they occur. On the other hand, at least some forms of sugar diabetes (diabetes mellitus) are genetic, but the disease may appear at different ages; and when it does appear, its expression may be more or less severe. Such variable penetrance and expressivity depend presumably on the environment, both external and genetic (*i.e.*, on the other genes the individual carries).

There has been considerable dispute about the hereditary basis of schizophrenia. There are apparently several genes the defective alleles of which may combine to cause, under suitable environmental conditions, a clinically diagnosable schizophrenia. Still stronger environmental components are found in diseases in which a "susceptibility" is inherited. Tuberculosis is caused by infection from a pathogenic bacterium; it is, therefore, basically an "environmental" disease. There is, nevertheless, good evidence that the carriers of certain genotypes contract the disease more easily than do the carriers of other genotypes. An even clearer relationship is that between falciparum malaria and sickle-cell anemia. One gene causes, in the homozygous condition, sickle-cell anemia, a disease usually fatal before adolescence. The same gene in the heterozygous condition gives at most only a mild anemia and in addition confers a relative immunity to falciparum malaria. Malaria is caused, of course, by a blood parasite; however, in populations in which practically everybody is infected, the genotype of a person may be the decisive variable. Evidence is rapidly accumulating that the susceptibility to widespread degenerative diseases, such as hypertension and coronary artery disease, is also genetically conditioned.

To "cure" schizophrenia or any other hereditary disease obviously does not mean removal of the responsible gene or genes. Health and disease, however, are conditions of the phenotype. The curative treatment places the carriers of the undesirable genes in environments in which the disease does not assert itself in the phenotype, and the persons concerned are enabled to lead reasonably normal and happy lives. The symptoms of the disease, not the cause of the disease, are treated, but even this is a great accomplishment.

HERITABILITY

Although hereditary diseases and malformations are, unfortunately, by no means uncommon in the aggregate, no one of them occurs very frequently. The characteristics by which one person is distinguished from another, such

as facial features, stature, shape of the head, skin, eye and hair colours, and voice, are not usually inherited in a clear-cut Mendelian manner, as are some hereditary malformations and diseases. This is not as strange as it may seem. Gene changes, or mutations, that produce morphological or physiological effects drastic enough to be clearly set apart from the more usual phenotypes are likely to cause diseases or malformations just because they are so drastic. The variations that occur among healthy persons are, as a general rule, caused by polygenes with individually small effects. The same is true of individual differences among members of various animal and plant species. Even brown-blue eye colour in man, which in many families behaves as if caused by two forms of a single gene (brown being dominant and blue recessive), is often blurred by minor gene modifiers of the pigmentation. Some apparently blue-eyed persons actually carry the gene for the brown eye colour, but several additional modifier genes decrease the amount of brown pigment in the iris.

The nature-nurture problem in man

The question geneticists must often attempt to answer is how much of the observed diversity between persons, or between individuals of any species, is due to hereditary, or genotypic variations, and how much of it is due to environmental influences. Applied to man, this is sometimes referred to as the nature-nurture problem. With animals or plants the problem is evidently more easily soluble than it is with man. Two complementary approaches are possible. First, individuals, or their progenies, are raised in environments as uniform as can be provided, with food, temperature, light, humidity, etc., carefully controlled. The differences that persist between such individuals or progenies probably reflect genotypic differences. Second, individuals with similar or identical genotypes are placed in different environments. The phenotypic differences then may be ascribed to environmental induction. Experiments combining both approaches have been carried out on several species of plants that grow naturally at different altitudes, from sea level to the alpine zone of the Sierra Nevada of California. Young yarrow plants (*Achillea*) were cut in three parts and the cuttings replanted in experimental gardens at sea level, at midaltitude (4,800 feet [1,460 metres]), and at a high altitude (10,000 feet [3,050 metres]) (Figure 7). The

From T. Dobzhansky, *Heredity and the Nature of Man*, (1964); Harcourt Brace Jovanovich, Inc.

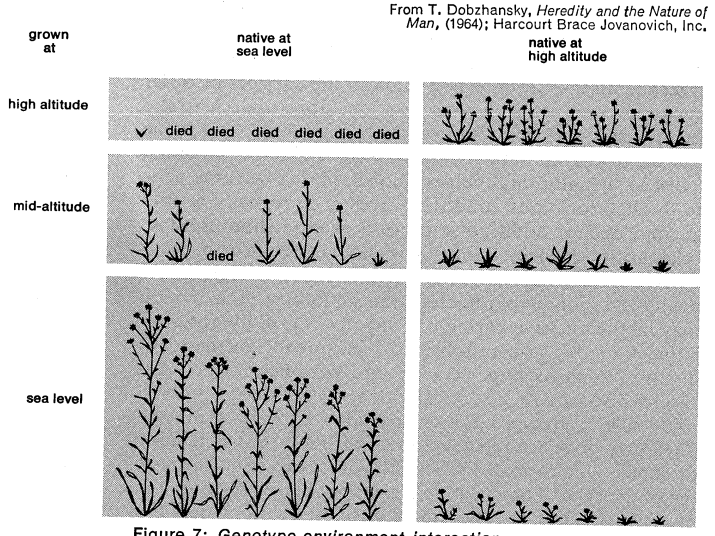


Figure 7: *Genotype-environment interaction.* Yarrow plants native in different habitats divided and replanted at different elevations.

plants native at sea level grow best in their native habitat, less well at midaltitudes, and die at high altitudes. On the other hand, the alpine race survives and develops better at the high-altitude transplant station than at lower altitudes.

With organisms that cannot survive being cut in pieces and placed in controlled environments, a partitioning of the observed variability into genetic and environmental components may be attempted by other methods. Suppose that in a certain population individuals vary in stature, weight, or some other trait. These characters can be measured in many pairs of parents and in their progenies. If the variation is due entirely to environment and not at all to heredity, then the expression of the character in the parents and in the offspring will show no correlation at all (heritability = zero). On the other hand, if the environment is unimportant and the character is uncomplicated by dominance, then the means of this character in the progenies will be the same as the means of the parents; with differences in the expression in females and in males taken into account, the heritability will equal unity. In reality, the heritabilities are found to lie mostly between zero and one. (One formal definition of heritability is the ratio between the additive genotypic variance and the phenotypic variance.) Some examples of heritabilities of traits in different animals are given in Table 2.

Table 2: Some Heritability Estimates

trait	correlation
Cattle	
Butterfat percentage	0.6
Milk yield	0.3
Pigs	
Body length	0.5
Weight at 180 days	0.3
Litter size	0.15
Poultry	
Egg weight	0.6
Annual egg production	0.3
Body weight	0.2
Viability	0.1
Drosophila melanogaster	
Abdominal bristle number	0.5
Body size	0.4
Ovary size	0.3
Egg production	0.2

Source: D.S. Falconer, *Introduction to Quantitative Genetics*, 1960.

Heritability estimates

It is important to understand clearly the meaning of heritability estimates. They show that, given the range of the environments in which the experimental animals lived, one could predict the average body sizes in the progenies of pigs better than one could predict the average numbers of piglets in a litter. The heritability is, however, not an inherent or unchangeable property of each character. If one could make the environments more uniform, the heritabilities would rise, and with more diversified environments they would decrease. Similarly, in populations that are more variable genetically the heritabilities increase, and in genetically uniform ones they decrease. In man the situation is even more complex because the environments of the parents and of their children are in many ways interdependent. Suppose, for example, that one wishes to study the heritability of stature, weight, or susceptibility to tuberculosis. The stature, weight, and liability to contract tuberculosis depend to some extent on the quality of nutrition and generally on the economic well-being of the family. If no allowance is made for this fact, the heritability estimates arrived at may be spurious; such heritabilities have indeed been claimed for such things as administrative, legal, or military talents and for social eminence in general. It is evident that having socially eminent parents makes it easier for the children to achieve such eminence also; biological heredity may have little or nothing to do with this.

A general conclusion from the evidence now available may be stated as follows: diversity in almost any trait, physical, physiological, or behavioral, is due in part to genetic and in part to environmental variables. In any array of environments, individuals with more nearly similar genetic endowments are likely to show a greater average resemblance than the carriers of more diverse genetic endowments. It is, however, also true that in different environments the carriers of similar genetic endowments may grow, develop, and behave in different ways.

V. Changing heredity

HEREDITY AS A SOURCE OF CONSTANCY AND CHANGE

The essence of heredity is reproduction of the carriers of the genetic information, the genes. A very remarkable property of the chromosomal deoxyribonucleic acids, DNA, is precisely that these substances have a structure that permits the "messages" they contain, spelled in the "genetic alphabet," to be transcribed exactly in their copies. A dynamic stability, due to the repeated copying of the genetic information, is obviously essential in the hereditary materials. Without it, the complex machinery of the body, developed in the evolutionary history, could not be maintained. On the other hand, if the stability were complete and the self-reproduction of the genes always faultless, evolution could not occur; and life would still be like its primordial source, the first living substance that arose from the inorganic world. Although heredity is a conservative principle, its conservatism is balanced by occasional deviations from the precision of the copying mechanism of the hereditary materials. Such deviations result in mutations (*q.v.*).

USEFUL AND HARMFUL CHANGES

The significance of mutants

Newly arisen mutants are usually detrimental to their carriers. The classical *Drosophila* mutants are, in a sense, a collection of unfit types, most of them unable to survive in nature. Some mutations give rise to serious, or even fatal, hereditary diseases and malformations. Other mutations, studied chiefly in *Drosophila* flies, produce only slight changes in the appearance or physiology of their carriers; they act usually as subvital conditions, which in human genetics are best described as constitutional weaknesses or frail health. Minor mutations seem to be more frequent than the drastic ones; in *Drosophila*, subvitals are certainly more frequent than lethals. There is, however, no doubt—and on this point all competent geneticists are agreed no matter how divergent their opinions may be on other matters—that any increase of the mutation rates in human populations must be injurious to public health. This applies to the mutations induced by high-energy radiations as well as to the spontaneous mutations that arise without known exposures to mutagenic agents.

The fact that most mutations are deleterious may seem to contradict the fairly generally accepted view that the mutation process is the mainspring of evolutionary changes. It should, however, be kept in mind that every species faces not just one standard environment but environments diversified in time and in space. The different variants of a gene arising by mutation may, then, be favourable in some environments and unfavourable in others. The statement that a given mutation is favourable or unfavourable is, therefore, really meaningless if the environment is not specified. Consider the mutation that confers resistance to streptomycin in the colon bacteria. This mutation is useful indeed in an environment that includes streptomycin, for the nonmutants cannot survive at all. On a medium without streptomycin the resistant mutants not only have no advantage but some also are at a severe disadvantage because they cannot live without streptomycin in the medium. What was a poison to the original strain became a necessity to the mutant.

To find useful mutants, experiments must be arranged in one of two ways—either the organisms are placed in unusual environments or one starts with material already carrying a deleterious genetic condition (some sort of a hereditary weakness or malformation). With microorganisms, extensively studied classes of mutants are those that change the nutritional requirements. By placing bacteria or yeasts on culture media with food substances that the original genotype is unable to metabolize, one may select mutants that confer on their possessors such ability. With *Drosophila*, mutants that had originally a poor viability are often observed to improve when kept in laboratory cultures for many generations. It is possible to demonstrate that the improvement is due to selection of favourable mutants modifying the effects of the original mutant in a favourable direction.

The fact remains, nevertheless, that most mutations are harmful, some of them harmful perhaps under all conditions and in all environments. Uncontrolled accumulation of mutations would spell biological disaster—extinction of a species or population. It is basically a consequence of the fact mentioned above, that most mutations are accidents, "misspellings" of the genetic "messages." The mutation process does not, however, constitute evolution; it only supplies the raw materials from which evolutionary changes may be compounded. The compounding agent is natural (or artificial) selection.

VI. Heredity and evolution

Influences on heredity

In the study of heredity the first question that arises is how the genotype of an individual is formed from the constituents of the genotypes of his parents. This is the genetics of individuals or basic genetics. One may also inquire how the genotype in a fertilized egg cell influences the developmental pattern of the organism and thus realizes its potentialities. This is developmental genetics. An individual, at least an individual of a sexually reproducing species, is not, however, biologically complete in itself. Its biological role is actualized through its membership in a reproductive community, a Mendelian population. A Mendelian population consists of individuals among whom matings may or do occur. An individual is mortal and temporary; a Mendelian population has a continuity through time. The genetic processes in Mendelian populations are the subject matter of population genetics.

A Mendelian population is said to have a gene pool. The gene pool is the sum total of the genes carried by the individual members of the population. The gene pool also continues through time. The genes of the individuals of the generation now living come from a sample of the genes of the previous generation; if these individuals reproduce, their genes will pass into the gene pool of the following generations. The Mendelian population and its gene pool in man have a very complex structure. Individuals born and living close together are more likely to meet and to mate than those living far apart. In a widely distributed species such as mankind, the likelihood of mating of individuals born on different continents was, until the development of modern means of travel, very small. The gene pool of the human species is, accordingly, divided into the smaller gene pools of races and populations living in different regions. Aside from the geographic divisions, there are also linguistic, religious, social, economic, and educational barriers that break the gene pools into further, often overlapping, subdivisions. The smallest subdivision is referred to as an isolate or panmictic unit; it consists of a relatively limited number of persons (or animals or plants) that may be regarded as potential mates. Few of these divisions may be sharp enough to decide where one gene pool subdivision ends and the other begins, and yet these subdivisions are biologically meaningful.

A biological species, in sexually reproducing organisms, is defined as the most inclusive Mendelian population. The gene pool of mankind is an entity, the limits of which are not in doubt, since no gene exchange between the human and any other related species takes place. Nor does the intraspecific differentiation impair the unity. There may never have been a marriage of, for example, an Eskimo and a Melanesian, but genetic communications between the Eskimo and Melanesian gene pools occur through the chains of geographically intermediate populations. A genetic change arising anywhere in the world, if favourable, may spread throughout mankind. This is how genetic changes may have transformed the ancestral prehuman species into the present one. The genetic unity of mankind makes any genetic damage (*e.g.*, that caused by exposure to high-energy radiation) a common concern of all mankind, regardless of whether the damage is inflicted more heavily on one portion of the human population than on another.

G.H. Hardy and W. Weinberg independently formulated a theorem that became the foundation of popula-

The basis for population genetics

tion genetics. Two or more gene alleles will have the same frequency in the gene pool generation after generation, until some agent acts to change that frequency. Consider a population that is, as most human populations actually are, a mixture of individuals with M, N, and MN blood types. An individual with M blood is a homozygote with two *M* genes (*MM*), and N individual has two *N* genes (*NN*), and an MN individual is a heterozygote (*MN*). Suppose that a population consists of 49 percent of individuals with M, 42 percent with MN, and 9 percent with N bloods. What will be the frequencies of the blood types in the next and the following generations? Assume for simplicity that (1) marriages are at random with respect to the blood groups, (2) people with different blood groups have neither advantages nor disadvantages in survival or in reproduction, (3) the genes *M* and *N* do not change frequently by mutation, and (4) the population is large enough so that chance fluctuations may be ignored. Also assume that all individuals produce equal numbers of sex cells with each of the pair of alleles they carry and that the sex cells of the parents combine at random in fertilization. Persons with M blood produce sex cells with the allele *M*, and these sex cells will amount to 49 percent of the total. Persons with MN blood produce equal numbers of *M*- and of *N*-bearing sex cells—*i.e.*, 21 percent of each. Finally, persons with N blood will give *N*-bearing sex cells, 9 percent of the total. The gene pool will, therefore, contain 49 + 21 = 70 percent of *M* and 21 + 9 = 30 percent of *N*-type sex cells, or, using decimals, 0.7 *M* and 0.3 *N*, respectively. These sex cells will combine to produce the following blood types in individuals: 0.7 × 0.7 = 0.49, or 49 percent of M; 0.3 × 0.3, or 9 percent of N; and 2 × 0.7 × 0.3, or 42 percent of MN. Generalizing, if the proportions of *M* and *N* genes in the gene pool are *p* and *q* respectively, the frequencies of the blood groups will be, generation after generation:

$$p^2MM + 2pqMN + q^2M = 1$$

This is the Hardy–Weinberg formula, which describes the genetic equilibrium status in populations. The genetic composition of a population can meaningfully be described in terms of the frequencies of various alleles of the genes in the gene pool. Different populations of the same species are likely to differ in the frequencies of some, probably of many, genes. If the gene frequencies are different, the populations are racially distinct; if the differences are large, one may decide to give these populations different racial (or subspecific) names.

NATURAL SELECTION AND DARWINIAN FITNESS

If all gene frequencies remained constant in populations indefinitely, evolution could not take place. Evolution is, in the last analysis, change of gene frequencies (see EVOLUTION). The four assumptions made above in the discussion of the Hardy–Weinberg equilibrium are, of course, oversimplifications. The mating may be selective rather than random with respect to a given gene. Suppose that persons with M, MN, and N bloods prefer to marry individuals with a blood group the same as themselves. Selective mating will not, by itself, change the gene frequencies, but it will disturb the Hardy–Weinberg equilibrium in another way. The relative frequencies of the homozygotes (MM and NN) will increase from generation to generation, while those of the heterozygotes (MN) will decrease. Eventually the population will consist of homozygotes only. Preferential mating of unlike genotypes will, on the contrary, increase the incidence of the heterozygotes, but it will not, no matter how long continued, eliminate the homozygotes.

The carriers of some genes may survive more often or be more fecund or both than the carriers of other genes. When the carriers of different genes are not equally efficient in transmitting these genes to the succeeding generations, the result is natural selection. When the inequality of the transmission rates of the genes is imposed by man's will, the result is artificial selection. In general, the genes that confer on their possessors a superior reproductive efficiency will increase in frequencies from generation to generation, and the reproductively inferior genes will become less frequent. Let a gene allele A_1 have a frequency *p* and its alternative gene A_2 the frequency *q* in the gene pool. The values *p* and *q* are fractions, and $p + q = 1$. The formula is then:

Gene frequency

$$(1) \qquad p^2A_1A_1 + 2pqA_1A_2 + q^2A_2A_2 = 1.$$

Suppose that the relative numbers of the surviving progeny left by the carriers of the genotypes A_1A_1 and A_2A_2 are in the ratio 1 : 1 − s (the value s is called the selection coefficient). If for every 100 offspring of A_1A_1 parents only 90 surviving offspring are left by A_2A_2 parents, then s = 1/10 or 0.1. The heterozygotes, A_1A_2, may leave as many progeny as A_1A_1 (if A_1 is dominant) or as many as A_2A_2 (if A_2 is dominant) or an intermediate number (if neither is dominant). The Hardy–Weinberg formula is then modified thus:

$$(2) \qquad p^2A_1A_1 + 2pqA_1A_2 + q^2(1 - s)A_2A_2 = 1;$$
$$(3) \qquad p^2(1 - s)A_1A_1 + 2pqA_1A_2 + q^2A_2A_2 = 1;$$
$$(4) \qquad p^2A_1A_1 + 2pq(1 - hs)A_1A_2 + q^2(1 - s)A_2A_2 = 1.$$

The value h is the coefficient of dominance. The heterozygotes may exhibit the quality of hybrid vigour (heterosis) and be reproductively superior to both homozygotes A_1A_1 and A_2A_2. The expression is then:

$$(5) \qquad p^2(1 - s_1)A_1A_1 + 2pqA_1A_2 + q^2(1 - s_2)A_2A_2 = 1.$$

And finally (though this is rare except in species hybrids) the heterozygotes may be at a disadvantage compared to both homozygotes. The situation is simplest if the heterozygotes (A_1A_2) are equal in reproductive efficiency to one of the homozygotes or intermediate between the two (formulas 2–4 above). Whichever gene, A_1 or A_2, confers a superior reproductive efficiency on its possessors will increase in frequency in the population. The increase will continue generation after generation; given enough time (*i.e.*, enough generations) the more efficient gene will eliminate and supplant the less efficient one entirely. How rapid or slow the gene frequency changes will be depends on the magnitude of the selection coefficients. Table 3 gives some examples for a dominant allele, a recessive allele, and for no dominance—*i.e.*, for the case when the fitness of the heterozygote is intermediate between the two homozygotes. The two alleles in the population with which the selection starts are assumed to be equally frequent, $p = q = 0.5$. The selection coefficients of 1 (lethal), 0.5, 0.1, and 0.01 are considered, and the gene fre-

Table 3: The Decrease of the Frequencies of Genes Discriminated Against by Different Selection Pressures
(the frequency of the gene in the initial population is 0.5)

generations	strong selection (s = 1.0)	intermediate selection (s = 0.5)	intermediate selection (s = 0.1)	weak selection (s = 0.01)
	recessive allele			
1	0.33	0.43	0.49	0.499
2	0.25	0.37	0.48	0.497
5	0.14	0.27	0.44	0.494
10	0.08	0.16	0.39	0.488
20	0.05	0.09	0.31	0.476
50	0.02	0.04	0.18	0.442
100	0.01	0.02	0.10	0.391
	dominant allele			
1	—	0.40	0.49	0.499
2	—	0.29	0.47	0.497
5	—	0.07	0.43	0.494
10	—	0.002	0.35	0.487
20	—	—	0.20	0.474
50	—	—	0.01	0.433
100	—	—	—	0.362
	intermediate heterozygote			
1	0.25	0.42	0.49	0.499
2	0.13	0.34	0.47	0.497
5	0.02	0.17	0.43	0.494
10	0.001	0.04	0.37	0.487
20	—	0.003	0.26	0.475
50	—	—	0.07	0.438
100	—	—	0.006	0.377

quencies after 1, 2, 5, 10, 20, 50, and 100 generations of such selection are given.

The homozygotes for a recessive gene allele may not reproduce at all. With natural selection this may happen because they are inviable (lethal) or sterile; with artificial selection the same result is accomplished if the breeder kills them or does not use them as parents. The recessive allele is then opposed by a selection s = 1. Table 3 shows that a gene with an initial frequency of 0.5 will decrease to 0.33 after one generation, to 0.25 after two, and to 0.01 after 100 generations. With weaker selection (s smaller than unity) the decrease of the recessive allele will, of course, be slower. Note, however, that in all cases the frequency change is more rapid when the gene is common than when it is rare. A dominant allele opposed by a selection s = 1 (a dominant lethal) disappears in a single generation, and even weaker selections against dominants are more efficient than similar selections against recessives. A selection in favour of a recessive is, of course, just the reverse of that against a dominant, and vice versa. The frequencies can be read from Table 3 by subtracting the frequencies given from unity. When the dominance is absent, the efficiency of selection is, as shown in Table 3, intermediate between those for recessives and for dominants.

The attainment of balanced polymorphism

A most interesting, and at first sight paradoxical, outcome of selection would arise if the heterozygote is superior to both homozygotes, as shown in equation 5. Neither the gene A_1 nor A_2 is allowed to crowd the other out or to disappear entirely. Instead, a genetic equilibrium is reached, and the population attains the state of the so-called balanced polymorphism. All three genotypes continue to occur in the population, with frequencies dependent on the relative magnitudes of the selection coefficients s_1 and s_2. This will be true even if one of the homozygotes is seriously incapacitated, inviable, or sterile. The possible importance of this in human populations is considered below.

Darwin's description of the process of natural selection as the survival of the fittest in the struggle for life is a metaphor. "Struggle" does not necessarily mean contention, strife, or combat; "survival" does not mean that ravages of death are needed to make the selection effective; and "fittest" is virtually never a single optimal genotype but rather an array of genotypes that collectively enhance population survival rather than extinction. All these considerations are most apposite to consideration of natural selection in man. Decreasing infant and childhood mortality rates do not necessarily mean that natural selection in the human species no longer operates. Theoretically, natural selection could be very effective if all the children born reached maturity. Two conditions are needed to make this theoretical possibility realized: first, variation in the number of children per family and, second, variation correlated with the genetic properties of the parents. Neither of these conditions is farfetched.

Darwinian fitness is sometimes referred to also as the adaptive value or the selective value; these terms are best treated as synonyms, although they may have somewhat different connotations. The Darwinian fitness of a genotype, or of a group of genotypes, is measured as the contribution of their carriers to the gene pool of the succeeding generation, relative to the contributions of other genotypes present in the same population. In the example given above, the Darwinian fitness of the genotype A_1A_1 was taken to be unity and the fitness of A_2A_2 as $1 - s$ or less than unity. The fitness is, of course, subject to change in different environments; the carriers of a genotype A_1A_1 may leave more surviving progeny than A_2A_2 in a certain environment, but the reverse may be the case in another environment. Darwinian fitness is reproductive fitness; bodily or mental vigour, health, and energy obviously contribute to this fitness but only insofar as they result in a superior reproductive capacity. Mules, no matter how strong and resistant, must be ranked zero in Darwinian fitness because they are sterile. The emphasis on reproductive success rather than on survival is characteristic of the modern concept of natural selection, as distinguished from the classical one. The difference is not, however, so great as it may seem at first glance; the carriers of a genotype evidently must survive in order to reproduce, and they must reproduce in order to survive in the next generation.

VARIETIES OF NATURAL SELECTION

There are several kinds of natural selection, rather different in their biological consequences and in their importance to man. The simplest of them is the normalizing selection, which was already known before Darwin but, of course, not under this name. Normalizing selection counteracts the accumulation in populations of hereditary diseases, malformations, and weaknesses. Suppose that a gene allele A_1, the carriers of which have a high Darwinian fitness, mutates to a state A_2, which lowers the fitness. If A_2 is a dominant lethal or a gene that renders its carriers sterile, then (as shown in Table 3 column s = 1.0) all the A_2 mutants will be eliminated in the same generation in which they arise. A new crop of mutants will, of course, appear in the next generation. If, however, the selection is not so completely efficient, some mutant genes will escape its dragnet and will be transmitted to the next generation. That generation will contain all the newly arisen mutants, a part of the mutants that arose in the preceding generation, a smaller part of those having arisen two generations ago, etc. How great a "genetic load" of uneliminated mutants a population can accumulate will depend principally on two factors—how often the mutation arises and how much it lowers the Darwinian fitness.

Normalizing selection

Simple formulas have been worked out to describe the situations that arise. Suppose that a deleterious mutation from $A_1 \rightarrow A_2$ occurs at a rate u per generation. Suppose further that the mutant is discriminated against by a selection coefficient s. If, then, the mutant A_2 is dominant to the original state, A_1, the frequency of A_2 in the gene pool will be u/s. If A_2 is recessive to A_1, its accumulated frequency will be much higher, namely $\sqrt{u/s}$. The reason deleterious recessive mutants are allowed to attain higher frequencies than equally deleterious dominants is simple: a recessive mutant may be carried in many heterozygotes, in which it does not express itself, and is consequently protected, or sheltered, from the weeding-out action of natural selection. With mutants that are neither dominant nor recessive, the accumulation will be between u/s and $\sqrt{u/s}$.

All human populations doubtless carry genetic loads consisting of harmful mutant genes. This cannot be blamed entirely on culture, civilization, or on any other specifically human attributes. Populations of *Drosophila* flies and of other sexually reproducing organisms also carry genetic loads. The accumulation of the genetic loads is a necessary consequence of the occurrence of mutations, most of which are harmful but not always harmful enough to be eliminated immediately after they are produced. Harmful mutations are accumulated until the numbers of the respective mutant genes become equal to the numbers eliminated by natural selection in the same population. The population is then said to be in the state of "genetic equilibrium." H.J. Muller has termed the elimination of harmful mutants "genetic death." Genetic "death" is sometimes cruel, sometimes rather benign. The death of a child from a severe hereditary disease and the genetically conditioned failure to have one more child are both genetic "deaths." The higher the mutation rates, the more harmful the mutants produced and the more frequent the genetic deaths. In populations that have reached genetic equilibrium, the total number of genetic deaths will be equal to the total number of the mutations subject to normalizing natural selection.

Heterotic balancing selection

A very different form of natural selection is heterotic balancing selection. It occurs when the Darwinian fitness of a heterozygote exceeds the fitness of both homozygotes, a situation mentioned above in equation 5. Heterotic balancing selection also leads to genetic equilibrium but not to an equilibrium between mutation and the normalizing selection. The balanced polymorphism that is es-

tablished is due to the selection favouring the heterozygotes against the homozygotes. In a sexually reproducing population the heterozygotes tend, however, to produce a fresh crop of homozygotes in every generation. The maintenance in human populations of such grave hereditary diseases as thalassemia and sickle-cell anemia is apparently due to this form of selection. The homozygotes for the sickle-cell and thalassemia genes are lethal; their Darwinian fitness is zero, and they are discriminated against by selection coefficients $s_2 = 1$. But while the homozygotes are lethal, the heterozygote for the sickle-cell gene has a higher fitness than the "normal" homozygote in countries in which falciparum malaria is endemic, apparently because of its relatively greater resistance to that disease.

In malarial environments, populations that contain some thalassemia and sickle-cell anemia genes, therefore, have advantages over populations free of these genes. The former populations are in less danger from the ravages of malaria, although they "pay" for this advantage by sacrificing in every generation some individuals who die of anemia. This advantage, of course, does not hold in countries in which falciparum malaria is rare or absent. Populations native in these countries have, as expected, few or no thalassemia or sickle-cell genes. The lethal diseases caused by homozygosity for these genes certainly bring about some "genetic deaths"; they are a part of the genetic load of the populations. But this genetic load, due to a disadvantage of being homozygous for certain genes, is very different from the mutational load controlled by the normalizing selection. The former is maintained by the heterotic balancing selection, while the latter is maintained by recurrent mutation.

Diversifying selection

Another form of natural selection is diversifying, or disruptive, selection. In many discussions and mathematical analyses of selection this simplifying assumption is adopted: that the environment in which a population lives is uniform and that the selection advantages and disadvantages of different genotypes are independent of their frequencies in the population. This simplification, however, flies in the face of reality. Many animals can subsist on a variety of foods; many plants grow on different soils; humans have to fill many different employments, functions, professions, and social roles. It is most likely that some genotypes will be fitter in some environments than they are in other environments. Diversifying selection will then favour different genotypes in different subenvironments, or ecological niches, that occur in the population.

A special form of selection occurs in mammals due to the incompatibility of certain maternal genotypes with those of their unborn children. The best studied case in man is that of a Rhesus-positive fetus in a Rhesus-negative mother (see BLOOD GROUPS). This selection should, theoretically, make the entire population either Rhesus positive or Rhesus negative. It does not appear to be doing this for reasons that have not been clarified. Another special kind of selection is that due to so-called meiotic drives, disturbances of the Mendelian segregation mechanisms, which result in sex cells carrying certain gene alleles being more or less frequent than expected on a random basis. The last to be mentioned, but in the long run possibly the most important form of selection, is directional selection. Suppose that the climate becomes warmer or cooler, that there appears a new source of food or a new predator or disease, or that there occur some other prominent environmental changes. Some genotypes will, then, become more favourable and others less favourable. Directional natural selection will operate to reconstruct the gene pool of the population in accord with the demands of the new environment.

NATURAL SELECTION IN OPERATION

When the antibiotic streptomycin is added in a sufficient concentration to a culture of colon bacteria, only the streptomycin-resistant mutants are able to reproduce, while the streptomycin-sensitive cells are eliminated. If an experimenter adds the streptomycin to a culture in order to obtain resistant strains of the bacteria, he is effecting artificial selection. Natural selection acts, however, very much in the same fashion. Widespread and often rather indiscriminate utilization of antibiotics against various infectious micro-organisms led to quite unintended consequences—the appearance of numerous "drug-fast" infections. In some instances (as with gonorrhea in some countries) the treatment with penicillin, originally very effectual, became less reliable or even useless as selection brought about a drug-resistant strain of gonococci.

Genetic resistance to DDT

Essentially the same process took place with considerably more complex organisms, various insect pests, in response to the introduction of certain highly potent insecticides. The first such insecticide, DDT, was used on a large scale during and after World War II. The discovery of DDT, followed by invention of many other insecticides of similar or even greater potency (see PEST CONTROL), led to widespread hope of eradication of all insect pests. This enthusiasm was short lived, since by 1947 DDT-resistant populations of the housefly were recorded in Italy and soon thereafter in other countries. It apparently takes the fly only two to three years to evolve a genetically conditioned resistance to DDT, and similar resistances arise against other insecticides as well. During the Korean conflict DDT-resistant lice emerged. Attempts to eradicate malaria-bearing mosquitoes by DDT sprays resulted in DDT-resistant mosquito populations. Laboratory experiments with *Drosophila* and with houseflies have shown that one can artificially select for insecticide resistance. Moreover, there are apparently genetically different resistant strains, some differing from the sensitive ones by a single gene with a strong effect and others by accumulation of several genetic changes, each taken separately increasing the resistance only slightly but giving strong resistance in the aggregate.

A black mutant of the peppered moth (*Biston betularia*) was found at Manchester, England, in 1848. Until that time the prevalent form of this species was light gray with dark speckles. The frequency of the black variant rapidly increased in the immediate area and reached about 99 percent by 1898. Dark coloration, melanism, has arisen and spread in many species of British moths belonging to different genera and families. The spread of the melanic forms was most rapid in and near industrial areas, where the foliage and the tree trunks are blackened with soot and other atmospheric pollutants. The spread of the dark forms is consequently known as industrial melanism. The light form of *Biston betularia* is protectively coloured when resting on tree trunks covered with lichens but conspicuous on blackened trunks on which the lichens have been largely destroyed by pollution. The selective agents are the insectivorous birds that feed on the moths; they find the light moths easily on dark tree bark and the black moths easily on light tree trunks. H.B.D. Kettlewell has verified the hypothesis by exposing known numbers of light and of melanic moths in localities with blackened and with unpolluted vegetation. A greater proportion of the blacks than of the whites remained alive and were recaptured on dark vegetation, and the opposite was observed on the unpolluted vegetation. Genetically, the difference between the light and the melanic forms has been shown to be due to a single gene, the allele for melanism being dominant to that for the light coloration.

Plant genes modified by man

Plants that grow in man-made or man-modified habitats are often different from their wild relatives and different in ways that adapt them to survive and reproduce in their respective habitats. *Camelina sativa*, a plant of the mustard family, is a common weed in fields of cultivated flax in Europe. The *Camelina* that contaminates the flax fields is, however, genetically different in several ways from that growing outside the fields. *Camelina* outside flax fields is a low-growing, branched plant with small seeds; that growing in flax fields is much taller, unbranched, and with larger seeds that resemble those of the flax in size and in specific gravity. The characteristics of *Camelina* growing in flax fields are such that it is harvested together with the flax, and its seeds remain with

those of flax during the winnowing process. The variety in the flax fields is doubtless derived from the original form now found growing freely; the remarkable adaptations that are found in the former have, then, developed since man started to plant flax on fairly large scales. Genetically, the varieties of *Camelina* differ from each other in several, probably in many genes. The process of natural selection—incidentally abetted by man—has, thus, created a new adaptive genetic system.

The examples of drug resistance in bacteria, pesticide resistance in insects, and industrial melanism in moths all involve directional natural selection, which, if continued long enough, leads to replacement of old gene alleles by new ones. Balancing natural selection has a different biological function—it maintains genetic diversity, or polymorphism, in populations. Selection pressures involved in balancing selection in nature are sometimes quite high. This is very useful to those engaged in research on selection, since it is evidently easier to detect strong selection than weak selection.

Striking examples of heterotic balancing selection have been found in nature in populations of some *Drosophila*. Many species of *Drosophila* are polymorphic for gene arrangements in their chromosomes; two or more kinds of chromosomes differing in inversions of blocks of genes occur in the same population. The carriers of the different chromosomes interbreed freely, and both homokaryotypes (two chromosomes of a pair with the same gene arrangement) and heterokaryotypes (the chromosomes of a pair with different arrangements) occur in different individuals. Some populations of *Drosophila pseudoobscura* in the western United States undergo cyclic seasonal changes in the frequencies of the chromosomal variants. The kinds of chromosomes denoted *ST* are relatively common in spring and autumn and rare in June; the chromosomes with the gene arrangement *CH* are, on the contrary, common in midsummer and relatively rare in spring and in autumn. The changes observed in nature can be reproduced in laboratory experiments. Experimental populations can be created with any desired frequencies of the chromosomal variants and can be left breeding freely for many generations. Samples are taken from time to time to determine the frequencies of the different kinds of chromosomes by microscopic examination, and the selection coefficients and the Darwinian fitnesses of the karyotypes can be calculated. As a rule, the heterokaryotypes show heterosis, fitness greater than in the homokaryotypes; selection coefficients of the order of 0.5 and even higher are often encountered.

GENETICS OF RACE AND SPECIES DIFFERENCES

The nature and origin of races and species are dealt with in other articles (see EVOLUTION; and SPECIES AND SPECIATION). Here it is necessary to consider only the genetic composition of the race and species differences in sexually reproducing and outbreeding organisms. In naturally occurring populations a species may split into races because of a gradual geographical separation and eventual rift between formerly interbreeding groups. Such races, which inhabit different territories, are called allopatric. If these races are brought together, they assume a sympatric status, interbreed, exchange genes, and fuse into a single genetically variable population. Races of man and of certain parasites are exceptional because they can coexist, at least for a time, sympatrically. In man social rather than geographical or biological factors slow down interbreeding and race fusion. Distinct species may, on the other hand, be either allopatric or sympatric. The exchange of genes between species populations is prevented not only by geographic distance (as with races) but also by genetically based reproductive isolating mechanisms. Reproductive isolation is achieved by a variety of means: differences in preferred habitats, in breeding seasons, in sexual attraction and courtship rites, in sexual structures (flowers in plants and genitalia in animal); incompatibility of sex cells; inviability of the hybrid progenies; sterility of the hybrids; and weakness of the gene recombination products of the gene complements of the species.

Types of races

Race differences are more often quantitative than qualitative; racially distinct populations of a species differ usually in the frequency of certain genes rather than the presence or absence of certain genes. Studies on blood groups in man have revealed some most instructive situations. The four "classical" blood groups O, A, B, and AB are due to three alleles of a gene. Most human populations have individuals of all four types, and even parents and children, as well as brothers and sisters, may belong to different blood groups. However, some blood groups, and hence the gene alleles that produce them, are more frequent in some countries than in others. The gene for A blood, for example, increases in frequency from east to west in Europe; B blood is most frequent in some populations of India, Tibet, Mongolia, and Siberia. A majority of American Indians apparently had, before the arrival of Europeans and Africans, the O blood group only; however, the tribes of the Blackfoot and Bloods had the highest known frequencies of A blood, which is also very frequent among the Lapps in northern Europe (see also BLOOD GROUPS).

Human races differ certainly in many genetic traits, not in blood groups alone. Many genetic differences are the rule also between races of animals and plants (races are often referred to as subspecies in wild and as breeds or varieties in domesticated forms). The blood groups are, however, useful as a model to elucidate the genetic nature of race differences in general. Races are genetically open systems, and gene exchange between races does take place. Species, by contrast, are genetically closed systems in which gene exchange is rare or absent. Race differentiation is reversible; hybridization or intermarriage may cause races to merge into a single population. It is an error to think that in a population resulting from race hybridization everybody will be alike; in point of fact, such hybrid populations show a remarkable diversity of individuals. Species differentiation is irreversible. Races are populations, and an individual may have a genetic endowment that can occur in two or more different races or that is not common in any race. An individual belongs, however, to only one species, unless he is a species hybrid. Mules, hybrids between the horse and donkey, are sterile because of abnormalities in the processes of sex-cell formation in their gonads. Sterility of hybrids between species, if viable hybrids between them can be obtained at all, is observed very frequently, though some experimentally obtained species hybrids have proved to be fertile.

The origin of racially distinct populations

What causes the development of the gene-frequency differences between populations that live in different territories; in other words, what makes these populations racially distinct? The probable explanation is that genetic differences between populations arise in most cases through natural selection in response to the local environments that prevail in the territories they inhabit. It is, however, very difficult to verify this explanation in many concrete instances of race differentiation. For example, it is probable that the dark skin pigmentation of many human populations that live, or have until recently lived, in tropical and subtropical countries protects them from sunburns. It is probable also that the light skin colours of the natives of Europe facilitate the acquisition of vitamin D in regions with deficient sunshine. The evidence for even these hypotheses is not as complete and conclusive as might be desired. But when it comes to such racial traits as hair form (straight, wavy, curly, or frizzly), shapes of the nose, of the lips, of the cheekbones, etc., no acceptable evidence of their adaptive significance is available. The situation is no better with races of animals and plants: while some differences are certainly, or at least probably, adaptive, for most of them the adaptive significance is unknown.

Attempts have been made to envisage factors other than natural selection that could be responsible for genetic differentiation of populations. Appeals are frequently made to pleiotropism of the gene action; a visible race difference may in itself be neither adaptive nor unadaptive, yet it may be only an outward sign of an underlying physiological difference that is adaptively important. An elegant

example is the coloration of onions—red and purple bulbs are resistant to the attacks of a smudge fungus, while white bulbs are highly susceptible. A race trait may also be important as a sexual recognition mark, or it may play a role in the courtship ritual.

A most interesting possibility that should be seriously considered is that some differences between populations may be due to random genetic drift. One of the assumptions made in deducing the Hardy–Weinberg equilibrium theorem was that the population is infinitely large, or at least large enough so that chance events will not be seriously disturbing. Suppose, however, that a species lives in many isolated colonies, some of them consisting of only tens or perhaps hundreds of individuals. In such small colonies the number of individuals that inherit the gene allele *A* rather than *a* may, owing to chance alone, be in some generation relatively higher or lower than in a preceding generation. Gene frequencies in the different colonies will drift apart; in some colonies the allele *A* may become rare or may be lost altogether, while in other colonies it will be the allele *a* that will diminish in frequency or be eliminated. How important this random genetic drift may be in race differentiation is controversial. That genetic drift does occur is certain; a simple example is that in small villages a sizable fraction of the inhabitants sometimes have the same surname, and different surnames are frequent in different villages. Increasing or decreasing frequencies of the surnames evidently go together with increases or decreases of the frequencies of certain genes that the ancestors of the people with these surnames carried. Another situation in which genetic drift is likely to take place is when just one pregnant female or a small group of individuals colonizes a hitherto uninhabited island or an isolated mountain or lake area. The genes brought by the founders of new colonies will often be small, atypical, and often unbalanced samples of the gene pools of the populations whence they came. When the populations of the new colonies expand, they will be found to differ genetically from each other and from the ancestral population. Natural selection will then come into operation, giving rise to new balanced gene pools. This "founder principle" was probably important in the development of some human populations. Many tribes and local races may be the descendants of small numbers of original migrants and settlers. Whether the random genetic drift alone can explain the origin of the gene complexes that differentiate races or species is very doubtful. The point is, however, that genetic drift and natural selection are not mutually exclusive alternatives; it is not one or the other but the interaction of both that brings about race differentiation. The founder principle is a special case of random genetic drift. The gene pool of a colony derived from a single immigrant or several pairs of immigrants may need a restructuring by natural selection to become properly adapted to the new environment.

THE SUDDEN ORIGIN OF NEW SPECIES

The question is sometimes asked whether a new biological species has ever been observed to arise or whether all experimentally observed evolutionary changes have been only on intraspecific levels. There are at least two types of species formation, and one of them is rapid enough to be experimentally observable. Races (subspecies) of the same species differ in frequencies not of one but of several or many genes. Distinct species differ doubtless in still more genes, although in just how many cannot be determined exactly. The complexity of the species differentials is, however, demonstrable in those rather exceptional cases when fertile hybrids between species can be obtained in experiments and a second hybrid generation raised in large enough numbers and carefully studied. In such interspecific hybrid progenies an extraordinary diversity is often found; no two individuals are quite alike, and some specimens exhibit peculiar combinations of traits that render them sometimes strikingly different from both parental species and from the first generation hybrids. All this indicates that distinct species carry different, harmoniously balanced systems of genes.

Complexity of species differences

Mixtures and recombinations of the components of these balanced systems give rise to genotypes, most of which are poorly adapted, but a few of which result in satisfactory or even superior fitness. The gene exchange between species is impeded or excluded altogether by reproductive isolating barriers. The biological function of these barriers is limitation or prevention of the appearance of ill-adapted gene combinations.

The process of race (subspecies) formation consists, then, of accumulation of genetic differences, most of them adaptive to the environments these races inhabit. The process of species formation (speciation) consists, in addition, of building reproductive isolating barriers, which takes many generations. It is not surprising, therefore, that this kind of speciation is not fully reproducible in laboratory experiments, although some beginning steps in this direction have been made. New species may, however, arise suddenly, in a single leap, by doubling the chromosome complement in hybrids between two other pre-existing species. This form of speciation is fairly common in many families of plants, much less so in animals (see SPECIES AND SPECIATION).

VII. Heredity and applied science

MEDICO-LEGAL APPLICATIONS AND GENETIC COUNSELLING

Heredity and variation are basic functions of living organisms, and they can be studied regardless of any prospects of technological or other practical applications of the knowledge thus gained. And yet the relevance and applicability of genetics to human affairs is expanding rapidly. It is within the realm of possibility that genetics will become the most important of all applied sciences if it comes to be used to improve man and to direct the evolution of mankind. Genetic engineering may, conceivably, become the pre-eminent form of engineering. These grand and imposing developments are, however, mostly in the future, but some relatively modest applications are even now being used.

Disputed paternity is one of the forensic problems that may sometimes be elucidated by genetic deductions. A man being sued in court for support of a child of whom he does not consider himself the father may seek a "paternity exclusion." In a notorious case in which justice obviously was not done, a well-known actor, whose blood was group O, was forced to pay the attorney's fees and support a child with group B blood, whose mother had A blood. Excepting the extremely unlikely possibility of mutation, an O father (*OO* genotype) and an A mother (either *AA* or *AO*) could not produce a B child because neither of them has a *B* gene to contribute. More recently courts have accepted blood tests as evidence in disputed paternity cases. Note should, however, be taken that while exclusion of paternity can be established, no known tests can affirm that a man X is the father of a child Y, because there may always be more than a single person who carries the same testable genes as X.

The importance of genetic counselling

A geneticist is often asked for an assessment of the chances of birth of genetically handicapped children to prospective parents. The variety of problems that come up in such inquiries is endless, and no geneticist in the world is competent enough to solve all of them. There are hundreds of diseases and malformations in the production of which genetic variables play a part, but all too often the mode of inheritance of these defects (*i.e.*, dominance, recessivity, sex linkage, penetrance, etc.) is not precisely known. The demand for genetic counselling is, however, so widespread and frequently so urgent that more and more so-called heredity clinics and counselling centres are appearing, attached usually to universities and medical schools. In some instances reasonably clear-cut answers can, of course, be given. Suppose, for instance, that a prospective parent is himself healthy but has one or more siblings with a defect or a disease known to be inherited as a simple dominant Mendelian trait. Provided that the dominant gene in question has complete penetrance, the parent can be assured that he is just as unlikely to transmit the respective defects to his children as are persons with no known relatives with these defects.

The prognosis is less reassuring if the parent himself is a carrier of a dominant genetic defect, for in this case the chances are even that any child of his will or will not inherit that defect. The genetic risks are less with recessive defects, such as albinism or phenylketonuria or infantile amaurotic idiocy. A couple that has already produced one affected child stands one chance in four of conceiving another child similarly affected (of course, three out of four chances are that the next child will be unaffected, but in cases of this sort the dire consequence is the pressing one). Whether the counsellor should recommend to a couple that they refrain from having more children is another question; perhaps the wisest course is to explain the situation as clearly as possible and leave the decision to the prospective parents themselves. If they do have more children, they will at least be aware of the risk and be prepared for what may be in store. If recessive defects are found in the relatives of one or both prospective parents, the advice has to be even less confident but more optimistic, since the likelihood of both prospective parents being heterozygous carriers may be quite small. The matter can be clarified better with those recessive diseases that (as is the case with phenylketonuria) render the heterozygous carriers recognizable by certain biochemical tests. If at least one of the parents to be is a noncarrier, the risk is reassuringly small. The predicament of a genetic counsellor is greatest with several widespread and grave diseases, which rather certainly have a genetic basis but the inheritance of which has not been sufficiently clarified. Here belong such genetic puzzles as schizophrenia, various allergies, and predispositions to heart disease and to cancer.

OUTBREEDING AND INBREEDING

Man is an outbreeding species. Marriages between close relatives are forbidden by custom and by law in most human societies. Whether the prohibition was introduced because people knew that incest or inbreeding is detrimental is questionable (see CONSANGUINITY). The prohibition may seem reasonable, however, in view of the evidence accumulated by practical breeders and geneticists. In normally outbreeding species, the progenies of matings in which the parents are close relatives tend to be less vigorous than the offspring of unrelated parents. This is a consequence of the genetic loads of deleterious recessive genes carried in populations. A recessive gene is more or less innocuous in the heterozygous condition, but its chances of becoming homozygous are greater in the offspring of parents who are close relatives (and therefore have more similar genotypes) than in families where the parents are not closely related.

The effects of outbreeding and inbreeding

It is worthwhile to stress at this point that the effects of outbreeding and inbreeding on the vigour, viability, and fertility of the offspring depend upon the reproductive biology of the species or the population in which they occur. Outbreeding is not invariably or necessarily invigorating, nor is inbreeding invariably or necessarily debilitating. Wheat is an example of a plant that is predominantly self-pollinating. Self-pollination is, of course, the closest possible form of inbreeding, but this inbreeding does not progressively weaken the vigour of a wheat strain. The converse, however, is observed with corn (maize). Inbreeding decreases the yield, and the intercrossing of inbred lines restores hybrid vigour in the progeny. One may say that hybrid vigour (heterosis) is the normal state of a corn population, while inbreeding and a prevalance of homozygosis is the normal state in wheat. In species and populations in which the reproductive biology is adjusted to outbreeding, consanguinity leads to a decline in the average vigour and to the appearance of relatively many individuals with hereditary diseases and malformations. Just what kind of a malformation or ill health, if any, will appear in a given progeny depends on what deleterious recessive genes were carried by the common ancestor of the relatives who mate. This will certainly be different in different families. The histories of the reigning houses of Egypt and the Peruvian Incas, where the monarchs allegedly married their sisters, are not sufficiently well documented to conclude that in these instances the incestuous marriages resulted in no weakening of the progeny. High childhood mortality may have gone unrecorded; furthermore, "sister" may have been, in some instances, a title bestowed upon a person rather than indicative of an actual relationship.

It is important to study how much inbreeding depression will occur in a given species or population, and a great deal of information in this direction is being accumulated. With rare exceptions, the highest degree of consanguinity that occurs in mankind with appreciable frequency is marriage of first cousins, resulting in homozygosis of 6.25 percent of the genes in the progeny. A genetic counsellor would hardly be justified in discouraging all marriages of people known to be relatives, and yet he may point out that the chance of genetic weaknesses in the progeny of such marriages is measurably greater than when unrelated persons marry.

GENETIC LOAD AND HYBRID VIGOUR

How hereditary variability is maintained in populations of sexually reproducing species is an unsettled and controversial question. The classical theory assumes that the pressure of mutation generating detrimental genetic variants is the main factor. The balance theory, while not denying that some variability is maintained by recurrent mutation, considers that a far from negligible portion of the variability is preserved by natural selection. These two theories, though they are not incompatible alternatives, have interestingly different theoretical and even philosophical implications.

According to the classical theory, a majority of the genes a species has are fixed, uniform, and homozygous in most, or even in all, individuals. The minority of the genes that are not fixed are represented by two or more variant alleles. One allele of each gene is "normal," and it confers a high fitness on its possessors; the others are mutant, abnormal, and more or less deleterious. This view is logically connected with the conception of the "optimal" genotype. The optimal genotype is the one that contains only normal genes. Any deviation from the optimum genotype will be detrimental. Nevertheless, the process of mutation unrelentingly generates detrimental mutant genes, lowers fitness, and creates a genetic load. A genetic load can, in this view, be defined as the deviation of the observed fitness from that produced by the optimum genotype. The weakness of this definition is obvious: the optimum genotype is operationally elusive; it cannot be located and its fitness cannot be measured.

The balance theory of population structure

The balance theory of population structure takes a somewhat more optimistic view. Not all genetic variability present in populations is a regrettable departure from the optimal genotype. Some of this variability is, on the contrary, an adaptive device that permits the population to secure a firmer hold on its environment. The environment is never uniform; it has a greater or lesser variety of ecological niches, opportunities for living that members of the population can exploit. A living species or population faces many diverse environments, not a single environment. Two genetic strategies are possible to cope with environmental complexity. First, there may be genotypes that react favourably in several environments. Second, there may be a variety of more or less specialized genotypes to match the variety of environments. Nature has used both strategies in the evolutionary process. A sexually reproducing population contains a variety of hereditary endowments. Most, and perhaps all, individuals may be heterozygous for many genes. Any gene that is rare will be carried in heterozygotes much more frequently than in homozygotes (except in highly inbred or self-fertilizing species). Natural selection operates in sexual populations chiefly with heterozygotes. It will tend to make the heterozygotes highly fit, even if the corresponding homozygotes are low in fitness. In other words, natural selection promotes hybrid vigour or heterosis and maintains the variant genes in a state of balanced polymorphism. It takes a variety of genetic endowments to make a world worth living in.

Natural selection cannot prevent the appearance of some individuals of inferior fitness. In a population in which no two individuals carry the same genes, it is not possible to describe the genetic load as the deviation from the optimal genotype. This latter will be represented by a single individual or not at all. An operationally meaningful measure of fitness is that of the average or norm for a given population. The pronounced negative deviants from this average may be considered as constituents of the genetic load and the positive deviants as the genetic elite. Between the genetic load and the genetic elite there is the adaptive norm comprising a majority of the genotypes that are formed. The division lines between the load, the adaptive norm, and the elite can only be established by a convention. For most purposes, the load and the elite should probably refer to fairly extreme negative or positive deviations from the population mean. One may, however, foresee that some genetic processes will cause these deviations to become more or less frequent. For example, a genetic radiation damage or a greater incidence of marriages between relatives would increase the expressed genetic load compared to what it is at present. Positive eugenics may attempt to expand the genetic elite in mankind.

The most impressive success scored to date, in attempts to utilize the genetic elite of an agricultural plant for increasing yields of a useful product, is represented by hybrid corn. Hybrid corn is obtained by artificially crossing inbred corn (maize) strains in order to exploit the phenomenon of hybrid vigour or heterosis.

GENETIC IMPROVEMENTS OF ANIMALS AND PLANTS

Hybrid corn is the most impressive but by no means the only achievement of genetics applied to agriculture. General accounts of ANIMAL BREEDING and PLANT BREEDING can be found in the articles under these titles. This section is concerned only with the genetic basis of the methods used in breeding programs. Attempts to emulate the methods that proved so successful with corn are being made in several species of crop plants as well as in domestic animals. The results look promising, especially with poultry and swine. It should be kept in mind that the inbreeding-heterosis technique is most likely to succeed with species that are normally cross-fertilizing and outbred and less so in habitually inbreeding or self-pollinating forms (like wheat, barley, and some other crop plants).

Artificial selection

The basic method of improvement is artificial selection. The genotypes that are selected are those that, when placed in suitable and economically feasible environments, induce in their carriers the qualities desired by the breeder. The improvement methods used in the breeding work must evidently be appropriate to the reproductive biology, the genetic population structure, and the economic value of the form to be improved. It is evidently impossible, for example, with large domestic animals such as cattle or horses to raise numbers of individuals from which selection is made comparable to that practicable with field crops. Collections of pedigrees, published in stud books for horses and in herd books for some breeds of cattle, are still regarded by commercial breeders as important in the choice of most prized animals. Progeny and sib testing are now being used more and more often as aids for evaluation of the breeding worth of individual animals. A quite different situation is encountered with some useful plants in which cross-pollination happens only rarely, and the normal method of propagation is self-pollination (as in most cereals, peas, beans, tomatoes, etc.) or parthenogenesis or asexual reproduction (sugar cane, bluegrass, bananas, many citrus fruits, etc.). Here the problem is to identify the individual, or a small group of individuals, with a desired combination of traits and to propagate them to form pure lines (progenies obtained by self-pollination) or clones (progenies obtained asexually). Some of the most valuable varieties of wheats are descended from a single seed progenitor.

Selection, artificial as well as natural, can operate only if genetic diversity exists among the materials available. In a genetically uniform pure line or a clone, selection is without effect. Provision of genetic variability is the prime concern for a breeder. A powerful means of inducing genetic variability is hybridization and Mendelian segregation in the hybrid progenies. Two lines, neither of which is particularly good, may produce valuable genotypes in segregating hybrid progenies. This is the reason that modern breeders frequently endeavour to collect primitive varieties from far-off lands and use them as materials for hybridization followed by selection. Many, if not most breeds and varieties, from thoroughbred horses to cultivated strawberries, are descended from hybrids of two or more ancient breeds or even of distinct species. The cultivated roses have genes from at least four wild ancestral species in their gene pool.

A promising new technique to speed up the improvement of useful plants and animals is induction of mutations that increase the amount of genetic variability available for selection. Although a decided majority of mutants are harmful, some exceptional ones may be useful.

IMPROVING MANKIND

Man has been remarkably successful in improving, for his benefit, the genetic endowments of his domesticated animals, cultivated plants, and even micro-organisms. The question inevitably presents itself whether man can succeed equally well in directing the evolution of his own species toward goals he regards as good and desirable. Man is the only organism on earth who is aware that he is a product and even the crowning achievement so far of the evolutionary process. The past evolution was a product of natural selection, a certain pattern of blind forces of nature. Can man gather enough knowledge and enough wisdom to substitute for these blind forces a conscious control of the ongoing evolution of mankind? The central idea of eugenics (*q.v.*) is that such a control is feasible. Ironically enough, thus far the progress of eugenical thought and public acceptance has been impeded more by its overzealous partisans than by its opponents. For decades it was held captive by race and class bigots, and its greatest perversion was its use as a sham justification of the genocidal policies of Nazi Germany. From time to time some fanatics predict mankind's degeneration and even extinction unless their pet eugenical schemes are put in practice without delay. It should, however, be evident that utmost care must be exercised in any attempts to tamper with the genetic endowment of mankind; for example, the implementation of some of the more bizarre eugenic recommendations of several decades ago would have been the worst sort of folly. And yet it is arguable that man will not be able to let his evolution drift uncontrolled much longer.

Negative and positive eugenics

Although there are not sharp dividing lines between them, negative and positive measures of genetic improvement can be distinguished. Negative eugenics is primarily directed toward elimination or, at any rate, reduction of the incidence of hereditary defects and diseases—in other words, toward control of the genetic load of mankind. Positive eugenics favours increasing the frequencies of superior hereditary endowments. Mankind, like any other biological species, has carried a genetic load since the dawn of time. It is claimed, however, that civilized living, technology, medicine, maintenance of the genetically handicapped, etc., increase man's genetic load. The total frequency of mutation in man is conservatively estimated as 10 percent; *i.e.*, at least 10 percent of the sex cells produced carry a newly arisen mutant in every generation. Since, in addition, the mutation rates of human genes are probably increasing as a result of exposure to high-energy radiations and other mutagens, some authorities have expressed alarm concerning the genetic prospect of the human species.

A variety of proposals have been advanced to decrease the genetic load or at any rate to check its increase. It is evident that one way to decrease the frequency of mutation is by avoiding all unnecessary exposure to high-energy radiations and other mutagens. Genetic counsel-

ling could be quite important if it were to become widespread or universal. If known or probable carriers of defective genes were informed of the probable consequences of their begetting children, perhaps some of them would draw the practical conclusions from the information made available to them. More radical proposals would have the carriers of defective genes sterilized or otherwise prevented from reproducing. The effectiveness of such a measure is in doubt because recessive defects are carried mainly in heterozygotes and would escape detection and sterilization, while the dominant hereditary diseases are due mostly to new mutations.

The basic difficulty is, however, that it may not always be easy to decide what artificial selection should select. Does anyone know what will be the best genotype for mankind centuries or millennia hence? In general, dominant defects give rise to the fewest doubts. One can hardly imagine circumstances in which such hereditary disorders as dominant muscular dystrophy, aniridia, multiple polyposis of the colon, or neurofibromatosis may be useful. With recessive defects the thorny question has to be faced whether they are maintained in populations only by the mutation pressure or also by increased fitness of the heterozygous carriers—*i.e.*, by attendant hybrid vigour or heterosis. Mankind free of a genetic load may be not only unattainable but also unacceptable. It may turn out to be a uniformly dull population none of whose members exhibit either great physical or mental vigour.

The claims made by some authorities that selection does not and cannot operate in modern mankind or that decreasing infant and childhood mortalities have done away with selection cannot be accepted without critical examination. The Darwinian fitness of a genotype is defined as its contribution to the gene pool of the following generation relative to the contributions of other genotypes. Differential mortality and differential fertility are, in principle, equivalent as far as selection is concerned. Natural selection can occur either because the carriers of different genotypes produce unequal average numbers of children or because different proportions of the children produced survive or because of any combination of these differences (except, of course, for the possibility that an advantage in fertility is exactly counterbalanced by a disadvantage in survival or vice versa). Decreasing birthrates and childhood mortalities have not necessarily abolished selection, but the direction of the selection in human populations may be undergoing radical changes. Whether selection is at present moving in a beneficial or a detrimental direction is a different issue. Some authorities viewed with alarm the fact that, at least in Western societies, persons with higher intelligence (as measured by intelligence, IQ, tests) have on the average fewer children than persons of lower intelligence. Predictions were made of inevitable decline of the average intelligence with time. Actual surveys have failed to bear out these predictions.

Euthenics and euphenics

Eugenical measures leading to genetic improvements must be accompanied by improvement of the environment in both physical and cultural aspects (euthenics) and by management of the developmental patterns of individual humans (euphenics). No genetic endowment can produce optimal results regardless of the environments in which its carriers are placed. The spectacular advances of the biological sciences have produced optimistic expectations of even more splendid advances to come. Several possibilities can be mentioned, although it is impossible to predict when, if at all, they will be realized. One has been labelled algeny, or "genetic surgery"—*i.e.*, deliberate modification of specific genes or implantation of desired new genes into chromosomes. Transformation and transduction have, indeed, been realized in some micro-organisms; and it is possible, at least in principle, that this kind of directed genetic change may someday be practicable in higher organisms, including man. The potency of genetic surgery for genetic improvement of mankind may, indeed, permit achievements far beyond the reach of the genetic techniques presently available. A variety of instruments for euphenic betterment (developmental engineering) may conceivably become available in the near future. Suppression of the action of undesirable genes and stimulation of that of desirable genes may be a most powerful means for controlling phenotypes that develop from genetically defective genotypes or for stimulating genotypes whose potentialities are hidden. Transplantation of organs to replace those injured by accidents or worn out is another possibility. Such organs may be taken either from people who die accidentally or because of faults in other organs or conceivably from animals bred especially for that purpose. Mechanical aids or prostheses (substitutes) for human organs or limbs may be improved. All this may increase human longevity and decrease the deterioration now unavoidable with aging. Last but not least and probably the most urgent is management and regulation of population growth by reliable, as well as emotionally and aesthetically acceptable, family-planning techniques. It is arguable that the avoidance of a disastrous "population explosion" is the prime necessity to make any genetic improvement of mankind really beneficial.

BIBLIOGRAPHY

Textbooks and popularizations: G.W. and M. BEADLE, *The Language of Life: An Introduction to the Science of Genetics* (1966), a popular presentation of basic facts of genetics; I.M. LERNER, *Heredity, Evolution, and Society* (1968), an excellent account of the present understanding of heredity and of its social and cultural implications; G.L. STEBBINS, *The Basis of Progressive Evolution* (1969), a well-written outline of genetic and evolutionary processes; M.W. STRICKBERGER, *Genetics* (1968), probably the most complete textbook of genetics in a single volume.

More advanced treatments of particular topics: T. DOBZHANSKY, *Mankind Evolving: The Evolution of the Human Species* (1962) and *Genetics of the Evolutionary Process* (1970); C. STERN, *Principles of Human Genetics*, 2nd ed. (1960); A.C. STEVENSON *et al.*, *Congenital Malformations* (1966); E.B. FORD, *Ecological Genetics*, 2nd ed. (1971); I.M. LERNER, *The Genetic Basis of Selection* (1958); E. MAYR, *Animal Species and Evolution* (1963); B. WALLACE, *Topics in Population Genetics* (1968).

(T.D.)

Herod I the Great

Herod I the Great, king of Judaea under the Romans and founder of the Herodian house, played a major role in Near Eastern affairs in the 1st century BC, but is probably best known as the tyrant portrayed in the New Testament.

Herod was born in 73 BC in southern Palestine; his father, Antipater, was an Edomite (an Arab from the region between the Dead Sea and the Gulf of Aqaba). Antipater was a man of great influence and wealth, who increased both by marrying the daughter of a noble from Petra (in southwest Jordan), at that time the capital of the rising Nabataean kingdom. Thus Herod was, although a practicing Jew, of Arab origin on both sides.

When Pompey (106–48 BC) invaded Palestine in 63 BC, Antipater supported his campaign and began a long association with Rome, from which both he and Herod were to benefit. Six years later Herod met Mark Antony, whose lifelong friend he was to remain. Julius Caesar also favoured the family; he appointed Antipater procurator of Judaea in 47 BC and conferred on him Roman citizenship, an honour that descended to Herod and his children. Herod made his political debut in the same year, when his father appointed him governor of Galilee. Six years later Mark Antony made him tetrarch of Galilee. In 40 BC the Parthians invaded Palestine, civil war broke out, and Herod was forced to flee to Rome. The senate there nominated him king of Judaea and equipped him with an army to make good his claim. In the year 37 BC, at the age of 36, Herod became unchallenged ruler of Judaea, a position he was to maintain for 32 years. To further solidify his power, he divorced his first wife Doris, sent her and his son away from court, and married Mariame, a Hasmonean princess. Although the union was directed at ending his feud with the Hasmoneans, a priestly family of Jewish leaders, he was deeply in love with Mariame.

King of Judaea

During the conflict between the two triumvirs Octavian

and Antony, the heirs to Caesar's power, Herod supported his friend Antony. He continued to do so even when Antony's mistress, Cleopatra, the queen of Egypt, used her influence with Antony to gain much of Herod's best land. After Antony's final defeat at Actium in 31 BC, he frankly confessed to the victorious Octavian which side he had taken. Octavian, who had met Herod in Rome, knew that he was the one man to rule Palestine as Rome wanted it ruled and confirmed him king. He also restored to Herod the land Cleopatra had taken. Herod became the close friend of Augustus' great minister Marcus Vipsanius Agrippa, after whom one of his grandsons and one of his great-grandsons were named. Both Emperor and Minister paid him state visits, and Herod twice again visited Italy. Augustus gave him the oversight of the Cyprus copper mines, with a half share in the profits. He twice increased Herod's territory, in the years 22 and 20 BC, so that it came to include not only Palestine but parts of what are now the kingdom of Jordan to the east of the river and southern Lebanon and Syria. He had intended to bestow the Nabataean kingdom on Herod as well, but by the time that throne fell vacant. Herod's mental and physical deterioration made it impossible.

Herod endowed his realm with massive fortresses and splendid cities, of which the two greatest were new, and largely pagan, foundations: the port of Caesarea Palaestinae on the coast between Joppa (Jaffa) and Haifa, which was afterward to become the capital of Roman Palestine; and Sebaste on the long-desolate site of ancient Samaria. In Jerusalem he built the fortress of Antonia, portions of which may still be seen beneath the convents on the Via Dolorosa, and a magnificent palace (of which part survives in the citadel). His most grandiose creation was the Temple, which he wholly rebuilt. The great outer court, 35 acres (14 hectares) in extent, is still visible as Al-Ḥaram ash-Sharīf. He also embellished foreign cities: Beirut, Damascus, Antioch, Rhodes, and other towns. Herod patronized the Olympic Games, whose president he became. In his own kingdom he could not give full rein to his love of magnificence, for fear of offending the Pharisees, the leading faction of Judaism, with whom he was always in conflict because they regarded him as a foreigner. Herod undoubtedly saw himself not merely as the patron of grateful pagans but also as the protector of Jewry outside of Palestine, whose Gentile hosts he did all in his power to conciliate.

Mental disorder

Unfortunately, there was a dark and cruel streak in Herod's character that showed itself increasingly as he grew older. His mental instability, moreover, was fed by the intrigue and deception that went on within his own family. Deeply in love with Mariame, he was prone to violent attacks of jealousy; his sister Salome (not to be confused with her great-niece, Herodias' daughter) made good use of his natural suspicions and poisoned his mind against his wife in order to wreck the union. In the end Herod murdered Mariame, her two sons, her brother, her grandfather, and her mother, a woman of the vilest stamp who had often aided Salome's schemes. Besides Doris and Mariame, Herod had eight other wives and had children by six of them. He had 14 children.

In his last years Herod suffered from arteriosclerosis. He had to repress a revolt, became involved in a quarrel with his Nabataean neighbours, and finally lost the favour of Augustus. He was in great pain and in mental and physical disorder. He altered his will three times and finally disinherited and killed his firstborn, Antipater. The slaying, shortly before his death, of the infants of Bethlehem was wholly consistent with the disarray into which he had fallen. After an unsuccessful attempt at suicide, Herod died at Jericho at the end of March or beginning of April in 4 BC. His final testament provided that, subject to Augustus' sanction, his realm would be divided among his sons: Archelaus should be king of Judaea and Samaria, with Philip and Antipas sharing the remainder as tetrarchs.

BIBLIOGRAPHY. FLAVIUS JOSEPHUS, the Jewish historian who was of priestly descent, wrote a detailed and vivid account of Herod and his times in his *Antiquities*, XV, XVI, XVII, 1–8. *Josephus*, available in many English editions, of which the Loeb, 9 vol. (1926–65), is the latest and best, is the foundation for all later work on Herod. A.H.M. JONES, *The Herods of Judaea* (1938), is a scholar's appreciation. The article "Herodes" in *Pauly-Wissowa Real-Encyclopädie*, suppl. vol. 2, pp. 1–199 (1913), is particularly valuable for its complete family tree. STEWART PEROWNE, *The Life and Times of Herod the Great* (1956), is the work of one who knew intimately the topography of all of Herod's many architectural creations. MICHAEL GRANT, *Herod the Great* (1971), is a beautifully illustrated book by a scholar of international standing.

(S.H.P.)

Herodotus

A Greek historian living in the 5th century BC, Herodotus wrote a history of the Greco-Persian Wars that stands as the first great narrative and critical history in the ancient world. No biography of Herodotus has come down from antiquity, nor is it likely that any was written until long after his death. Scattered references have survived from later classical authors, and a short article has survived in a 10th-century Byzantine lexicon called the *Suda*. The most important source on Herodotus' life is the *History* itself.



Herodotus, Roman herm probably copied from a Greek original of the first half of the 4th century BC. In the Museo Archeologico Nazionale, Naples.
By courtesy of the Soprintendenza alle Antichita della Campania, Naples

It is believed that Herodotus was born (perhaps in 484 BC) of a prominent family in Halicarnassus, in Asia Minor. He was exiled from that city by the tyrant Lygdamis but returned later to help overthrow the tyranny and restore constitutional government. Also, it is said that Herodotus was related to the epic poet Panyasis, that he spent a considerable time on the island of Samos, and that eventually, despite his services to his native city, he became so unpopular that he decided to leave Halicarnassus—this time permanently. Later, he joined the new colony sponsored by Athens in Thurii, in south Italy, where he finally died and was buried in the market place. Not all these details are certain, but the general outlines are generally accepted.

It is significant that Herodotus should have been an Asiatic Greek by birth and upbringing and equally significant that he left Asia behind him for Athens and the West, because both parts of his experience contributed to his intellectual development. From Asia and the tradition of the Ionian philosophers, he learned to make use of *historiē*, the new method of scientific inquiry that

consisted first in asking a question, then in looking for information relevant to that question, and, finally, in drawing a conclusion from the data collected. Without his sojourn in Athens and, in general, without the breadth of view he acquired by life as an exile, Herodotus might never have come to ask himself the questions that made him a historian: "Why did the Persian Wars come about, and what deeds were accomplished on both sides that are worthy of being remembered?" The historical work of Herodotus, therefore, is intimately connected with the facts of his personal life, while many of those facts must be deduced from the *History*.

Interest in geography

When he began collecting information, he had not decided to write about the Persian Wars, but he probably had in mind a geographical work something like that of Hecataeus of Miletus. Hecataeus had written a book known as the *Periegesis*, enough of which remains to show that he described the different lands around the Mediterranean sea in clockwise order, beginning with Spain and ending with the north coast of Africa. Even before Hecataeus, the philosopher Anaximander had shown an interest in physical geography, constructing the first Greek world map. No doubt Hecataeus hoped to improve on Anaximander, just as Herodotus set out to improve on the results obtained by Hecataeus. There are a number of passages in the text of the *History* indicating that when Herodotus began to collect his information, he was not thinking about the Persian Wars. For example, he undertook a journey in the Black Sea region without making any effort to cover the ground traversed by Darius in his Scythian campaign—a campaign he was later to describe in great detail; then again, his trips to Tyre and to the island of Thasos were both motivated by his curiosity about the legendary Heracles (Hercules), a theological rather than a historical question.

Best known of all his journeys is the one he made to Egypt. Later on, when he had made up his mind to write about the Persian Wars, he could not bear to leave out the account of his investigations in Egypt, which fill the entire second book of the *History*. He probably went there with a copy of Hecataeus' *Periegesis* in his portmanteau along with an improved version of Anaximander's map. A French scholar, Camille Sourdille, writing in 1910, worked out the generally accepted version of Herodotus' itinerary in Egypt. Herodotus entered Egypt in the western delta at Canopus sometime in August, departing from Pelusium in the eastern delta before December. In between, he visited quite a number of famous places, including Memphis and the site of the great pyramids. He travelled up the Nile all the way to the first cataract, stopping briefly in Thebes. Unfortunately, it is not known what year this occurred, though his reference to the skulls of the Egyptian and Persian dead still lying on the battlefield at Papremis shows he was there some time after 460 BC, when the battle had been fought.

Scientific imagination

More important are the inferences drawn from his remarks, for they show that his primary interest was in the geography and monuments rather than in the history of Egypt. A good example of how his mind worked is his discussion of the causes for the annual Nile flood, in which he criticizes earlier explanations and offers one of his own. But he was more interested in the long-term effects of the annual inundation. Others had noted marine fossils (not only in Egypt) far inland and concluded that land can be created by the silt of rivers; the Egyptian Delta had already been recognized as the work of the Nile. But Herodotus went further. Using such information as he had about the height to which the Nile rose in earlier times when it flooded the land, as compared with the greater height needed to accomplish this purpose in his own day, he tried to calculate how long it took for the silt brought down by the river to build up, not merely the delta but the rest of Egypt as well. Thus, he speculated that if the Nile changed course slightly and emptied into the Red Sea, which was about the size of Egypt, it would create a new land there within 5,000 years. His data were unsatisfactory, but his method was scientific. He did not stop there. In summarizing the political history of Egypt, he put the pyramid kings, who belong back in the Old Kingdom, at a very late date. This has been explained in various ways, but the answer may well be a simple one. His calculations had persuaded him that the pyramids could not have been built so early because the area north of Memphis would still have been under water. He was quite wrong, of course, but magnificently wrong, because he formulated for the first time the relationship existing between man and his environment during the historical period.

Herodotus, however, is not always and completely a rationalist, as is shown in the respect he paid to the oracle of Delphi despite the many indications of rather fallible behaviour to be found in his pages. Somehow, Herodotus manages to reconcile belief in a divinely controlled universe with an equally strong belief in natural causes. To understand the kind of man he was, one must consider his attitude toward the supernatural, for G. De Sanctis, a 20th-century Italian historian, and others found him less skeptical toward the religious views of his day than his predecessor Hecataeus of Miletus. Perhaps the historian's comparatively long stay in Athens modified the radical philosophical views to which he had been exposed earlier in life. Be that as it may, Herodotus often seems to anticipate the efforts of later writers in attempting to reconcile the contradictions between faith and reason. A good example is his account of Croesus of Lydia. Croesus is doomed by the fact that the founder of his line, Gyges, usurped the throne; the Delphic oracle had predicted the overthrow of Gyges' succession after five generations, but this prediction had been forgotten by Croesus' time. Then there is another irrational element in the story. The acropolis of Sardis was to be made impregnable by having a lion cub carried around the perimeter. This was done except for a small section, where access was so difficult that no enemy could be imagined foolhardy enough to try that side. And that, of course, is how the Persians got in. Croesus, then, is in a hopeless situation, his fate has been decreed. But while Herodotus preserves these supernatural explanations, he is not satisfied. He offers a rational explanation side by side with the irrational. After fighting a drawn battle with Cyrus, Croesus withdraws to Sardis and dismisses his allies on the mistaken assumption that Cyrus also will retire. But Cyrus did no such thing, and when he suddenly appeared before the walls of Sardis, it was too late for Croesus to recall his allies. Did Sardis fall as the result of divine vengeance for the sins of Gyges or because Croesus made a mistake in judging his opponent?

Ambivalence of reason and faith

Another illustration of Herodotus' ambivalence is his use of dreams. Many of the dreams cited in the *History* have a prophetic character, such as Astyages' dream about his daughter Mandane or Cyrus' dream presaging the future greatness of Darius. Both prophecies are fulfilled despite determined efforts to thwart them. But the dream of Xerxes is somewhat different. The king is accosted by a figure who appears to him as he sleeps and threatens him with dire consequences if he persists in calling off the invasion of Greece. Xerxes promptly consults his uncle Artabanus, who sets his mind at ease. He tells his nephew that dreams can have natural causes, that men tend to dream at night about matters that have been occupying their minds during waking hours, and that there is nothing supernatural in such nocturnal manifestations. But, in the end, Artabanus is forced to withdraw his opposition to the war *by a dream*. Herodotus leaves it to the reader to decide whether the gods have determined on the defeat of Xerxes or whether Xerxes was a fool to allow his good sense to be overruled by a phantom.

Debt to Periclean Athens

It may well have been in Athens that Herodotus decided to write about the Persian Wars. Having made up his mind, he worked up the ethnographic, geographic, and mythological materials he had collected into a new framework, in which all past history leads up to a struggle between East and West, culminating with the invasion of Xerxes in 480 BC. In the process he also makes a clear distinction between history and prehistory, the dividing line for Greece coming at about 600 BC. Rules

of evidence for legendary figures, such as Minos of Crete or the heroes at Troy, are different than those applied to historical personages, such as Croesus or Peisistratus. Herodotus' intellectual development owed much to Periclean Athens, where he was able to read parts of his unpublished *History* to appreciative audiences. Sophocles is said to have been one of his friends, the plays of Aeschylus seem to have influenced him, and among his listeners—perhaps—was the young Thucydides.

Herodotus was unique in the ancient world. Other famous historians, such as Thucydides, Xenophon, and Polybius, great though they were, lacked his qualities of cynicism without despair and humour without bitterness. Above all, he alone was a truly Greek historian, not an Athenian or an Achaean but a Greek, whose birthplace was soon forgotten and had to be redetermined by later scholars. The first historian set an example of universality that few if any of his successors have been able to match.

BIBLIOGRAPHY. FELIX JACOBY, "Herodotos" in *Pauly-Wissowa Real-Encyclopädie*, suppl. 2, pp. 205–520 (1913), the most thorough discussion of Herodotus' life and work (in German); JOHN L. MYRES, *Herodotus: Father of History* (1953), an expert appraisal of Herodotus as a literary artist; P.E. LEGRAND, *Hérodote: Introduction*, 2nd ed. (1955), a readable and illuminating interpretation of the life and character of Herodotus (in French); GIETANO DE SANCTIS, *Studi di Storia della Storiografia greca*, pp. 1–71 (1951), a provocative discussion of how Herodotus put his *History* together, what his purpose was, and the rationalism of his predecessor Hecataeus; T.S. BROWN, "Herodotus Speculates About Egypt," *Am. J. Philology*, 86:60–76 (1965).

(T.S.B.)

Hero Worship

Hero worship is the cult of real or imaginary persons of the past at their tombs or relics or at other cult places. This narrow and strict definition has, however, often been broadened to include such phenomena as honour and veneration, as well as worship proper.

Heroes and heroic literature. The term hero—derived from the Greek noun *hērōs*, the etymology of which is uncertain—has, in the course of time, acquired a variety of meanings: the main figure in a literary work; one who accomplishes remarkable deeds or whose life exhibits the qualities of extraordinary courage, valour, and fortitude; the main character in a tale or epic of the kind usually known as heroic literature; or a person worthy of veneration and honour. Because the use of the noun hero and the adjective heroic have been determined by their Greek origins, it may be useful to survey the terms briefly before considering their application by extension to other cultures and cultural phenomena.

Homer and Hesiod

In the Homeric poems the term hero signifies any freeman of that early age described in the *Iliad* and the *Odyssey*. It is applied more specifically to the outstanding characters in those epics: superior beings who excelled in war and adventure and who prized virtues and values such as courage and loyalty. It is difficult to generalize about heroes and the heroic even in Greek culture because there is considerable variety in the treatment of the subject. The Homeric heroes are frequently in touch with the gods who intervene in their affairs and determine their destinies. This is not the case with most of the heroes of the Greek tragedians. Some heroes are of divine parenthood on one side; this semidivine origin serves to account for the supernatural powers of many heroes. Altogether the supernatural plays a prominent role in many heroic tales (*e.g.*, magic weapons, oracles, encounters with powerful sorcerers or witches, battles with dragons and monsters). Furthermore, the hero's exploits are not necessarily confined to this world; his adventures may lead him to the underworld (the realm of the dead) or to the realm of the gods.

The idea that heroes lived in a bygone period, to which the teller of the heroic stories is looking back, gave rise to the notion of a heroic age. This notion is found in Hesiod's *Works and Days*, in which a race of heroes intervenes between the third and the fourth of the four periods into which the history of the world is divided. Very possibly Hesiod's account is an attempt to harmonize an ancient mythological tradition regarding four cosmic ages with the epic tradition of the poets that dealt with the heroic exploits of heroic beings in a heroic age. Something like a heroic age is also assumed by the epics and tales of other cultures.

The stories or cycles of stories describing the exploits of the heroes and great figures of the heroic age are usually subsumed under the heading heroic literature. This type of literature poses a great many problems, both as regards form (poetry, prose, epic) and literary and social history. Among the many questions that are still being discussed are the relation of heroic written literature to earlier oral traditions; the relation of the longer compositions to shorter units; the social background of this literary genre; and the historical realities reflected by the heroic traditions. Some specimens of heroic literature are demonstrably late or are deliberate imitations of earlier models: Virgil's *Aeneid* uses the Homeric example to give an imaginary account of the origin of Rome, in keeping with the new Roman ideology that emerged during the reign of Augustus (27 BC–AD 14). According to some scholars, the heroic poem reflects a phase of culture and a social order dominated by a powerful warrior class of a feudal type. The warrior's heroic exploits were celebrated in songs and lays (ballads) transmitted orally by wandering bards. According to this view, the heroic lay is the literary successor to the religious hymns and myths whose subjects were gods rather than human heroes; it can claim some basis of historical truth, even if this historical nucleus was transformed by and overlaid with imaginative and legendary elements.

Two preliminary questions should be mentioned briefly. The one concerns the possibility and the limitations of a large-scale comparative analysis of the heroic literatures of diverse cultures. Whether, for example, Greek and Roman epic or Old High German (*e.g.*, *Hildebrandslied*) and Middle High German (*e.g.*, *Nibelungenlied*) heroic sagas can be usefully compared and to what purpose is a moot point. The other concerns the fact that heroic literature exists at different levels, referred to by a variety of technical terms (myth, saga, legend, folktale), but there is no unanimity in the use of this terminology.

Relationship of heroes and gods

Nature and sacral status of the hero. The superhuman qualities of many heroes poses the question of their relationship to the gods, especially as the distinction between *Göttermythologie* (myths of gods) and *Heldensage* (heroic saga) is not so neat and rigid as suggested by some scholars. The Greek mythographer Euhemerus of Messina, who lived about 300 BC, maintained that the gods that were the objects of popular worship had originally been great conquerors, heroes, or sages subsequently venerated by posterity. Euhemerus' theory, although inadequate as an explanation of the myths and worship of all gods, has at least the merit of drawing attention to the fact of the veneration or even deification of the illustrious departed. This links hero worship with ancestor worship on the one hand and such phenomena as the deification (Greek *apotheōsis*) and cult of the emperors at Rome on the other. The cult of many lesser or minor gods, unlike that of the great deities, has often been accounted for on lines similar to Euhemerus' theory; even here, however, the problem is more complicated. Many scholars, although agreeing that the cult of ancestors and that of heroes were somehow related, have differed as to the nature of the connection. Ancestors—*i.e.*, powerful departed souls whose hostility could be dangerous, whose favour could be beneficial, and who should therefore be propitiated—enjoyed a cult in the narrow circle of their family and descendants. A hero would be honoured by a wider circle than the family—*e.g.*, by a city that he had founded or in which his remains had been buried. Some scholars have believed that the cult of the dead derived from that of heroes, whereas others have thought that the cult of kings led to that of illustrious ancestors, thence to that of the illustrious dead in general, and, finally, to the worship of

purely legendary figures. The question became one of the divine or human origin of heroes; of whether they were the product of historic reality or of the imagination. Frequently the problem has been posed in terms of a clear-cut dichotomy: heroes are deified (or semideified) men, and their cult is that of the dead elevated to divine rank; or, conversely, they are "faded" gods, and their cult is what was left after their demotion from the status of full divinity.

Another approach, relating the hero to cult and ritual, views heroes not as degraded gods but as personifications of ancient and at times forgotten rituals. This theory, popularized in its more extreme form by Lord Raglan in his *Hero* in 1936, holds that all traditional narratives, including myths and heroic legends, originate in ritual. Tales such as those of Troy or King Arthur are connected with ancient rites that have fallen out of use although the narrative has survived. This explanation of heroes in terms of ritual drama is the very opposite of hero worship as understood by the definition given at the beginning of this article.

Varieties of cultic forms. *Ancient Greece.* Hero worship is well attested in ancient Greece, though the actual facts may not always accord with the accounts and interpretations offered by the later Greek authors. The association of hero worship with the cult of the dead is confirmed by a number of technical ritual details. Thus, cults connected with heroes have a funerary and chthonic, or "underworld," character, unlike the rites offered to the heavenly gods. The manner of slaughtering the sacrificial animal is different, and so is the word itself for sacrificing (*thuein* for the gods and *enagizein* for hero cults). The sacrifice for heroes and for the dead is not offered on a "high altar"; the low, round, and hollow altar is constructed so that the blood flows away into the earth. The whole carcass is burned, whereas a sacrifice to the gods is shared between them and the humans. On the other hand, these differences must not be pressed too much. Some Greek hero cults have no funerary character. There are gods with surnames of heroes. A temple precinct is defiled by the presence of a tomb but not by the tomb of a hero, and many traditions speak of the burial of heroes in sanctuaries. Many heroes prove their divine origin by the rituals or virtues associated with them. The patrons of rain, fertility, and healing cults are, more often than not, archaic spirits with traditional cults that were subsequently personified, given an individual biography, or attached to the names and biographies of existing (legendary or historical) heroes. Heroic legends are often deliberately etiological in intent; *i.e.*, they account for the origin and details of cultic centres, practices, and liturgies. Personification may also lead to identification in the course of literary traditions: Iphigeneia of Brauron, who protects women in labour, and the Iphigeneia sacrificed at Aulis by her father, Agamemnon, leader of the Greeks in the Trojan War, were surely not the same originally, though they were later identified.

Types of Greek heroes

It appears that heroes are as heterogeneous in their origins as in their developed qualities, and thus it is impossible to generalize about their cults. The English scholar L.R. Farnell—in his work *Greek Hero Cults and Ideas of Immortality*, published in 1920—distinguished seven types of hero: priestly hero-gods of cultic origin (*e.g.*, Trophonius, an architect in Greek legend whose oracle—which was consulted only by a means of peculiar ceremonies—was at the spot where he was swallowed up by the earth); sacral heroes or heroines associated with a god (*e.g.*, Iphigeneia); secular figures who became fully divinized (*e.g.*, Heracles); epic heroes like Hector and Achilles in Homer's *Iliad;* fictitious genealogical heroes like Ion, who was constructed to account for the origin of the Ionian people of ancient Greece; functional and cultural demons—often anonymous and generally of local importance only; and, occasionally, real men who were subsequently granted cults. The classification is far from satisfactory, yet it helps to show the variety of types even within one culture and to emphasize the fact that not all heroic types are connected with worship. Thus, the heroes of Homer, although sufficiently close to the gods to have the latter intervene in their affairs, are not described as objects of a cult. On the other hand, the Greeks themselves developed ideas and legends associating heroic tales with cultic data. This pattern easily converged with the habit of considering the illustrious dead as a source of power and a potent force to be used for the benefit of the community. Hence, the relationship to the tomb, interred body, or relics as a centre from which radiated a certain power also developed. During the war between Sparta and Arcadia in the middle of the 6th century BC, the Spartans were told that they would be victorious if they could possess themselves of the body of Orestes, the hero who avenged the death of his father Agamemnon by killing his mother and her lover. Founders and colonizers of cities receive heroic honours in approved Greek fashion (*i.e.*, animal sacrifices, athletic games, paeans). Pausanias, a Greek traveller in the 2nd century AD, still saw the tomb of Leonidas, the king of Sparta, as the site of a cult. Then, as now, popular sports champions would also be "idolized." Euthymus of Locri won a boxing contest in Olympia in 484 BC; six centuries later Pausanias still saw his statue and could even relate a heroic legend to the effect that Euthymus had freed the city of Temesa from a dreadful ghost and ultimately gained "a shrine and a cult."

The concept of the hero and, thus also, that of hero worship was developed and transformed by the Greek poets, who increasingly moralized it. In the tragedies of Euripides, heroic honours are the reward of moral excellence; and these honours are mainly the institution of a cult. In fact, the heroic life epitomizes certain values without which life is not worth living. The hero prefers death to an unworthy life and accepts suffering, trials, and tribulations as long as these are compatible with his conception of fortitude and virtue. According to Plato, the struggles and sufferings of the heroic life are due to man's desire for immortal glory. The ideal hero fears disgrace only. Because the heroic life usually ends with a heroic death, the hero is, almost by definition, a tragic hero. It is this tragic quality that, among other things, distinguishes the heroes of the great epics from those of folktales.

Non-Greek traditions

Primitive religions. The use of the terms hero and hero worship in cultural contexts other than the Greek tradition is somewhat problematic. The chief characters in many primitive myths cannot be described as heroes. Sometimes they are mythical figures (divine, semidivine, human, or animal) whose activities account for the origins of certain cosmic facts or cultural achievements or social institutions; they are often described as culture heroes, but neither are they heroes in the classical Greek sense nor do they enjoy cults and worship. In some instances the mythical ancestors are deemed to return on certain ritual occasions, but it may be doubted whether the term hero worship is applicable here. Even when the myth tells of the origin of things and institutions through the death of a primordial mythical being, the terms hero and hero worship are, strictly speaking, inappropriate.

Shamanism and Chinese and Japanese religions. In some forms of religion, illustrious or powerful individuals may be identified after their death with powerful spirits or promoted to divine rank. Thus, many of the mighty spirits in some forms of Asian shamanism (a religion characterized by belief in a world of unseen gods, demons, and ancestral spirits responsive only to the priests, called shamans) are identified with departed generals and warlords. In Chinese and Japanese religions, the differences between men, spirit powers, and gods are a matter of degree only; and thus worship of the ancestors and the illustrious departed easily merges with spirit cults and regular divine worship. Among the thousands of gods or divine beings, some are the personifications of natural forces, others are mythological deities, and others again are deified kings, heroes, scholars, and benefactors. Because the celestial hierarchy is similar to that on earth, deserving individuals could be elevated to divine rank by imperial decree. One of the best known examples

is the Chinese hero Kuan Ti, also known as Kuan Yü or Wu Ti, who after his death in the 3rd century AD advanced successively—throughout the T'ang, Sung, Ming, and Ch'ing dynasties—to ever higher honours. Thus, in 1594 he received the title "Faithful and Loyal Great Ti, God of War." He is said to have had 1,600 state temples and thousands of local shrines in modern (pre-Communist) China. The 9th-century statesman and scholar Sugawara Michizane (also known as Kanko) may be cited as a Japanese example.

Christianity, Islām, and Buddhism. The Christian cult of saints, though shaped by a specifically Christian ideal of holiness, owes much to Greek forms of hero worship. The original Christian hero is the martyr, and his cult is centred on his tomb and his relics. On the more popular level, saints are considered to be a source radiating power and blessing; in more strictly Christian terms, they are an inspiring reminder of Christian life and a concrete point of contact with the person of the martyr who—somehow still connected with his tomb or relic, yet at the same time in heaven—can act as an intercessor with God. Later, the place of martyrs was taken by confessors (*i.e.*, those who professed their faith but were not required to die a martyr's death) and by saints whose lives exhibited heroic virtue in acts of penance, ascetic mortification, and charity. Whereas the souls of the ordinary departed are prayed for, those of the saints are prayed to. The cult of saints, and especially of tombs of holy men, is also known in popular Islām, though the orthodox disapprove of the practice as heartily as Protestants disapprove of the Catholic attitudes toward saints. The saint as a hero meriting worship is known also in Eastern religions. In Buddhism the ascetic and renouncer is the real victor and hero, for it is he that has overcome the world; the Buddha is the real *cakravartin* (world conqueror).

Modern forms. In addition to the narrow definition of hero worship given at the beginning of this article, the term is also used in a wider sense to refer to attitudes of reverence, honour, and admiration—at times enthusiastic—toward individuals who have excelled in the virtues applauded in a given society. These may range from the most barbarous to the most refined or spiritualized. Heroes can be warriors, rulers, prodigies of sexual prowess, world redeemers, and saints. The study of heroes and of later interpretations of traditional heroic figures can provide a clue to the dominant values—and the changes in the values—of a group. Hiawatha, a legendary chief of the Onondaga tribe of North American Indians, functioned as the Iroquois symbol of civilization and human progress. Perhaps the most striking recent example of a hero cult is that which formed around "Che" Guevara—the Cuban guerrilla leader—who became a kind of mythical symbol with which the revolutionary youth of the radical left wished to identify in ideal as well as in action.

Significance. Some writers, analyzing and comparing the types of heroes described in the various heroic traditions of the world, celebrated in legends, or venerated in cults, claim to find certain regularities or structures characteristic of the hero. These are said to exhibit a basic pattern of the heroes' biographies (*e.g.*, supernatural or marvellous birth, endangered childhood, prodigious feats in manhood, successful overcoming of dangers and trials, acquisition of a hidden treasure, winning a bride or liberating a captive damsel, heroic death) and to express something essential to man's understanding of himself. If all important truths present themselves under the guise of the figures of religion and mythology, then the figure of the hero signifies man's aspiration toward a fuller development of his spirit and transcendence of his incomplete state. The hero, according to this view, symbolizes man's urge and struggle to transcend the limitations of his existence and to conquer a fuller and more total life. The heroic adventures are thus seen as symbolic representations of an essentially spiritual quest.

BIBLIOGRAPHY. THOMAS CARLYLE, *On Heroes, Hero-Worship, and the Heroic in History* (1841, reprinted 1966), popularized the notion of hero worship and contributed to the modern interest in the subject. LORD RAGLAN, *The Hero* (1936), argues the thesis that the heroes of legend and myth are not historical figures but personifications of ancient rituals. JOSEPH CAMPBELL, *The Hero with a Thousand Faces*, 2nd ed. (1968), attempts to bring out the basic problem underlying the many forms of the heroic and to interpret these as a symbol of man's spiritual quest. The classical discussions of hero worship in ancient Greece are F. DENEKEN's article "Heros" in W.H. ROSCHER (ed.), *Ausführliches Lexicon der griechischen und römischen Mythologie* (1886–90); ERWIN ROHDE, *Psyche*, 8th ed. (1921; Eng. trans., 1925); EDUARD MEYER, *Geschichte der Alterthums*, vol. 2 (1893); and FRIEDRICH PFISTER, *Der Reliquienkult im Altertum*, 2 vol. (1909–12), who also discusses the work of his predecessors. The best known English work on the subject is L.R. FARNELL, *Greek Hero Cults and Ideas of Immortality* (1921, reprinted 1970). MARIE DELCOURT, *Légendes et cultes de héros en Grèce* (1942), is an excellent brief survey of the question. The transition from Greek to Christian hero worship is discussed by HIPPOLYTE DELEHAYE, *Les Origines du culte des martyrs*, 2nd ed. rev. (1933); and A.J. FESTUGIERE, *La Sainteté* (1942), which also has a good chapter on the moral aspects of the Greek hero. Problems of heroic literature are discussed in C.M. BOWRA, *Heroic Poetry* (1952).

(R.J.Z.W.)

Herschel Family

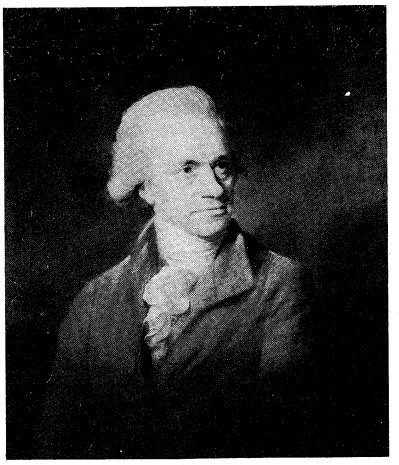
Members of the Herschel family transformed the science of astronomy by their indefatigable studies of the stars and nebulae. Sir William, who became famous for the discovery of Uranus, built the largest telescopes of his day and, in collaboration with his sister, Caroline Lucretia, developed pioneer theories of cosmology. His son, Sir John, revised and completed his father's work and became a prominent figure in British science. The Herschels applied the bold concept that vast changes occurred in the universe in the course of time, explaining in this way the differences to be observed among stars and nebulae.

WILLIAM AND CAROLINE HERSCHEL

Frederick William (Friedrich Wilhelm) Herschel was born on November 15, 1738, in Hanover, Germany, where his father was an army musician. Following the same profession, the boy William played in the band of the Hanoverian Guards. After the French occupation of Hanover in 1757, he escaped to England, where at first he earned a living by copying music. But he steadily improved his position by becoming a music teacher, performer, and composer, until in 1766 he was appointed organist of a fashionable chapel in Bath, the well-known spa.

Telescope construction. By this time, the intellectual curiosity he had acquired from his father led him from the practice to the theory of music, which he studied in Robert Smith's *Harmonics*. From this book he turned to Smith's *A Compleat System of Opticks*, which introduced him to the techniques of telescope construction and whetted his appetite for viewing the night sky. Combining obstinacy with boundless energy, William was not content to observe the nearby Sun, Moon, and planets, as did nearly all astronomers of his day, but was determined to study the distant celestial bodies as well, and he realized he would need telescopes with large mirrors to collect enough light, larger, in fact, than opticians could supply at reasonable cost. He was soon forced to grind his own mirrors. They were ground from metal disks of copper, tin, and antimony in various proportions. In 1781 his ambitions outran the capacities of the local foundries, and so he prepared to cast molten metal into disks in the basement of his own home; but the first mirror cracked on cooling, and on the second attempt the metal ran out onto the flagstones, after which even he accepted temporary defeat. His later and more successful attempts produced ever-larger mirrors of superb quality—his telescopes proved far superior even to those used at the Greenwich Observatory. He also made his own eyepieces, the strongest with a magnifying power of 6,450 times.

At Bath, he was helped in his researches by his brother Alexander, who had come from Hanover, and his sister, Caroline, who was his faithful assistant through much





(Left) William Herschel, oil painting by L. Abbott, 1785. (Centre) Caroline Herschel, engraving by Joseph Brown, 1847. (Right) John Herschel, pencil drawing by H.W. Pickersgill (1782–1875). (Left, right) In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

of his career and an excellent astronomer in her own right. News of this extraordinary household began to spread in scientific circles. He made two preliminary telescopic surveys of the heavens. Then, in 1781, during his third and most complete survey of the night sky, William came upon an object that he realized was not an ordinary star.

Discovery of Uranus. It proved to be the planet Uranus, the first planet to be discovered since prehistoric times. William became famous almost overnight. His friend Dr. William Watson, Jr., introduced him to the Royal Society of London, which awarded him the Copley Medal for the discovery of Uranus, and elected him a Fellow. Watson also helped him to secure in 1782 an annual pension of £200 from George III; Sir Joseph Banks, in 1787, helped to secure an annual pension of £50 for Caroline. He could thus give up music and devote himself exclusively to astronomy. At this time William was appointed as an astronomer to George III, and the Herschels moved to Datchet, near Windsor Castle.

Although he was 43 years old when he became a professional astronomer, William worked night after night to develop a "natural history" of the heavens. A fundamental problem for which Herschel's big telescopes were ideally suited concerned the nature of nebulae, which appear as luminous patches in the sky. Some astronomers thought they were nothing more than clusters of innumerable stars the light of which blends to form a milky appearance. Others held that some nebulae were composed of a luminous fluid. When William's interest in nebulae developed in the winter of 1781–82, he quickly found that his most powerful telescope could resolve into stars several nebulae that appeared "milky" to less well equipped observers. He was convinced that other nebulae would eventually be resolved into individual stars with more powerful instruments. This encouraged him to argue in 1784 and 1785 that all nebulae were formed of stars and that there was no need to postulate the existence of a mysterious luminous fluid to explain the observed facts. Nebulae that could not yet be resolved must be very distant systems, he maintained; and, since they seem large to the observer, their true size must indeed be vast—possibly larger even than the star system of which the Sun is a member. By this reasoning, William was led to postulate the existence of what later were called "island universes" of stars.

William's early work on nebulae

Theory of the evolution of stars. In order to interpret the differences between these star clusters, it was natural for William to emphasize their relative densities, which he did by contrasting a cluster of tightly packed stars with others in which the stars were widely scattered. These formations showed that attractive forces were at work: with the passage of time, he maintained, widely scattered stars would no doubt condense into one or more tightly packed clusters. In other words, a group of widely scattered stars was at an earlier stage of its development than one whose stars were tightly packed. Thus, William made change in time, or evolution, a fundamental explanatory concept in astronomy. In 1785 he developed a cosmogony—a theory concerning the origin of the universe: the stars originally were scattered throughout infinite space, in which attractive forces gradually organized them into even more fragmented and tightly packed clusters. Turning then to the system of stars of which the Sun is part, he sought to determine its shape on the basis of two assumptions: (1) that with his telescope he could see all the stars in our system, and (2) that within the system the stars are regularly spread out. Both of these assumptions he subsequently had to abandon. But in his studies he gave the first major example of the usefulness of stellar statistics in that he could count the stars and interpret this data in terms of the extent in space of the Galaxy's star system. Other astronomers, cut off from the evidence by the modest size of their telescopes and unwilling to follow William in his bold theorizing, could only look on with varying degrees of sympathy or skepticism.

In 1787 the Herschels moved to Old Windsor, and the following year to nearby Slough, where William spent the rest of his life. Night after night, whenever the Moon and weather permitted, he observed the sky in the company of Caroline, who recorded his observations. Caroline, on her own, discovered eight comets in the years 1786 to 1797. On overcast nights, William would post a watchman to summon him if the clouds should break. Often in the daytime, Caroline would summarize the results of their work while he directed the construction of telescopes, many of which he sold to supplement their income. His largest instrument, too cumbersome for regular use, had a mirror made of speculum metal, with a diameter of 48 inches and a focal length of 40 feet. Completed in 1789, it became one of the technical wonders of the 18th century.

William's achievement, in a field in which he became a professional only in middle life, was made possible by his own total dedication and the selfless support of Caroline. He seems not to have considered the possibility of marriage until after the death in 1786 of a friend and neighbour, John Pitt, whose widow, Mary, was a charming and pleasant woman. Before long, William proposed marriage; he and Mary would live in the Pitt house, while Caroline would remain at Observatory House in Slough. But Mrs. Pitt was shrewd enough to realize that William's commitment would be to Observatory House, which they made their principal home after their marriage on May 8, 1788. William and Caroline continued their labour in astronomy, but as the rigours of observing took their toll of William's health, he came to appreciate more and more the comforts that Mary's

sensible management brought to his home. It was some time before Caroline was reconciled to being displaced in William's affections, after so many years as his constant companion and assistant.

Theory of the structure of nebulae. William's grand concept of stellar organization received a jolt on November 13, 1790, when he observed a remarkable nebula, which he was forced to interpret as a central star surrounded by a cloud of "luminous fluid." This discovery contradicted his earlier views. Hitherto William had reasoned that many nebulae that he was unable to resolve (separate into distinct stars), even with his best telescopes, might be distant "island universes" (such objects are now known as galaxies). He was able, however, to adapt his earlier theory to this new evidence by concluding that the central star he had observed was condensing out of the surrounding cloud under the forces of gravity. In 1811 he extended his cosmogony backward in time to the stage when stars had not yet begun to form out of the fluid.

This example of William's theorizing is typical of his thinking: an unrivalled wealth of observations interpreted by means of bold though vulnerable assumptions. For example, in dealing with the structural organization of the heavens, he assumed that all stars were equally bright, so that differences in apparent brightness are an index only of differences in distances. Throughout his career he stubbornly refused to acknowledge the accumulating evidence that contradicted this assumption.

William's star catalogs

Herschel's labours through 20 years of systematic sweeps for nebulae (1783–1802) resulted in three catalogs listing 2,500 nebulae and star clusters that he substituted for the 100 or so milky patches previously known. He also cataloged 848 double stars—pairs of stars that appear close together in space, and measurements of the comparative brightness of stars. He observed that double stars did not occur by chance as a result of random scattering of stars in space but that they actually revolved about each other. His 70 published papers include not only studies of the motion of the solar system through space and the announcement in 1800 of the discovery of infrared rays but also a succession of detailed investigations of the planets and other members of the solar system. He was knighted in 1816 for his outstanding contributions to British science.

JOHN HERSCHEL

John Frederick William, the only child of William and Mary Herschel, was born on March 7, 1792. John, a frail child, who struggled against ill health all his life, was to become another object of Caroline's love. When he was eight, he was sent to school at nearby Eton, but his mother, seeing him involved in a fight, took him away to be educated privately in less robust surroundings.

In 1809 John entered Cambridge University, where he gained all the advantages of education and friendship that his father had missed in his youth. He joined with Charles Babbage, the mathematician and inventor of the computer, and George Peacock, also a mathematician and later a theologian, in resolving "to do their best to leave the world wiser than they found it." In 1812 they founded the Analytical Society of Cambridge in order to introduce continental methods of doing the mathematical calculus into English practice. They did so by replacing the cumbersome symbolism of Newton with the more efficient type invented by the German philosopher and mathematician Gottfried Wilhelm Leibniz. John's exceptional abilities were quickly recognized: in 1812 he submitted his first mathematical paper to the Royal Society, for which he was elected a fellow in the following year. In 1813 he earned first place in the university mathematical examinations.

Later that year he asked his father's advice about a career. William recommended the church, for "a clergyman is in possession of dignified leisure" to cultivate his mind. But John was attracted to the law and began to study for the bar in London. He soon realized that he had made a wrong choice. Taking advantage of his residence in London, he established contact with his friends in the scientific community, including the astronomer James South. In 1815 he applied for the professorship of chemistry at Cambridge but lost the election by only one vote. His final break with the legal profession occurred that summer when he became seriously ill; after convalescence he returned to Cambridge as a mathematics teacher. But, because his famous father, in spite of failing health, stubbornly continued his research, John felt a filial duty to join him as his assistant. "My heart dies within me," John wrote Babbage in 1816 when he left Cambridge.

In later years, John's apprenticeship in astronomy did not prevent him from making important contributions to chemistry and the physics of light and particularly to mathematics, for which he was awarded the Copley Medal of the Royal Society in 1821. But, through his work with his father, he gained the full benefit of the aged astronomer's unrivalled experience in the construction and use of large telescopes. This apprenticeship laid the foundation of John's subsequent achievements. In 1820 he was among the founders of the Royal Astronomical Society.

Observation of double stars. John Herschel's first major task in astronomy was the re-observation of the double stars that were cataloged by his father. The movements of these pairs of stars about each other offered the best hope of investigating the gravitational forces operating in the universe. John was fortunate to find in James South a collaborator who was able to afford the refined instruments best suited for this work. The catalog that they compiled between 1821 and 1823 and that they published in the *Philosophical Transactions* in 1824 earned them the Gold Medal of the Royal Astronomical Society and the Lalande Prize in 1825 from the Paris Academy of Sciences. This publication was their only joint undertaking.

In the 1820s John also travelled extensively in Europe, visiting foreign scientists; he was abroad when William's life came peacefully to a close in Slough, on August 25, 1822. John was secretary of the Royal Society in 1824–27. In 1825 he turned his attention to the nebulae his father had studied so closely, writing to Caroline that "The curious objects . . . I shall now take into my especial charge—nobody else can see them." His studies of double stars led him in 1826 to the very important problem of measuring the "parallax of the fixed stars," which is the apparent shift in angular position of a given star as viewed together with its neighbour or against the background of the sky; a shift caused by the Earth's revolution around the Sun. Stellar parallax is therefore empirical proof that the Earth does revolve. John's estimation of annual variations of position angles for many stars was an important step in the resolution of this problem.

Caroline's last years

After William's death, Caroline moved back to Hanover for the last 25 years of her life. During her retirement she prepared a catalog of the 2,500 nebulae and star clusters her brother had discovered. For this accomplishment she was awarded in 1828 the Gold Medal of the Royal Astronomical Society.

On March 3, 1829, John married Margaret Brodie, the daughter of a Presbyterian minister; all who knew them agreed that theirs was an idyllic union. In 1831 he was knighted.

Voyage to the Southern Hemisphere. His sense of obligation to complete his father's work in astronomy led him to consider a journey to the Southern Hemisphere to survey the skies not visible in England, but an extended stay would be necessary, and he was reluctant to leave his aging mother for so long. When she died in January 1832, John quickly began planning his expedition. The revision and extension of his father's catalogs, which he carried out at Observatory House, beginning in 1825, was brought to completion and published in 1833. In November of that year, John and his family set sail for the Cape of Good Hope with a large reflecting telescope for observing faint nebulae, similar in size to William's favourite instrument. It had three interchangeable mirrors, one made by William, one by John under William's

directions, and the third by John alone, all with 20-foot (six-metre) focal lengths. He also had a refracting telescope of seven-foot (two-metre) focal length for observing double stars.

The family established their home at Feldhausen, a Dutch farmhouse six miles southeast of Cape Town. John spent four years of intense scientific activity, the clear southern skies allowing much more rapid progress in observing than was possible in England. When the family embarked for home in March of 1838, John had recorded the locations of 68,948 stars, amassed long catalogs of nebulae and double stars, described many details of the Great Nebula in the constellation Orion, as well as the Magellanic Clouds—actually two galaxies visible only in the Southern Hemisphere—and observed Halley's Comet and the satellites of Saturn. In addition, his descriptions of sunspot activities and his measuring of solar radiation by means of a device he had invented contributed to the development of systematic studies of the Sun as an important part of astrophysics. But preparing his results for publication took longer; on his return he was made a baronet, was lionized by the scientific world, and was obliged to accept time-consuming official duties. The year after returning from Africa, he heard about the remarkable invention of the French photographer Louis-Jacques-Mandé Daguerre to produce pictures on light-sensitive paper. Seeing at once the technical possibilities of photochemistry, he published important studies of his own on this new subject and succeeded in preparing a photograph of a solar spectrum in colour.

In 1840 the Herschels bought a spacious house in Collingwood, Kent, and there John laboured at his desk, taxing his health with additional commitments, which included the composing of the *Outlines of Astronomy* for educated laymen (1849). Growing from an earlier book, this highly successful science text went through many editions, including Arabic and Chinese. He maintained his wide scientific interests, especially in the chemistry of photography in which he conducted many experiments. But the bulk of his time was occupied with the *Results of Astronomical Observations, Made during the Years 1834–38 at the Cape of Good Hope*, which he published in 1847. His aunt Caroline died in Hanover on January 9, 1848.

John's public career

Inherited wealth and a serene family life contributed to his success in his life's work. But suddenly he began to seek appointment to public office. Whatever the reasons, the change cost him dear. At the end of 1850 he was appointed master of the Mint, and, although his integrity, humanity, and intelligence made him an obvious choice for the post, he was not sufficiently robust to cope with the strain of administering a large organization. Furthermore, he continued to undertake even more burdensome tasks, especially the plans for the scientific and technical sections for the Great Exhibition of 1851. He was forced to live in London away from his family and to work all hours of the day and night. His health deteriorated, he became depressed, and in 1854 he suffered a nervous breakdown.

The following year he was incapacitated for several weeks, and in 1856 he resigned his post at the Mint. He spent his remaining years with his family at home in Kent, working on the catalogs of double stars and of nebulae and star clusters, but in the face of ever-deteriorating health. He died at Collingwood on May 11, 1871.

BIBLIOGRAPHY. Two surveys of William Herschel's life and work are J.B. SIDGWICK'S highly readable *William Herschel: Explorer of the Heavens* (1953); and A. ARMITAGE'S more systematic *William Herschel* (1962). A charming picture of his personality, with extensive quotations from manuscripts, is given in *The Herschel Chronicle: The Life-Story of William Herschel and His Sister Caroline Herschel* (1933), ed. by his granddaughter, CONSTANCE A. LUBBOCK. His complete published papers are reprinted in *The Scientific Papers of Sir William Herschel*, ed. by J.L.E. DREYER, 2 vol. (1912); and the most significant are reprinted and critically analyzed by M.A. HOSKIN in *William Herschel and the Construction of the Heavens* (1963). The standard biography of John Herschel is *The Shadow of the Telescope* by GUNTHER BUTTMANN (1970), which includes a bibliography. *Herschel at the Cape* (1969), is a pedantic edition of his diaries and letters, 1834–38.

(M.A.H.)

Herzen, Aleksandr

Aleksandr Ivanovich Herzen (also spelled Gertsen), a social philosopher, political journalist, and memorialist, was one of the founders of the radical tradition of the 19th-century Russian intelligentsia. He originated the theory of a unique Russian path to Socialism, or peasant Populism; he established abroad the first uncensored Russian periodical, *Kolokol* (*The Bell*); and he chronicled his career in one of the great works of Russian prose, *My Past and Thoughts*.



Herzen, oil painting by Nicolay Nicolayevitch Gué, 1867. In the State Tretyakov Gallery, Moscow.
By courtesy of the State Tretyakov Gallery, Moscow

Heritage and early influences

Born in Moscow on April 6 (March 25, old style), 1812, Herzen was the illegitimate son of a wealthy nobleman, Ivan Alekseyevich Yakovlev, and a German girl of humble origins. Reared in his father's house, he received an elite and far-ranging education from French, German, and Russian tutors. Still, the "taint" of his birth, as he regarded it, made him resentful of authority and, ultimately, of the autocratic, serf-based Russian social order. This resentment bred a cult of the Decembrists, a revolutionary group that staged an uprising in 1825. Herzen and his friend, Nikolay Ogaryov, who, like Herzen, was influenced by the dramas of the German playwright Friedrich Schiller, took a solemn oath to devote their lives to continuing the Decembrists' struggle for Russian freedom.

Attending the University of Moscow between 1829 and 1833, Herzen evolved from "romanticism for the heart to idealism for the head" and became an adept of the German philosopher Friedrich Schelling's *Naturphilosophie*.

Eventually Herzen and Ogaryov and their circle fused the pantheistic idealism of Schelling with the Utopian Socialism of the French social philosopher Henri de Saint-Simon to produce a philosophy of history in which the "World Spirit" evolved ineluctably toward the realization of freedom and justice.

This metaphysical politics was enough, however, to lead to the arrest of the entire circle in 1834. Herzen was exiled for six years to work in the provincial bureaucracy in Vyatka (now Kirov) and Vladimir; then, for an indiscreet remark about the police, he spent two more years in Novgorod. The misery of this period was relieved by an extravagantly romantic courtship and an exceptionally happy marriage with his cousin, Natalya Zakharina, in 1838.

Herzen's eight-year experience with injustice and the acquaintance it afforded with the workings of Russian government hardened his radicalism. He abandoned the nebulous idealism of Schelling for the thought of two

other contemporary German philosophies—the "realistic logic" of G.W.F. Hegel and the materialism of L.A. Feuerbach. He thus became a "Left-Hegelian," holding that the dialectic (development through the reconciliation of conflicting ideas) was the "algebra of revolution" and that the disembodied truths of "science" (*i.e.,* German idealism) must culminate in the "philosophy of the deed," or the struggle for justice as proclaimed by French Socialism. In later life Herzen explained that this metaphysical approach to politics was inevitable for his generation, since the despotism of Nicholas I made action impossible and thus left pure thought as the only free realm of expression.

Armed with these philosophical weapons, Herzen returned to Moscow in 1842 and immediately joined the camp of the Westernizers, who held that Russia must progress by assimilating European rationalism and civic freedom, in their dispute with the Slavophiles, who argued that Russian development must be founded on the Orthodox religion and the fraternal peasant commune. Herzen contributed to this polemic two able and successful popularizations of Left-Hegelianism, *Diletantizm v nauke* ("Dilettantism in Science") and *Pisma ob izuchenii prirody* ("Letters on the Study of Nature"), and a novel of social criticism, *Kto vinovat?* ("Who Is to Blame?"), in the new "naturalistic" manner of Russian fiction.

Soon, however, Herzen fell out with the other Westernizers because the majority of the group were reformist liberals, whereas Herzen had by now embraced the anarchist Socialism of the French social theorist Pierre-Joseph Proudhon. At this point, in 1846, Herzen's father died, leaving him a considerable fortune; and the following year Herzen left Russia for western Europe—as it turned out, for good.

Years in Paris

Herzen went immediately to the capital of European radicalism, Paris, hoping for the imminent triumph of social revolution. The revolutionary upheavals of 1848 that he witnessed in Paris and Italy soon disabused him: he became convinced that the Western "matadors of rhetoric" were too imbued with the values of the past to level the existing social order, that Europe's role as a progressive historical force was finished, that Western institutions were in fact "dead." He concluded further that, contrary to the teachings of the Hegelians, there was no "rational" inevitability in history and that society's fate was decided instead by chance and human will. He developed these themes in two brilliant but rather confused works, *Pisma iz Frantsii i Italii* ("Letters from France and Italy") and *Stogo berega* (*From the Other Shore*). His disillusionment was vastly increased by his wife's liaison with the radical German poet Georg Herwegh and by her death in 1852.

Loss of faith in the West, however, provoked a spiritual return to Russia: though "old" Europe, "fettered by the richness of her past," had proved incapable of realizing the ideal of Socialism, "young" Russia, precisely because its past offered nothing worth conserving, now seemed to Herzen to possess the resources for a radical new departure. And Herzen (borrowing an idea from his old foes, the Slavophiles) found these resources above all in the collectivist peasant commune, which he viewed as the basis for a future Socialist order. This new faith in Russia's revolutionary potential was expressed in letters to the French historian Jules Michelet and the Italian revolutionist Giuseppe Mazzini in 1850 and 1851.

Move to London

In 1852, Herzen moved to London and the following year, with the aid of Polish exiles, he founded the "Free Russian Press in London," the first uncensored printing enterprise in Russian history. In 1855, Nicholas I died, and soon thereafter Alexander II proclaimed his intention of emancipating the serfs. Responding to this unprecedented "thaw," Herzen rapidly launched a series of periodicals designed to be smuggled back to Russia: "The Polar Star" in 1855, "Voices from Russia" in 1856, and a newspaper, *Kolokol* (*The Bell*), created in 1857 with the aid of his old friend Ogaryov, now also an émigré. Herzen's aim was to influence both the government and the public toward emancipation of the peasants, with generous allotments of land and the liberalization of Russian society. To this end, he moderated his political pronouncements, speaking less of Socialist revolution and more of the concrete issues involved in Alexander's reforms. For a time he even believed in enlightened autocracy, hailing Alexander II in 1856 (in words that echoed the famous dying tribute of Julian the Apostate to Christ) with: "you have conquered, oh Galilean!" *The Bell* soon became a major force in public life, read by the Tsar's ministers and the revolutionary opposition.

Soon, however, the ambiguity of Herzen's position between reform and revolution began to cost him support. After 1858, moderate liberals, such as the author Ivan Turgenev, attacked Herzen for his Utopian recklessness; and after 1859 he quarrelled with the political writer N.G. Chernyshevsky and the younger generation of radicals, whose intransigent manner appeared to him as very dangerous to reform. He also lost faith in the government; when the Emancipation Act was finally enacted in 1861, he denounced it as a betrayal of the peasants.

He therefore veered again to the left, and called on the student youth to "go to the people" directly with the message of Russian Socialism. Furthermore, on the urging of the anarchist Mikhail Bakunin, he threw the support of *Kolokol* behind the unsuccessful Polish revolt of 1863. He immediately regretted this rashness, for it cost him the support of all moderate elements in Russia without restoring his credit among the revolutionaries. *Kolokol's* influence declined sharply. In 1865 Herzen moved his headquarters to Geneva to be near the young generation of Russian exiles, but in 1867 public indifference forced *Kolokol* to cease publication.

Decline of *The Bell*

Amidst these political reverses, Herzen turned his energies increasingly to his memoirs, *My Past and Thoughts*, which were designed to enshrine both his own legend and that of Russian radicalism. A loosely constructed personal narrative, interspersed with sharp vignettes of both Russian and Western political figures and with philosophical and historical digressions, it provides a masterful fresco of contemporary European radicalism. At times witty, irreverent, and playful in style, and at other times lyrical, passionate, and rhapsodical, it is one of the most original and powerful examples of Russian prose. Published principally between 1861 and 1867, its scope and quality have placed it alongside the great Russian novels of the 19th century in artistic stature.

In 1869 Herzen wrote letters *K staromu tovarishchy* ("To an Old Comrade"; Bakunin), in which he expressed new reservations about the costs of revolution. Still, he was unable to accept liberal reformism completely, and he expressed interest in the new force of the First International, Marx's federation of working-class organizations. This wavering position between Socialism and liberalism, which characterized so much of his career, proved to be his political testament. He died a few months later, on January 21 (January 9, O.S.), 1870, in Paris. The ambiguities of his position have made it possible ever since for both Russian liberals and Socialists to claim his legacy with equal plausibility.

BIBLIOGRAPHY. Works on Herzen include MARTIN MALIA, *Alexander Herzen and the Birth of Russian Socialism, 1812–1855* (1961), for his ideology and politics; E.H. CARR, *The Romantic Exiles: A Nineteenth-Century Portrait Gallery* (1933), for his personal life; *My Past and Thoughts: The Memoirs of Alexander Herzen*, 5 vol., trans. by CONSTANCE GARNETT (1924–26; rev. by HUMPHREY HIGGENS, 4 vol., 1968); and the introduction by ISAIAH BERLIN to Herzen's *From the Other Shore* (1956). Collected editions of his works (in Russian) have been edited by M. LEMKE, 22 vol. (1915–25); and by the SOVIET ACADEMY OF SCIENCES, 30 vol. (1954–66).

(M.E.Ma.)

Herzl, Theodor

Theodor Herzl was the founder of the so-called Zionist movement to establish a Jewish homeland. Although he died more than 40 years before the establishment of the State of Israel, he was an indefatigable organizer, propagandist, and diplomat who had much to do with making Zionism into a political movement of worldwide significance.



Herzl, 1904.
By courtesy of the Zionist Archives and Library, New York

Herzl was born in Budapest, Hungary, on May 2, 1860, of well-to-do middle class parents. He first studied in a scientific secondary school, but to escape from its anti-Semitic atmosphere he transferred in 1875 to a school where most of the students were Jews. In 1878 the family moved from Budapest to Vienna, where he entered the University of Vienna to study law. He received his license to practice law in 1884 but chose to devote himself to literature. For a number of years he was a journalist and a moderately successful playwright.

In 1889 he married Julie Naschauer, daughter of a wealthy Jewish businessman in Vienna. The marriage was unhappy, although three children were born to it. Herzl had a strong attachment to his mother, who was unable to get along with his wife. These difficulties were increased by the political activities of his later years, in which his wife took little interest.

Conversion to Zionism

A profound change began in Herzl's life soon after a sketch he had published in the leading Viennese newspaper, *Neue Freie Presse*, led to his appointment as the paper's Paris correspondent. He arrived in Paris with his wife in the fall of 1891 and was shocked to find in the homeland of the French Revolution the same anti-Semitism with which he had become so familiar in Austria. Hitherto he had regarded anti-Semitism as a social problem that the Jews could overcome only by abandoning their distinctive ways and assimilating to the people among whom they lived. At the same time, his work as a newspaperman heightened his interest in, and knowledge of, social and political affairs and led him to the conviction that the answer to anti-Semitism was not assimilation but organized countereforts by the Jews. The Dreyfus affair in France also helped crystallize this belief. French military documents had been given to German agents, and a Jewish officer named Alfred Dreyfus had been falsely charged with the crime. The ensuing political controversy produced an outburst of anti-Semitism among the French public. Herzl said in later years that it was the Dreyfus affair that had made a Zionist out of him. So long as anti-Semitism existed, assimilation would be impossible, and the only solution for the majority of Jews would be organized emigration to a state of their own.

Herzl was not the first to conceive of a Jewish state. Orthodox Jews had traditionally invoked the return to Zion in their daily prayers. In 1799 Napoleon had thought of establishing a Jewish state in the ancient lands of Israel. The English statesman Benjamin Disraeli, a Jew, had written a Zionist novel, *Tancred*. Moses Hess, a friend and coworker of Karl Marx, had published an important book, *Rom und Jerusalem* (1862), in which he declared the restoration of a Jewish state a necessity both for the Jews and for the rest of humanity. Among the Jews of Russia and eastern Europe, a number of groups were engaged in trying to settle emigrants in agricultural colonies in Palestine. After the Russian pogroms of 1881, Dr. Leo Pinsker had written a pamphlet, "Auto-Emanzipation," an appeal to western European Jews to assist in the establishment of colonies in Palestine. When Herzl read it some years later, he commented in his diary that if he had known of it, he might never have written *The Jewish State*.

Herzl's first important Zionist effort was an interview with Baron Maurice de Hirsch, one of the wealthiest men of his time. De Hirsch had founded the Jewish Colonization Association with the aim of settling Jews from Russia and Romania in Argentina and other parts of the Americas. The 35-year-old journalist arrived at the Baron's mansion in Paris with 22 pages of notes, in which he argued the need for a political organization to rally the Jews under a flag of their own, rather than leaving everything to the philanthropic endeavours of individuals like the Baron. The conversation was notable for its effect on Herzl rather than on the Baron de Hirsch, who refused to hear him out. It led to Herzl's famous pamphlet *The Jewish State*, published in February 1896 in Vienna. The Jewish question, he wrote, was not a social or religious question but a national question that could be solved only by making it "a political world question to be discussed and settled by the civilized nations of the world in council." Some of Herzl's friends thought it a mad idea, but the pamphlet won favourable response from eastern European Zionist societies. In June 1896, when Herzl was en route to Constantinople (Istanbul) in the hope of talking to the Ottoman Sultan about obtaining the grant of Palestine as an independent country, his train stopped in Sofia, Bulgaria; hundreds of Jews were present at the station to greet Herzl and to hail him as a leader.

Although he remained in Constantinople for 11 days, he failed to reach the Sultan. But he had begun the career as organizer and propagandist that would end only with his death eight years later. He went to London in an effort to organize the Jews there in support of his program. Not all the Jewish leaders in England were happy to see him because his political approach was not in tune with their ideas, but at public meetings in the East End he was loudly cheered. He was a tall, impressive figure with a long black beard and the mien of a prophet. Despite his personal magnetism, he found that his efforts to influence Jewish leaders in England were of little avail and therefore decided to organize a world congress of Zionists in the hope of winning support from the masses of Jews in all countries. He proposed to hold the congress in Munich, but as the Jews there—who were mostly assimilated—opposed it, he settled upon Basel, Switzerland. The congress met at the end of August 1897, attended by about 200 delegates, mostly from central and eastern Europe and Russia along with a few from western Europe and even the United States. They represented all social strata and every variety of Jewish thought—from Orthodox Jews to atheists and from businessmen to students. There were also several hundred onlookers, including some sympathetic Christians and reporters for the international press. When Herzl's imposing figure came to the podium, there was tumultuous applause. "We want to lay the foundation stone," he declared, "for the house which will become the refuge of the Jewish nation. Zionism is the return to Judaism even before the return to the land of Israel." One of Herzl's most faithful supporters was the writer Max Nordau, who gave a brilliant address in which he described the plight of the Jews in the East and in the West. The three-day congress agreed upon a program, henceforth to be known as the Basel Program, declaring that "Zionism aspires to create a publicly guaranteed homeland for the Jewish people in the land of Israel." It also set up the Zionist Organization with Herzl as president.

The First Zionist Congress

The seven remaining years of his life were devoted to the furtherance of the Zionist cause, although he remained literary editor of the *Neue Freie Presse* in order to earn a living. He established a Zionist newspaper, *Die Welt*, published as a German-language weekly in Vienna. He negotiated unsuccessfully with the Sultan of Turkey for the grant of a charter that would allow Jewish mass

The Sinai and Uganda projects

settlement in Palestine on an autonomous basis. He then turned to Great Britain, which seemed favourable to the establishment of a Jewish settlement in British territory in the Sinai Peninsula. When this project failed, the British proposed Uganda in East Africa. This offer, which he and some other Zionists were willing to accept, aroused violent opposition at the Zionist congress of 1903, particularly among the Russians. Herzl was unable to resolve the conflict. He died of a heart ailment at Edlach, near Vienna, on July 3, 1904, at the age of 44. He was buried in Vienna, but, in accordance with his wish, his remains were removed to Jerusalem in 1949 after the creation of the Jewish state and entombed on a hill west of the city now known as Mt. Herzl.

After the First Zionist Congress in Basel, Herzl had written in his diary:

> If I had to sum up the Basel Congress in one word—which I shall not do openly—it would be this: At Basel I founded the Jewish state. If I were to say this today, I would be greeted by universal laughter. In five years, perhaps, and certainly in 50, everyone will see it.

While the Jewish state was the product of many complex historic forces, including two world wars and the labours of Herzl's many followers, it was he who organized the political force of Jewry that was able to take advantage of the accidents of history. Through the strength of his personality, he aroused the enthusiasm of the Jewish masses and gained the respect of many statesmen of his time, in spite of the opposition of some Jewish leaders to his plans.

BIBLIOGRAPHY. ALEX BEIN, *Theodore Herzl* (1940; orig. pub. in Hebrew; Eng. trans. from the German of 1934), is the best biography available in English.

(D.B.-G.)

Hesiod

One of the earliest Greek epic poets and the first man in Western civilization to embody precepts or instructions in poetry, Hesiod, through his works, serves as a useful corrective to Homer's more glamorous portrayal of the world. Hesiod has an essentially serious outlook on life and is an artist who deals with the gloomier side of existence, relating, in his *Theogony* (the title not being by Hesiod himself and meaning a history of the generation or birth of the gods), the bloody power struggle among the divine dynasts Uranus, Cronus, and Zeus, while his *Works and Days* demonstrates that, in Hesiod's immediate circle at any rate, mankind's situation on earth was equally deplorable during what he calls the "age of iron."

By courtesy of the Landesmuseum, Trier



Hesiod, detail of a mosaic by Monnus, 3rd century AD. In the Rheinisches Landesmuseum, Trier.

Not a great deal is known about the details of Hesiod's life. A native of Boeotia, a district of central Greece, he flourished around 700 BC. Hesiod himself provides information about his life in his *Works and Days*, in which he vehemently criticizes a covetous brother who has taken the lion's share of their inheritance and who has bribed the rulers of the community to support him. In an attempt to remind them of their duty, Hesiod urges reverential awe of Justice, whom he regards as a goddess. To placate her is to attain happiness, the only possible solution, he thinks, to the miseries man currently suffers.

On one occasion Hesiod participated in a song contest at some funeral games, contests held in honour of one Amphidamas, and was awarded a tripod (a common prize at the time and, like the modern cup, prestigious rather than valuable). This occasion, Hesiod says, was his only overseas trip; the festival took place at Chalcis, which is situated on the island of Euboea some 800 or 900 yards (700 or 800 metres) off mainland Greece. Otherwise, he spent most of his life in his home village of Ascra, near Mt. Helicon, originally as a shepherd, until the Muses (companions of Apollo and inspirers of poetry) appeared, endowing him with a poet's voice and bidding him "sing of the race of the blessed gods immortal."

Genuine works. Of Hesiod's two extant epics, the *Theogony* is clearly the earlier. In it, following the Muses' instructions, Hesiod recounts the history of the gods, beginning with the emergence of Chaos, Gaea (Earth), and Eros. Gaea gives birth to Uranus (Heaven), the Mountains, and Pontus (the Sea); and later, after uniting herself to Uranus, she bears many other deities. One of them is the Titan Cronus, who rebels against Uranus, emasculates him, and afterward rules until he in turn is overpowered by Zeus. This story of crime and revolt is interrupted by many additional pedigrees of gods, including the progeny of Night, a deity sprung directly from the original primordial Chaos; here Hesiod has brought together the forces of evil (Death, Strife, Hunger, and Woe) that beset the life of man. Elsewhere, in addition to mythical family relations, Hesiod presents new ones that are the product of his own speculation. Thus, the names of the 50 sea-maidens (the Nereids) fathered by the sea-god Nereus indicate various qualities of the Sea. In a different way, the story describing the first woman, Pandora, sent by Zeus to bedevil man, brings out Hesiod's firm belief in the supreme and irresistible power of Zeus, This power is most majestically displayed in the Titanomachia, the battle between the Olympian gods, led by Zeus, and the Titans, who support Cronus. After his victory Zeus organizes his kingdom.

Content of the *Theogony*

Hesiod's authorship of the *Theogony* has been questioned but is no longer doubted, though the work does include sections inserted by poets and rhapsodes (professional reciters of poetry) before the degeneration of the epic as an art form in the late 6th century BC. The story of Typhoeus' rebellion against Zeus has almost certainly been added by someone else, while the somewhat overlapping accounts of Tartarus, the hymn on Hecate, and the progeny of the sea monster Keto are highly suspect. The discovery of a Hurrian theogony similar to Hesiod's seems to establish a link with the Near East. But with Hesiod the succession Uranus–Cronus–Zeus resembles a classical Greek tragic trilogy. Thus, the Erinyes (the deities of vengeance) are born when Uranus is overthrown by Cronus, while their own hour for action comes when Cronus is about to be overthrown by Zeus. These and other similar features plausibly represent Hesiod's own contributions to the inherited story.

Hesiod's other epic poem, the *Works and Days*, has a more personal character. It is addressed to his grasping brother Perses in an attempt to make him stop his evil practices. In the first part, Hesiod recounts two myths illustrating the necessity for honest, hard work in man's wretched life. One takes up and continues the story of Pandora, who out of curiosity opens a jar, loosing multifarious evils on humanity; the other traces man's decline since the Golden Age, down through the Silver, Bronze, and Heroic to Hesiod's own miserable era of iron.

Personal character of the *Works and Days*

Hesiod also speaks to Perses directly, urging him to abandon his schemes and thenceforth to gain his livelihood through strenuous and persistent work: "Before success the immortal gods have placed the sweat of our brows." Hard work is for Hesiod the only way to prosperity and distinction.

In the second half of the poem, Hesiod describes with much practical detail the kind of work appropriate to

each part of the calendar and how to set about it. The description of the rural year is enlivened by a vivid feeling for the rhythm of human life and the forces of nature, such as the overpowering winter storm, which drives man back into his home, and the parching heat of summer, during which he must rest.

The poem ends with a series of primitive taboos and superstitions, followed by a section explaining which days of the month are auspicious for sowing, threshing, shearing, and the begetting of children. It is difficult to believe that either of these sections could have been composed by Hesiod.

Spurious works. Such was the power of Hesiod's name that epics by other poets were soon attributed to him; these are often included in editions of his works. The *Precepts of Chiron*, the *Astronomy*, the *Ornithomanteia* (Divination by Birds), the *Melampodeia*, which described a contest between two seers, and the *Aigimios* are today little more than names. There are numerous extant fragments of the *Catalogues of Women*, which deals primarily with women who through union with gods become mothers of heroes and ancestresses of noble families. Papyruses deciphered since the 1890s, and especially in the 1950s and 1960s, have added much to knowledge of its content and have made it possible to arrive at a clearer idea of its organization. There is no evidence for the theory that the oldest parts are by Hesiod. The story of Alcmene, Heracles' mother, is extant in an expanded form as the *Shield of Heracles*, probably dating from the early or middle 6th century. In its present form the *Contest Between Homer and Hesiod*, ending in Hesiod's victory, postdates the emperor Hadrian (2nd century).

BIBLIOGRAPHY

Works: *The Works and Days, Theogony*, and *The Shield of Herakles*, trans. by R. LATTIMORE (1959), a brisk, modern translation, perhaps appreciated best when sampled along with *Hesiod, the Homeric Hymns, and Homerica*, trans. by H.G. EVELYN-WHITE, rev. ed. (1959), an antique but accurate translation, with parallel text. More exotic translations include *Fable of the Hawk and the Nightingale*, trans. by ROBERT GRAVES (1959); and *The Georgicks of Hesiod*, trans. by G. CHAPMAN (1618 and 1858). The most recent edition of the original Greek text of Hesiod's epics and a selection of "fragments" (*i.e.*, passages preserved from the *Catalogues of Women* . . .), ed. by R. MERKELBACH, F. SOLMSEN, and M.L. WEST was published in 1970.

Biography and criticism: Scholarly analyses of Hesiod's poems and discussions of various facets of these works are provided by F. SOLMSEN, *Hesiod and Aeschylus* (1949); and M.L. WEST in the "Prolegomena" of *Hesiod Theogony* (1966). The best appreciation of his poetic individuality is in HERMANN FRANKEL, *Dichtung und Philosophie des frühen Griechentums* (1951). Important for Hesiod's place in the history of Greek thought is WERNER JAEGER, *Paideia: The Ideals of Greek Culture*, trans. from the German by GILBERT HIGHET, 2nd ed., vol. 1 (1945, reprinted 1970). Near-Eastern theologies comparable with Hesiod's *Theogony* are summarized by West; for suggestions about the relationship, see P. WALCOT, *Hesiod and the Near East* (1966).

(F.So.)

Hessen

A *Land* (state) of Germany, that, following partition of the nation after World War II, became part of the Federal Republic of Germany, Hessen (Hesse) had an area of 8,151 square miles (21,111 square kilometres) and a population of some 5,400,000 in the early 1970s. The state lies between the Upper Rhine Plateau and the Thüringer Wald (Thuringian Forest) and was formed in 1945 through the amalgamation of former Prussian provincial units. Its capital is Wiesbaden.

Landscape and people. Hessen consists mainly of richly wooded hill land cut by two fault troughs, the west and east Hessen depressions, which have been important corridors since earliest times. Along these troughs most of the state's main towns were founded in the Middle Ages. The area is dominated to the south by the Vogelsberg, a great basaltic mass, and the Rhön, a mountainous mass rising to the Wasserkuppe (3,117 feet [950 metres]), Hessen's highest mountain. The Spessart and the Odenwald regions both belong in part to Hessian territory.

The Vogelsberg and the Rhön

Hessen is drained by the Rhine and its tributaries the Main and the Lahn, by the Weser, and by the Werra and Fulda. The greater part of the *Land* lies in the country of hills and lowlands that is drained north by the Fulda and Eder rivers above their confluence to form the Weser River. Beechwoods and conifers cover the highlands, and cultivated land lies on the limestone uplands and on the loess soils of the lowlands proper. This is a predominantly unspoiled agricultural countryside with small historic towns.

In the early 1970s, according to the available figures, there were five towns with populations of over 100,000: Frankfurt am Main (669,000), Wiesbaden (250,100), Kassel (214,200), Darmstadt (142,100), and Offenbach am Main (117,300). About 61 percent of the population is Protestant and 33 percent Roman Catholic.

The economy. The state is well wooded, and small-scale farming is still widespread; of over 2,400,000 persons who were employed at the 1970 census, more than 150,000 were engaged in agriculture and forestry. Nearly 1,200,000 were involved in power supply, mining, manufacturing, and building. Potatoes and sugar beets are the chief crops produced in quantity, more than 1,200,000 and 880,000 metric tons respectively in the early 1970s. Wheat is the most widely grown crop, covering more than 320,000 acres (130,000 hectares)—more than double the space devoted to potatoes, which, in turn, have almost two and a half times the acreage accorded to sugar beets.

Poultry outnumbers humans by nearly a million, numbering almost 6,225,000. Pigs total nearly 1,400,000, while cattle number about 900,000. The southwest of Hessen is primarily industrial but is also an area of intensive agriculture (which occupies about one-fifth of its workers). The plains along the Rhine and Main rivers are a mosaic of vineyards, orchards, and fields of grain, potatoes, and tobacco. This lowland has an early spring and a warm summer, and the total rainfall is only about 20 inches (500 millimetres) a year. Market gardening is especially important near the cities. The surrounding hills have a three-year rotation of rye, oats, and potatoes, and livestock farms concentrate on the production of butter and cheese.

Natural resources

Natural resources are small; there are some low-grade iron ores in the Tanus Mountains, which are of little economic significance; salt mines near Fulda; and small, brown-coal deposits near Frankfurt am Main and Kassel. The industries depend on the Rhine waterway and its extensions up the Main and Neckar. The Rhine–Main area, centred on Frankfurt am Main, Mainz, and Wiesbaden, is one of the great industrial regions of Germany. Kassel, Offenbach, Wiesbaden, and Darmstadt are other big manufacturing centres. Vehicles, machinery, chemicals (especially at Höchst, a western quarter of Frankfurt am Main), electrical goods, scientific instruments, and textiles are among the products of these and other towns. New industries have developed since World War II, stimulated by the arrival of refugees; these manufactures include the making of glass, toys, and musical instruments. Book publishing is a prominent economic activity.

Transportation. In the early 1970s, there were nearly 9,900 miles (16,000 kilometres) of classified roads, about 450 miles of *Autobahn*, more than 2,150 miles of federal highway and more than 4,400 miles of first-class highway. Motor vehicles numbered more than 1,700,000. A symbol of progress in transportation is the cloverleaf interchange on the "Frankfurt Cross" *Autobahn*. The Post Office, Federal Railway, and private companies all run bus services. The Rhine provides the chief waterway, and its economic importance cannot be overestimated. In air traffic Hessen is paramount, Frankfurt's Rhine–Main Airport being western Europe's busiest after Heathrow (London) and Orly (Paris). In the early 1970s, it was handling more than 9,300,000 passengers a year. Rail travel in Hessen, as throughout the Federal Republic of Germany, is largely electrified and has international European links.

Administration. Hessen is divided into two administrative districts (*Regierungsbezirke*)—Kassel and Darmstadt—which are subdivided into municipal and rural dis-

tricts (*Stadtkreise* and *Landkreise*). Education is free and widely organized. There are part-time, full-time, and advanced vocational schools and schools for public-health occupations. In addition, 23 institutes for training technicians and 19 colleges of engineering exist.

There are three universities, Frankfurt am Main, Giessen, and Marburg, the latter boasting the largest student hostel in the Federal Republic. These are supplemented by a technical university at Darmstadt, one Protestant and three Roman Catholic theological colleges, eight teacher-training colleges, one college of music, and two colleges of fine arts.

There are nearly 350 hospitals with over 60,000 beds. The social-welfare public assistance includes aid to tuberculars and war victims.

Justice is administered by a constitutional court (*Staatsgerichtshof*), a court of appeal, and various other courts dealing with regional jurisdiction, labour, finance, administrative matters, and cases coming under social headings.

Cultural life. Old traditions remain very much alive in Hessen, especially in the thriving city of Frankfurt am Main. Hessen provided the chief source for the brothers Grimm when they collected their fairy tales—particularly in the Weser Gebirge (hills), on either side of the Weser River and around Kassel. On the Weser's banks are many ruined castles, old churches, and palaces, and there are also burgher houses and town halls throughout the state.

Frankfurt am Main has the Cathedral of St. Bartholomew, the Goethe Museum (1961) standing beside the Goethehaus, the 18th-century poet's birthplace, a town hall (Römer), a picture gallery, and an opera house. Kassel, where an international modern art exhibition is held every four years, boasts a wallpaper museum (Wilhelmshöhe Castle), a museum and gallery with a large collection of works by the 17th-century Dutch painter Rembrandt, and the Brothers Grimm Museum. Wiesbaden and Darmstadt also have very respectable art galleries. Interesting half-timbered buildings and houses in later styles are prolific throughout the state.

BIBLIOGRAPHY. Little material specifically on Hessen is in English. An exhaustive bibliography of ancient to contemporary works is in *Schrifttum zur Geschichte und geschichtlichen Landeskunde von Hessen*, ed. by KARL E. DEMANDT, 3 vol. (1965–68); see also the same author's *Geschichte des Landes Hessen* (1959). A more elementary but useful history of the postwar *Land* and its predecessor states is KURT FINKE, *Hessen: Vergangenheit und Gegenwart* (1970). A firsthand account of the early post-war period appears in DEXTER L. FREEMAN, *Hesse: A New German State* (1948). The standard geography series *Harms Landeskunde* devotes vol. 1, ed. by JULIUS WAGNER, to *Hessen* (1961). Detailed information on Hessen's population, economy, political structure, judiciary, education, and modern historical development are given in a series of volumes of the Hessisches Statistisches Landesamt, Wiesbaden. For a handy ready-reference with abundant information, see the *Hessenlexikon* (1965). A well-illustrated, informative treatment of the Hessian economy appears in *Hessen um Rhein und Main*, 2nd ed. by KLAUS ERHR. VON VERSCHUER (1966). Hessen's lead in reform of primary and secondary education is covered in ROBERT GEIPEL, *Bildungsplanung und Raumordnung* (1968); health and social care are discussed in the HESSISCHES SOZIALMINISTERIUM, *Heute für Morgen: Hessens Gesundheits- und Sozialwesen* (1970). Aspects of Hessian ethnology are dealt with in INGEBORG WEBER-KELLERMANN, *Volksleben in Hessen* (1970). For works in English reflecting portions specifically devoted to Hessen, see the bibliography of the article GERMANY, FEDERAL REPUBLIC OF.

Heterocyclic Compounds

Heterocyclic compounds are a major class of organic (carbon-containing) chemical compounds, characterized by the fact that the atoms in their molecules are joined into rings, or circles, containing at least one atom of an element other than carbon. These compounds are of great importance because many of the biochemical materials essential to life belong to the class. Nucleic acids, for example, the chemical substances that carry the genetic information controlling inheritance, consist of long chains of heterocyclic units held together by other types of materials. Many naturally occurring pigments, vitamins, and antibiotics are heterocyclic compounds, as are most hallucinogens, substances that produce hallucinations. Modern society is dependent on synthetic heterocycles for use as pharmaceuticals, pesticides, and herbicides, as well as dyes and plastics.

The most common heterocycles are those with five- or six-membered rings, containing atoms of nitrogen, oxygen, or sulfur. The best known of the simple heterocyclic compounds are pyridine, pyrrole, furan, and thiophene. Pyridine and pyrrole are both nitrogen heterocycles; that is, their molecules contain nitrogen atoms in the rings with carbon. Because the molecules of many biological materials consist in part of pyridine and pyrrole rings, strong heating of such materials produces small amounts of pyridine and pyrrole. In fact, both of these substances were discovered in the 1850s in an oily mixture of substances formed by strong heating of bones. Today, however, pyridine and pyrrole are generally obtained from coal tar or are prepared by synthetic reactions. The chief commercial interest in both compounds lies in their conversion to other substances, chiefly dyestuffs and drugs. Pyridine is used also as a solvent, a waterproofing agent, a rubber additive, an alcohol denaturant, and a dyeing adjunct. Furan is an oxygen-containing heterocycle, used primarily for conversion to other substances (including pyrrole). Furfural, a close chemical relative of furan, is obtained from oat hulls and corncobs and is used in the production of ingredients of nylon. Thiophene, a sulfur heterocycle, resembles benzene in its chemical and physical properties. It is a frequent contaminant of benzene from natural sources and was first discovered during purification of benzene. Like the other compounds, it is used primarily for conversion to other substances. Furan and thiophene both were discovered in the latter part of the 19th century.

The common simple heterocyclic compounds

Each of the molecules of pyridine contains a ring of six atoms, five of which are carbon atoms and one of which is a nitrogen atom. Pyrrole, furan, and thiophene molecules contain five-membered rings, composed of four atoms of carbon and one atom of nitrogen, oxygen, or sulfur, respectively.

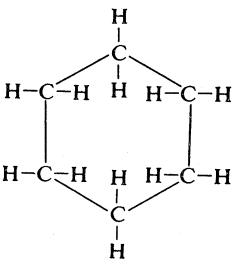
In general, the physical and chemical properties of heterocyclic compounds are best understood by comparing them with ordinary organic compounds that do not contain hetero-atoms (noncarbon atoms). In this article, the comparative aspects of heterocyclic compounds will be considered first, followed by a brief survey of the different classes of heterocyclic compounds, arranged according to the size and nature of the ring.

GENERAL ASPECTS OF HETEROCYCLIC SYSTEMS

Comparison with carbocyclic compounds. The molecules of organic chemical compounds are built up from a framework or backbone of carbon (C) atoms to which are attached hydrogen (H), oxygen (O), or other heteroatoms. Carbon atoms have the unique property of being able to join with one another to form chains of atoms. When the ends of the chains are joined together, ring—that is, cyclic—compounds result; such substances often are referred to as alicyclic or carbocyclic compounds. Substitution of one or more of the ring carbon atoms in the molecules of a carbocyclic compound with a heteroatom gives a heterocyclic compound.

Carbocyclic compounds

A typical carbocyclic compound is cyclohexane (C_6H_{12}), the molecular structure of which is indicated by the following formula

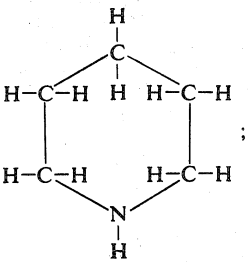


in which the letters represent atoms of the elements of which they are symbols and the lines represent bonds (linkages) between the atoms. For convenience such formulas are often written in the simplified polygonal form, such as

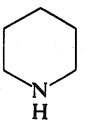


for cyclohexane, in which each angle of the polygon represents a carbon atom (it being understood that hydrogen atoms are joined to the carbon atoms as required).

When one of the carbon atoms of cyclohexane is replaced with an atom of nitrogen, the compound piperidine (a chemical relative of pyridine, above) is produced. The structural formula of piperidine is written as follows:



; or for convenience:

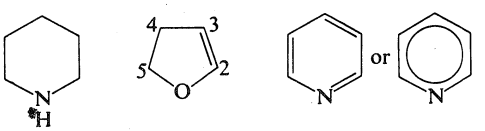


Other heterocyclic compounds can be envisioned as having been produced (similarly) from cyclohexane by substitution with other hetero-atoms or from other carbocyclic compounds by substitution with nitrogen or other hetero-atoms.

Paraffinic, olefinic, and aromatic compounds

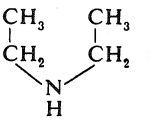
The simplest organic compounds are the hydrocarbons, compounds of carbon and hydrogen only. Hydrocarbons are classed as paraffinic if all the potential bonds of the carbon atoms are saturated—that is, if the four possible bonds of each carbon atom are joined singly to another carbon atom or to a hydrogen atom. They are classified as olefinic if they contain a double bond (also called an unsaturated linkage) between any two of the carbon atoms; and they are classed as aromatic if they contain a ring of alternating single and double bonds. Compounds with unsaturation are highly reactive—that is, they readily undergo additions of atoms or groups of atoms to the carbon atoms of their double bonds, giving each carbon four groups joined to it. Aromatic compounds, though unsaturated, are extremely stable and do not undergo the addition reactions characteristic of other unsaturated compounds. The stability and unreactivity of the aromatic system are associated with presence of three pairs of electrons, the so-called pi electrons, associated with the three double bonds of the ring. Together these electrons, comprising the so-called aromatic sextet, form an unusually stable structure, associated with the aromatic ring as a whole rather than with the individual atoms.

Heterocycles, too, may be classified as paraffinic, olefinic, and aromatic. Thus, piperidine is a heteroparaffinic compound containing no double bond, whereas 4,5-dihydrofuran is a hetero-olefinic compound, and pyridine is a typical hetero-aromatic substance, as shown in the following structural formulas, in the first of which the double bonds are shown and in the second of which the aromatic sextet is indicated by a circle.

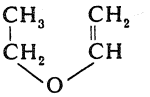


piperidine 4,5-dihydrofuran pyridine

This classification relates the chemistry of heterocycles directly with that of nonheterocyclic derivatives, which are usually better known. In general, synthetic methods and physical and chemical properties of the saturated heteroparaffinic and the partly unsaturated hetero-olefinic compounds closely resemble those for their acyclic (noncyclic) analogues. Thus, piperidine may be considered as a cyclic secondary amine (organic nitrogen compound) and has much in common with the acyclic amine, diethylamine, which is represented as follows:



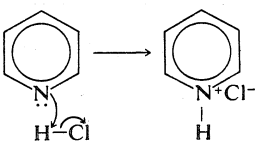
Similarly, 4,5-dihydrofuran mirrors many of the properties of the unsaturated ether, ethyl vinyl ether, written as follows:



It is within the area of hetero-aromatic compounds that most of the novel chemistry of the class is to be found, and for that reason hetero-aromatics will be emphasized in this article.

Chemical reactions. Every chemical reaction involves the forming or breaking of a chemical bond: for heterocycles the bonds in question are generally covalent bonds—linkages consisting of a pair of electrons shared by two atoms. Reactions generally can be placed in one of several categories, depending upon the origin, or disposition, of the electron pair of the bond that is formed or broken. The first of these categories is heterolytic reactions, in which the electron pair is supplied by one of the reactants (known as the nucleophile) to the other reactant (known as the electrophile). In the formation of pyridinium chloride, for example, which occurs as shown in the following equation:

Heterolytic reactions



in which, as is customary, the structure of the starting material is written at the left, the product is written at the right, and an arrow in between the two indicates the reaction. In these structural formulas the presence of a free electron pair on the nitrogen atom is indicated by two dots, the movements of electrons are shown by curved arrows, and plus and minus signs indicate positive and negative charges. In this reaction, the pyridine molecule behaves as a nucleophile, the electron pair that initially resided on the nitrogen atom finally being shared between that atom and the hydrogen atom of the hydrogen chloride (which acts, therefore, as an electrophile). In homolytic reactions, an electron-pair bond is formed (or broken) and in this process each of the reagents donates (or receives) a single electron. In some reactions, cyclic transition states are involved, which must be classified separately.

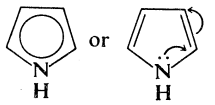
In a heterolytic bond-forming reaction the reagent that accepts the electrons is known as an electrophilic reagent; for instance, in the coordination reaction with hydrogen chloride (above), the nitrogen atom of pyridine acts as a nucleophilic reagent. This behaviour is characteristic of pyridine. The nitrogen atom in pyrrole, by contrast, does not act as a nucleophilic centre, a distinction that can be rationalized by the different type of electronic interaction present in these two hetero-aromatic molecules (as explained below).

The nature of hetero-aromaticity. Aromaticity denotes the significant stabilization of a ring compound by cyclic conjugation in which six pi electrons generally participate. A nitrogen atom in a ring can carry a positive or negative charge, or it can be in the neutral form. Ring oxygen and sulfur atoms can occur either in the neutral form or with a positive charge. A fundamental distinc-

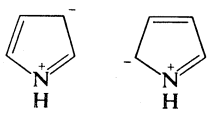
tion is usually made between: (1) those hetero-atoms that enter into aromatic conjugation by means of a lone electron pair in an orbital perpendicular to the plane of the ring, and (2) those hetero-atoms that enter into conjugation because they are connected to another atom by means of a double bond.

Pyrrole- and pyridine-type hetero-atoms

An example of an atom of the first type is the nitrogen atom in pyrrole. In pyrrole, the aromatic electron sextet is made up by the participation of two electrons from each carbon–carbon double bond, as well as the two electrons that comprise the unshared pair of the nitrogen atom. As a consequence, there tends to be a net flow of electron density from the nitrogen to the carbon atoms as the nitrogen electrons are drawn into the aromatic sextet; this movement may be indicated as follows,

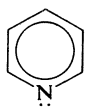


in which the curved arrows depict the movement, or tendency toward movement, of the electron pair. Alternatively, the pyrrole molecule may be described as a resonance hybrid—that is, a molecule whose true structure can only be approximated by two or more different forms, called resonance forms. With pyrrole, resonance forms such as the following make important contributions to the overall structure.

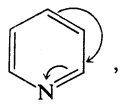


In these forms the nitrogen atom is depleted of electrons, as indicated by the positive charge.

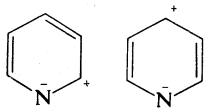
An example of a hetero-atom of the second type is the nitrogen atom in pyridine. Pyridine also has a pi electron sextet; but the nitrogen atom contributes only one electron to it, one further electron being contributed by each of the five carbon atoms. In particular, the lone electron pair on the nitrogen atom is not involved, as indicated in the following formulation.



Moreover, because of the greater attraction for electrons (electronegativity) of the nitrogen atom as compared to carbon atoms, there is a tendency for an electron flow toward the nitrogen atom, rather than away from it, as in pyrrole. This movement of electrons is shown as below

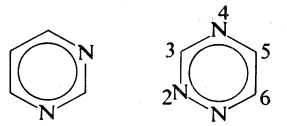


or it can be expressed in terms of contributions of the resonance forms as follows:



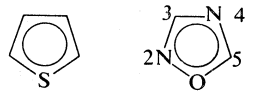
Quite generally, hetero-atoms may be referred to as pyrrole-like or pyridine-like, depending upon whether they fall into the first or second class, respectively, above. The pyrrole-like hetero-atoms $-NR-$ (R being a hydrocarbon group), $-N^--$, $-O-$, and $-S-$ tend to donate electrons into the pi electron system, whereas the pyridine-like hetero-atoms $-N=$, $-N^+R=$, $-O^+=$, and $-S^+=$ tend to attract the pi electrons of a double bond.

A six-membered hetero-aromatic ring contains one or more pyridine-like hetero-atoms (usually nitrogen), as is the case with the compounds pyrimidine and 1,2,4-triazine.



pyrimidine 1,2,4-triazine

A six-membered hetero-aromatic compound cannot normally contain any pyrrole-like hetero-atoms. A five-membered hetero-aromatic ring, however, always contains one pyrrole-like nitrogen, oxygen, or sulfur atom, and it may also contain up to four pyridine-like hetero-atoms, as shown by the compounds below.



thiophene 1,2,4-oxadiazole

The quantitative measurement, or even the precise definition, of aromaticity is difficult; two methods have been widely used to measure the aromaticity of carbocyclic compounds. The first method depends on the determination of the heat given off on complete combustion of the aromatic compound—say, benzene—compared with that expected for the corresponding carbocyclic compound containing three conventional double bonds—say, a cyclohexatriene. Because the heat of combustion is related to the energy content of the molecule, comparisons of the above kind reveal the reduction of energy associated with the aromatic system. The second method depends on measurement of magnetic ring currents set up in aromatic systems by external magnetic fields. The measurement is made with a nuclear magnetic resonance spectrometer, and the degree of observed magnetic effect is related to the degree of aromaticity. Both methods are difficult to apply to hetero-aromatic systems because of complications arising from the presence of the hetero-atoms. Table 1 gives some values for aromatic stabilization determined by combustion or related methods. Unfortunately, high errors are involved, as indicated by the wide variations in some of the values.

Measurements of aromaticity

Table 1: Aromatic Stabilization Energies for Benzene and Some Hetero-aromatic Rings
(kcal/mol)

benzene 36	pyridine 32	pyrazine 24
pyrrole 14–31	thiophene 24–31	furan 17–23
pyrazole 27–41	imidazole 12–32	1,2,4-triazole 20–49

With regard to its chemical reactivity, an aromatic compound is characterized by extra stability of the conjugated system it contains; this characteristic, in turn, is denoted by the tendency of the compound to react by substitution (replacement of a hydrogen atom) rather than by addition to the double bonds. From a reactivity

Table 2: Boiling Points of Saturated Heterocycles and Carbocycles of the Same Ring Size
(°C at 1 atm)

ring size	number and orientation of hetero-atoms	hetero-atom type NH	O	S	saturated cycloalkane
3	one	56	11	55	−34
4	one	63	48	94	13
5	one	87	66	121	49
6	one	106	88	141	80
6	two (1, 2)	165*	—	—	80
6	two (1, 3)	150	106	—	80
6	two (1, 4)	145	101	200	80
7	one	138	120	174	119

*Corrected to atmospheric pressure by the method of Hass and Newton, *Handbook of Chemistry and Physics*, 51st ed., edited by R.C. Weast, 1970.

Table 3: Melting and Boiling Points of Hetero-aromatic Compounds* (°C at 1 atm)

ring system (with location of substituent)	substituent										
	H	CH_3	C_2H_5	CO_2H	$CO_2C_2H_5$	$CONH_2$	NH_2	OH	OCH_3	Cl	Br
Benzene	80	111	136	122†	211	130†	184	43†	37†	131	155
Pyridine-2	115	128	148	137†	243	107†	57†	107†	252	171	193
Pyridine-3	115	144	163	235†	223	129†	65†	125†	179‡	150	173
Pyridine-4	115	145	171	306†	219	156†	157†	148†	93†	147	174
Pyrrole-1	130	114	129	95†‡	180	166†	—	—	—	—	—
Pyrrole-2	130	148	181	205†‡	39†	174†	—	—	—	—	—
Pyrrole-3	130	158	179	148†	78†‡	152†	—	—	—	—	—
Furan-2	31	64	92	133†	34†	142†	68†	80†	110	78	103
Furan-3	31	65	—	122†	179	168†	—	58†	—	80	103
Thiophene-2	84	113	133	129†	218	180†	214	217	154	128	150
Thiophene-3	84	115	135	138†	208	178†	—	—	—	136	157
Pyrazole-1	70†	127	137	—	213†	—	—	—	—	—	—
Pyrazole-3	70	205	209	214†‡	160†	—	285	164†	—	—	—
Pyrazole-4	70	207	—	275†	—	—	81†	118†	—	77†	97†
Isoxazole-3	95	118	138	149†‡	—	134†	—	—	—	—	—
Isoxazole-5	95	122	—	149†	—	174†	—	—	—	—	—
Imidazole-1	90	199	226	—	218†	—	—	—	—	—	—
Imidazole-2	90	141†	80	164†‡	—	—	—	250†‡	—	—	207†
Imidazole-4	90	56†	—	275†‡	157†	215†	—	—	—	—	130
Pyrimidine-2	123	138	—	270†	—	—	127†	320†	—	65†	—
Pyrimidine-4	123	141	—	240†‡	—	—	151†	164†	—	—	—
Pyrimidine-5	123	153	—	270†	38†	212†	170†	210†‡	—	—	75†
Pyrazine-2	57†	135	—	229†‡	—	189†	—	119†	187†	160	180

*Melting points above 30° C are indicated by a dagger†; those below 30° C are not included.
A dash indicates the compound is unstable, unknown, or that the data are not readily available.
‡Indicates that the compound melts with decomposition.

standpoint, therefore, the degree of aromaticity is measured by the relative tendency toward substitution rather than addition. Judging by this criterion, pyridine is more aromatic than furan, but it is difficult to say how much more aromatic it is.

Physical properties. Physical properties are important as criteria for judging the purity of heterocycles, just as for other organic compounds. The melting point was once the criterion most widely used, but now the optical spectra (based on light absorption), mass spectra (based on relative masses), and magnetic-resonance spectra (based on nuclear properties) have been increasingly used. Nevertheless, knowledge of the boiling point or melting point is still helpful in following purification of a compound. Organic compounds generally show great regularity regarding their physical properties, and heterocycles are no exception.

Melting and boiling points. The boiling points of certain saturated heterocycles compared with the corresponding cycloalkanes are listed in Table 2. The melting or boiling points of common hetero-aromatic ring compounds and their substituted derivatives are compared with those for benzene and its derivatives in Table 3. It can be seen from the table that replacement of a two-carbon unit (with molecule weight equal to 26) by a sulfur atom (atomic weight 32) has little effect on the melting or boiling point (compare corresponding benzene and thiophene compounds). On the other hand, replacement of a two-carbon unit by an oxygen atom (atomic weight 16) lowers the boiling point by about 40° C (compare corresponding benzene and furan compounds), which is to be expected because of the decreased molecular weight (lighter compounds being more volatile). The introduction of nitrogen atoms into the benzene ring is accompanied by less regular changes. Replacement of a two-carbon unit by an imino (NH) group, or of a single carbon by a nitrogen atom, increases the boiling point. Furthermore, if these two changes are made simultaneously, the boiling point is increased by an especially large amount, probably the result of association by hydrogen bonding (a weak form of attachment through certain types of hydrogen atoms) between the nitrogen atom and the imino group.

Effects of substituents

The effects of substituent groups in hetero-aromatic rings show considerable regularity. Thus, methyl and ethyl groups attached to ring carbon atoms usually increase the boiling point by about 20–30° C and 50–60° C, respectively, whereas conversion of an imino (NH) group into an amino (NR, in which R is a hydrocarbon chain) group (*e.g.*, pyrazole → 1-methylpyrazole) significantly decreases the boiling point because of decreased ease of association by hydrogen bonding (the active hydrogen having been replaced by a hydrocarbon group). Carboxylic acids and amides are all solids; carboxy derivatives of compounds containing a ring nitrogen atom usually melt at higher temperatures than those containing ring oxygen or sulfur atoms because of hydrogen bonding. Compounds containing both a ring nitrogen atom and a hydroxyl or amino group are usually relatively high melting solids. Chloro compounds usually have boiling points similar to those of the corresponding ethyl compounds.

Ultraviolet, infrared, nuclear magnetic resonance and mass spectra. Spectroscopic studies of heterocyclic

Table 4: Ultraviolet Spectral Characteristics of Hetero-aromatic Compounds

	neutral form		form with single positive charge			neutral form		form with single positive charge	
	wavelength (nm)	intensity ($\log_{10}\epsilon$)	wavelength (nm)	intensity ($\log_{10}\epsilon$)		wavelength (nm)	intensity ($\log_{10}\epsilon$)	wavelength (nm)	intensity ($\log_{10}\epsilon$)
Pyrrole	210	4.20	241	3.90	Pyridine	257	3.42	256	3.70
Furan	208	3.90	—	—	Pyridazine	247	3.04	238	3.21
						300	2.51		
Thiophene	231	3.87	—	—	Pyrimidine	243	3.51	242	3.64
Pyrazole	210	3.53	217	3.67	Pyrazine	261	3.77	266	3.86
						300	2.93		
Isoxazole	211	3.60	—	—	Indole	270	3.77	280	3.68
					Quinoline	275	3.51		
Isothiazole	244	3.72	—	—		299	3.46	313	3.79
						212	3.52		
Thiazole	233	3.57	—	—	Isoquinoline	306	3.38	270	3.30
						319	3.47	332	3.63

compounds, as have those of other organic compounds, have become of great importance as means of identification of unknown materials, as criteria of purity, and as probes for investigating the electronic structures of molecules, thereby explaining and helping to predict their reactions. The pattern of light absorption in the ultraviolet region of the spectrum (the so-called ultraviolet spectrum) of a compound is characteristic of the pi electron system of the molecule; *i.e.*, of the arrangement of double bonds within the structure. The ultraviolet spectra of hetero-aromatic compounds (given in Table 4) show general similarity to those of benzenoid compounds, and the effects of substituents can usually be rationalized in a similar way. The infrared spectrum of an organic compound, with its complexity of bands, provides an excellent "fingerprint" of the compound (far more characteristic than the melting point), and it can also be used to identify certain common groups, such as carbonyl (C = O) and imino (N — H) groups, as well as various of the heterocyclic ring systems (Table 5). Although magnetic resonance spectra were not widely used until about 1955, they have since become indispensable to any serious study of heterocyclic chemistry. Proton resonance spectra (the most common type), for example, yield information regarding the number of hydrogen atoms present in the molecule, as well as their chemical environment and their relative orientations in space (Table 6). Still later, mass spectra have been used to determine not only the complete molecular formula of the heterocyclic compound but also, from the way the molecule is fragmented, the arrangement of many of the atoms.

Measurements of pi electrons

Table 5: Ranges for Some Characteristic Infrared Absorption Maxima for Ring-Stretching Modes of Common Hetero-aromatic Compounds
(cm^{-1})

compound	absorption frequencies		
Pyrroles	1560–1530	1510–1480	1410–1390
Furans	1600–1560	1520–1470	1410–1370
Thiophenes	1540–1505	1440–1405	1370–1340
Pyridines	1610–1590	1580–1570	1485–1465

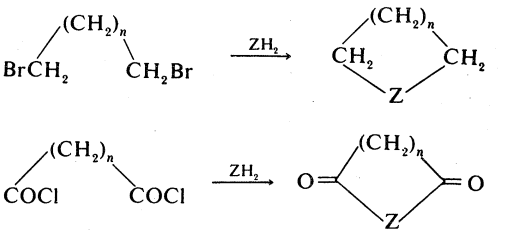
Table 6: Chemical Shifts in Proton Resonance Spectra of Various Heterocycles
(parts per million on delta scale)

compound	chemical shift at position		
	2	3	4
Aziridine	1.48	—	—
Oxirane	2.54	—	—
Thiirane	2.27	—	—
Azetidine	3.58	2.32	—
Pyrrole	6.62	6.05	—
Furan	7.40	6.30	—
Thiophene	7.19	7.04	—
Pyridine	8.50	7.06	7.46
Pyridazine	—	9.17	7.68

Synthesis and modification of heterocyclic rings. The important methods for synthesizing heterocyclic compounds may be classified under five headings, of which the first three are ways of forming new heterocyclic rings from precursors containing one less ring, the fourth covers the formation of a heterocyclic ring either from another heterocycle or from a carbocyclic ring, and the fifth includes the modification of substituents on a pre-existing heterocyclic ring. The key step in the formation of rings from noncyclic precursors (the first instance above) often is the formation of the carbon hetero-atom linkage (C — Z, in which Z represents an atom of nitrogen, oxygen, or sulfur). The actual ring closure, or cyclization, however, may involve the formation of a carbon–carbon bond. In any case, ring formation reactions are subdivided into three categories according to whether the cyclization reaction occurs primarily as a result of (1) nucleophilic or (2) electrophilic attack, or (3) by way of a cyclic transition state.

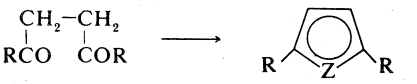
Classification by methods of ring closure

Nucleophilic ring closure. To prepare compounds containing one hetero-atom, a compound containing two halides (chloride [Cl], bromide [Br], or iodide [I]) or acyl halides (that is, halide derivatives of carboxylic acids) will react with the dihydro form of the hetero-atom, ZH_2 (or an equivalent reagent such as sodium hydroxide or sodium sulfide), to give nonaromatic heterocycles, as shown in the two examples below:



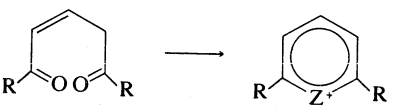
in which *n* equals only an integer

Diketones (ketones being compounds with oxygen atoms joined to carbon by double bonds) also can react with dihydro-Z compounds to give heterocycles. Diketones with the keto groups separated by two carbon atoms, for example, can be cyclized to form aromatic pyrroles, furans, and thiophenes as shown:



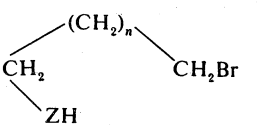
in which R is a hydrocarbon chain

With diketones separated by three carbons, such as the unsaturated compound below, six-membered rings may be formed:

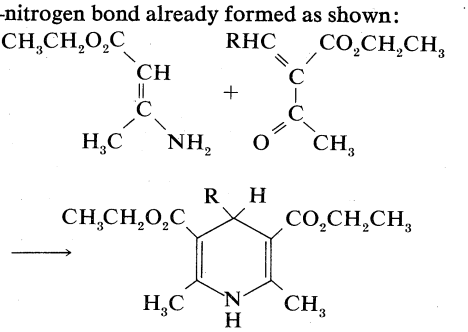


In each of these reactions the hetero-atom Z acts as a nucleophile in attacking the positively charged carbon atom produced by electron withdrawal due to the halogen atom (in the first instance above) or to the oxygen atom (in the second).

Usually, reactions like these proceed by means of intermediates in which only one of the C — Z bonds has been formed. In reactions of the first type in this section, for instance, compounds like the one shown below may be formed first.

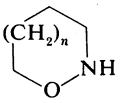


It is frequently possible to synthesize such intermediate compounds by other routes, and they then cyclize readily. One form of pyridine synthesis, for example, involves the condensation of an intermediate with the carbon–nitrogen bond already formed as shown:

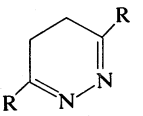


Heterocycles containing two adjacent nitrogen atoms and oxygen atoms, or both, also may be prepared from precursors of the type shown in the first example of this

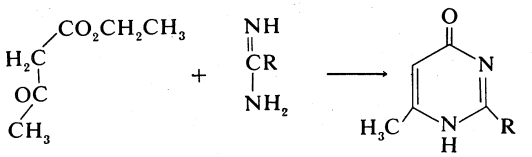
section by using hydrazine (N_2H_4), hydroxylamine (NH_2OH), or hydrogen peroxide (H_2O_2) in place of the dihydro-Z compound (ZH_2). With hydroxylamine, for example, the compound shown below is formed.



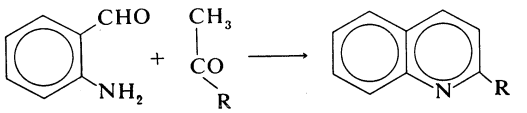
Similarly, two adjacent hetero-atoms can be introduced by reactions with diketones (the second reaction here), producing such compounds as the following:

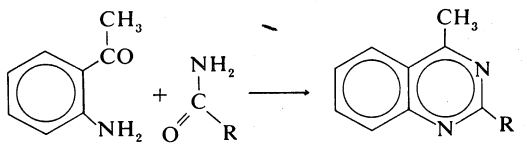


When it is desired to prepare a compound containing two nonadjacent hetero-atoms, appropriate components can be put together as illustrated in the following synthesis of pyrimidines:

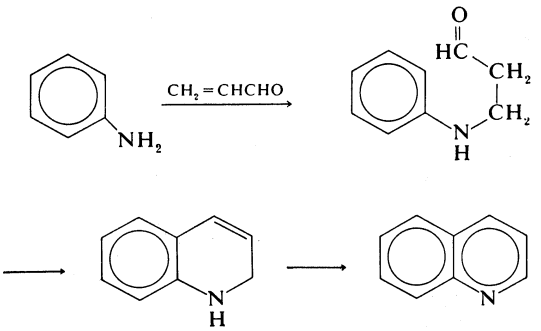


Ring synthesis reactions, in which the hetero-atom acts as a nucleophile, also involve orthodisubstituted benzenes (ortho substituents being on adjacent carbon atoms in the benzene ring), as in the preparation of many heterocycles fused to benzene rings. The formation of quinoline and quinazoline rings are examples of this type:



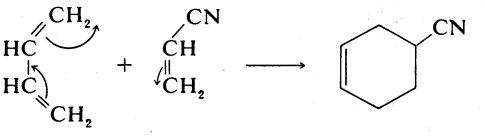


Electrophilic ring closure. Formation of heterocyclic rings by reactions in which the hetero-atom acts as an electrophilic (electron-seeking) reagent are rare, because the nitrogen, oxygen, and sulfur atoms are themselves electron-rich centres that react generally as nucleophiles. Electrophilic ring-closure reactions are known, however, in which a heterocyclic ring is formed by a reaction in which a carbon atom of the ring acts as an electrophile. Usually such reactions involve ring closure onto a benzene ring (or other aromatic system), an electron-rich system that is generally subject to attack by electrophilic reagents. An example of ring closures of this type is the formation of quinoline from aniline and acrolein, a dehydration product of glycerol. As shown below, the initial product of the reaction is a dihydroquinoline, which must be dehydrogenated to the fully aromatic product, quinoline itself:

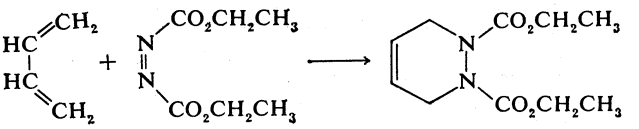


Ring closure by way of cyclic transition states. A most important method for the synthesis of carbocyclic six-membered rings is the so-called Diels–Alder diene reaction (named for its discoverers, two German Nobel Prize winning chemists, Otto Diels and Kurt Alder). In this reaction, as illustrated below, a diene (a compound with two double bonds) reacts with a dienophile (a diene-seeking reagent) to yield a cyclohexene:

The Diels–Alder reaction



In this reaction, it is not possible to determine whether the electrons move clockwise (as shown by the curved arrows) or whether they move in the reverse, counterclockwise fashion. For this reason, in reactions of this type the movements of electrons are thought to be synchronous rather than sequential, and the reactions are said to possess cyclic transition states. Heterocycles can also be prepared by the diene synthesis, as illustrated by this preparation of a tetrahydropyridazine:



Of far greater use, however, is the related method referred to as the Huisgen cyclic dipolar addition reaction; it is an important method for the preparation of many types of five-membered rings, especially those containing several hetero-atoms. Pyrazoles and isoxazoles, and many less common heterocycles, also are prepared by this method.

Conversion of one heterocyclic ring into another. There are many reactions of theoretical, and a few of practical, importance in which one heterocyclic ring is converted into another. The ring-atom exchange reactions of pyridine derivatives (see below, *Six-membered rings with one hetero-atom*) are good examples. In addition, ring-atom rearrangement or "shuffling" can be brought about with light (photochemically) in five- and six-membered hetero-aromatic compounds, and ring contraction by extrusion of an atom or group can occur under certain conditions.

Modification of an existing ring. Dehydrogenation of saturated or partially saturated rings to hetero-aromatic compounds by heating with sulfur or treatment with a palladium catalyst is analogous to similar reactions with carbocyclic compounds. The hydrogenation of hetero-aromatic rings is, by contrast, usually more difficult, for the hetero-atoms tend to poison the catalyst. Finally, the modification of substituents on heterocyclic rings is of highest importance in synthesis; reactions by which substituents may be altered are among the most useful in heterocyclic chemistry.

Nomenclature. Naming heterocyclic compounds is complicated because of the existence of many common names, in addition to the internationally agreed systematic nomenclature. A brief account of systematic nomenclature is given here, but for common names the reader is referred to the systematic survey in the next section.

Common and systematic nomenclature

The types of hetero-atoms present in a ring are indicated by prefixes; in particular "oxa," "thia," and "aza" denote oxygen, sulfur, and nitrogen atoms, respectively. The number of hetero-atoms of each kind are indicated by number prefixes joined to the hetero-atom prefixes as "dioxa" and "triaza." The presence of different hetero-atoms is indicated by combining the above prefixes, using the following order of preference: oxa first, followed by thia, then aza. Ring size and the number of double bonds are indicated by suffixes as shown in Table 7. In

Table 7: Suffixes Used in Naming Monocyclic Heterocycles

ring size	rings with nitrogen			rings without nitrogen		
	maximum unsaturated	one double bond	saturated	maximum unsaturated	one double bond	saturated
3	-irine	—	-iridine	-irene	—	-irane
4	-ete	-etine	-etidine	-ete	-etene	-etane
5	-ole	-oline	-olidine	-ole	-olene	-olane
6	-ine	—	—	-ine	—	-ane
7	-epine	—	—	-epine	—	-epane

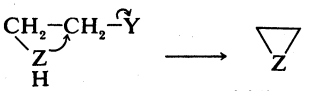
this table, "maximum unsaturation" refers to rings possessing the largest possible number of noncumulative (that is, with no atom in common) double bonds but with oxygen and sulfur atoms in their normal, divalent forms. In addition, partially saturated rings are indicated by the prefixes "dihydro," "tetrahydro," and so on, according to the number of excess hydrogen atoms. The positions of hetero-atoms, extra hydrogen atoms, and substituents are indicated by arabic numerals, for which the numbering starts at an oxygen atom, if one is present, or at a sulfur or nitrogen atom, and continues in such a way that the hetero-atoms are assigned the lowest possible numbers. Other things being equal, numbering starts at a nitrogen atom that carries a substituent rather than at a multiply bonded nitrogen. In compounds with maximum unsaturation, if the double bonds can be arranged in more than one way, their positions are defined by indicating the nitrogen or carbon atoms that are not multiply bonded, and consequently carry an "extra" hydrogen atom (or substituent), as follows: 1H —, 2H —, and so on.

Numbering systems

MAJOR CLASSES OF HETEROCYCLIC COMPOUNDS

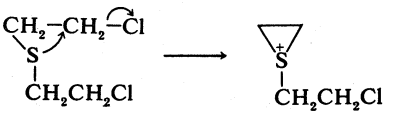
The major classes of heterocycles containing the common hetero-atoms (nitrogen, oxygen, sulfur) are reviewed first, in order of ring size (from smallest to largest), leaving compounds with other hetero-atoms to a final section. The classification by ring size is convenient because heterocyclic rings of a given size have many common features. For heterocyclic (as for carbocyclic) rings, the following broad generalizations can be made: 3- and 4-membered rings are strained and thus readily opened; they are also readily formed: such heterocycles are well-known reactive intermediates. Five- and six-membered rings are readily formed and are very stable: these sizes of ring also allow the development of aromatic character. Seven- and larger-membered rings are stable but not readily formed and are therefore relatively little investigated.

Three-membered rings. The three-membered-ring heterocycles containing single atoms of nitrogen, oxygen, and sulfur—aziridine, oxirane (or ethylene oxide), and thiirane, respectively—and their derivatives can all be prepared by nucleophilic reactions, of the type shown below.



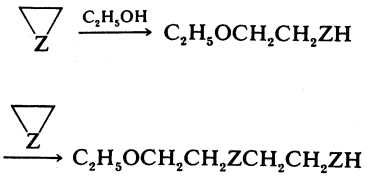
aziridine: Z=NH
oxirane: Z=O
thiirane: Z=S
(Y equals a group displaced by the heteroatom Z)

Thus, aziridine itself is formed by heating β-aminoethyl hydrogen sulfate with base (in which case Y is $-OSO_3H$). A reaction of this type is involved in the action of mustard gas, or dichlorodiethyl sulfide, a poison gas used in World War I. The mustard gas itself ring-closes to give a thiirane derivative, the biologically active agent, according to the reaction below.



Commercially, oxirane and aziridine (to a lesser extent) are important bulk industrial chemicals; oxirane is prepared on a large scale by the direct reaction of ethylene with oxygen.

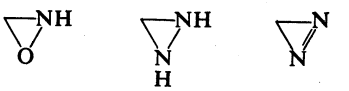
The chemical reaction characteristic of these three-membered rings is susceptibility to attack by nucleophilic reagents to open the ring as shown below:



The first step shown is the reverse of the formation reaction above. As indicated, a second molecule of the three-membered ring may react with the first product. Further reaction leads to long, chainlike molecules (polymers) of the type $C_2H_5O(CH_2CH_2Z)_nCH_2CH_2ZH$. Polymers and copolymers of oxirane and aziridine are useful plastics. Although aziridine, oxirane, and thiirane are heteroparaffinic, they are much more reactive than the corresponding normal amines, ethers, or sulfides because of the strain inherent in the three-membered ring. This behaviour reflects the comparable enhanced reactivity of the related carbocyclic compound cyclopropane.

Since 1950, three different classes of three-membered-ring compounds with two hetero-atoms have been discovered. They are derivatives of the parent ring systems oxaziridine, diaziridine, and diazirine, the chemical structures of which are shown below:

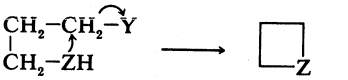
Three-membered rings with two hetero-atoms



oxaziridine diaziridine diazirine

These compounds have no practical importance, but they are stable enough for easy handling and offer interesting possibilities for future use.

Four-membered rings. Azetidine, oxetane, and thietane—the four-membered rings containing, respectively, nitrogen, oxygen, and sulfur atoms—are prepared by nucleophilic displacement reactions such as those used to prepare the corresponding three-membered rings, as shown:

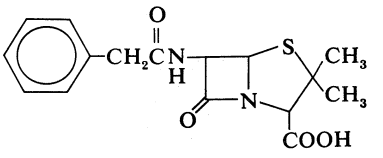


azetidine: Z=NH
oxetane: Z=O
thietane: Z=S

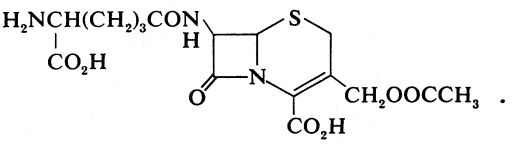
With the four-membered rings, however, the reactions proceed less readily than do the analogous reactions for the formation of the three-membered rings. Similarly, the ring-opening reactions of the four-membered heterocycles qualitatively resemble those of the corresponding three-membered rings, but they occur rather less readily. The most important heterocycles with four-membered rings are two related series of antibiotics, the penicillins and cephalosporins, both of which contain the azetidinone ring:



The chemistry of azetidinones, or β-lactams, as they are also called, was explored thoroughly during the intensive research into penicillin structure and synthesis that took place during World War II, a practical synthesis of penicillin not being achieved, however, until 1959. The complete chemical structures of a representative penicillin and a cephalosporin are shown below:

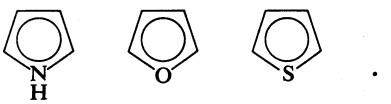


benzylpenicillin



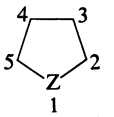
cephalosporin C

Five-membered rings with one hetero-atom. The parent aromatic compounds of the family—pyrrole, furan, and thiophene—have the structures shown:

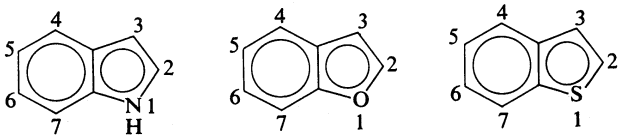


pyrrole furan thiophene

Five-membered heterocycles with one hetero-atom conform to the general structure:



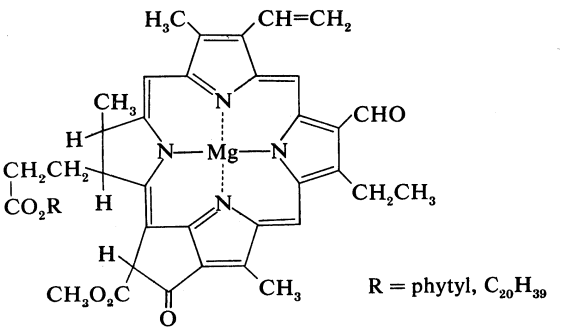
with numbering to indicate substituents as shown. The saturated analogues are called pyrrolidine, tetrahydrofuran, and thiophane, respectively. The bicyclic compounds containing a benzene ring fused to a pyrrole, furan, or thiophene ring are known, respectively, as indole, benzofuran, and thionaphthene; their structures are as shown:



indole benzofuran thionaphthene

The common numbering system is shown only for indole.

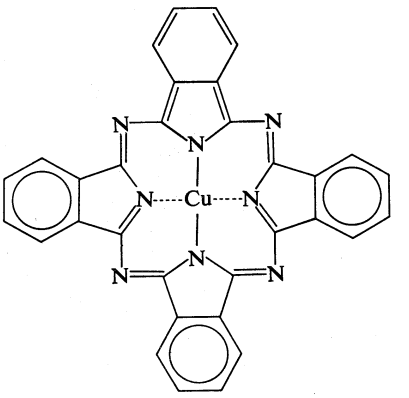
As mentioned earlier, pyrrole occurs in bone oil, in which it is formed by the pyrolytic decomposition of proteins (induced by strong heating). Reduced pyrrole rings are found in the essential amino acids proline and hydroxyproline, which are components of most proteins. Pyrrole derivatives are of widespread natural occurrence: the blood pigment hemoglobin and related compounds, the chlorophylls, and also vitamin B_{12} all are formed from four pyrrole nuclei arranged in a large ring, such as that of chlorophyll b, shown below:



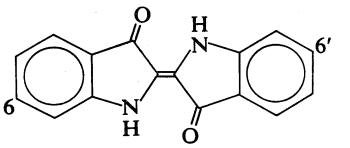
R = phytyl, $C_{20}H_{39}$

chlorophyll b
(plant pigment)

The phthalocyanins are a group of synthetic pigments that contain four benzopyrrole rings linked together in a large ring; a typical member of the family is Monastral blue with the structure shown:



The important vat dye indigo was formerly obtained from natural sources but is now synthesized on a large scale.

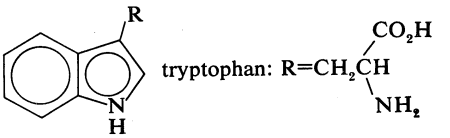


indigo

Tyrian purple, a dye used in classical times, is 6,6′-dibromoindigo (with bromine atoms at the numbered carbons).

Indole compounds

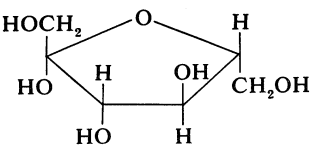
Other indole derivatives occur in nature. Tryptophan is an essential amino acid found in most proteins, and one of its degradation products is skatole, found in feces. Indole-3-acetic acid is a plant growth hormone, and many plant alkaloids contain indole rings. The structures of these compounds are:



tryptophan: R=$CH_2CH(NH_2)CO_2H$

skatole: R=CH_3
indole-3-acetic acid: R=CH_2CO_2H

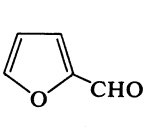
All carbohydrates, the biochemical family that includes the sugars and starches, are composed of one or more sugar (monosaccharide) units. These sugars are polyhydroxyaldehydes or -ketones (carbonyl compounds), which frequently exist as furanosides, that is, as a cyclized form with a five-membered, furanose, ring—*e.g.*, fructose, or fruit sugar—as the fructofuranose shown below:



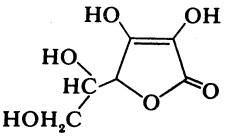
fructofuranose

Dehydration (removal of water) of certain carbohydrates yields furan derivatives; of great commercial importance is the conversion of a carbohydrate in corncobs into fu-

ran-2-aldehyde, or furfural, which is extensively used as a solvent in the manufacture of plastics and in the preparation of other furan derivatives. Many other reduced furans occur naturally, including vitamin C. The structures of furfural and vitamin C are as follows:

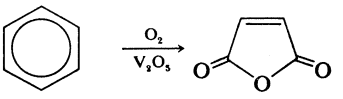


furfural

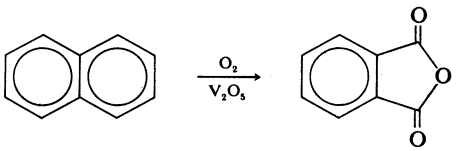


ascorbic acid (vitamin C)

Other derivatives of furan of industrial importance are the solvent tetrahydrofuran prepared from furfural, and maleic and phthalic anhydrides, important constituents of resins and plastics. These compounds are prepared in quantity by the oxidation of benzene and naphthalene, respectively, as shown:

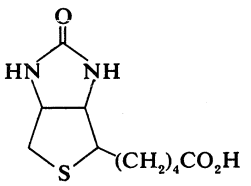


benzene maleic anhydride



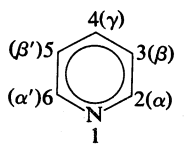
naphthalene phthalic anhydride

Thiophene and related compounds are found in coal tar and crude petroleum. The most important biologically occurring derivative is the vitamin biotin, with the structure:

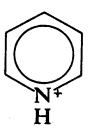


biotin (vitamin H)

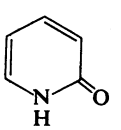
Six-membered rings with one hetero-atom. The nomenclature used for the various monocyclic nitrogen-containing six-membered ring compounds is as shown below, with numbering as shown for pyridine, Arabic numerals being preferred to Greek letters (though both systems are used).



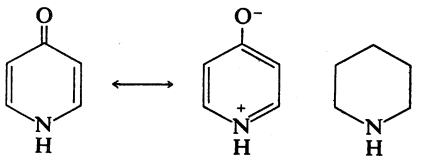
pyridine



pyridinium ion



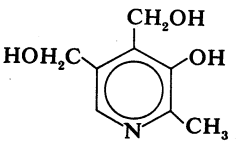
2-pyridone, 2(1H)-pyridone, or 1,2-dihydro-2-oxopyridine



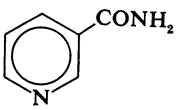
4-pyridone piperidine

The pyridones are aromatic compounds because of contributions to the resonance hybrid from charged canonical forms such as that shown for 4-pyridone. Mono-, di-, and trimethylpyridines are called picolines, lutidines, and collidines, respectively, with the position of the methyl groups denoted by the appropriate numbers; *e.g.*, 2, 4, 6-collidine. Pyridine-2-, -3-, and -4- carboxylic acids also have widely used trivial names; picolinic, nicotinic (derived from nicotine, of which it is an oxidation product), and isonicotinic acid, respectively. Pyridine and the picolines, lutidines, and collidines occur in coal tar and bone oil. Pyridine derivatives are also of great biological importance: vitamin B_6 is pyridoxine, and another B-group vitamin is nicotinamide. Their structures are as follows:

Pyridine and derivatives

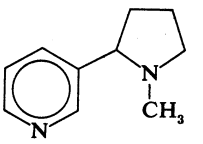


pyridoxine (vitamin B_6)

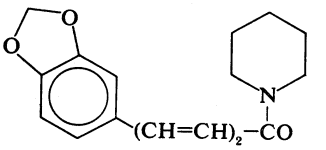


nicotinamide

Coenzymes I and II are derived from nicotinamide, and codecarboxylase from pyridoxine. Many alkaloids are derived from pyridine or piperidine, among them nicotine (in tobacco) and piperine (in pepper), with the structures shown.

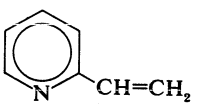


nicotine

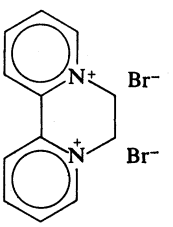


piperine

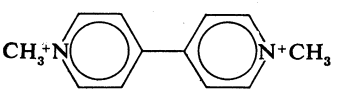
Pyridine, which is now prepared catalytically on a large scale from tetrahydrofurfuryl alcohol and ammonia, is an important solvent and intermediate used to make other compounds. Vinyl-pyridines, such as



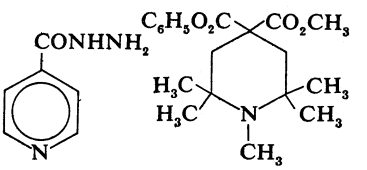
are important monomers used to make plastics. Diquat and paraquat are important herbicides, which selectively kill grassy plants. Pharmaceutically important pyridines include the tuberculostat isoniazid, the local anesthetic α-eucaine, the sulfa drug sulfapyridine, and the respiratory stimulant coramine. The chemical structures of these compounds are as follows:



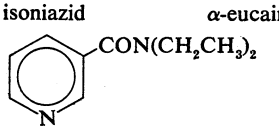
diquat



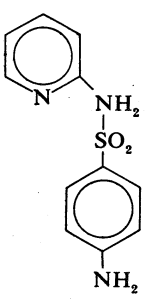
paraquat (herbicide)



isoniazid α-eucaine

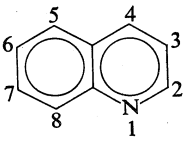


coramine

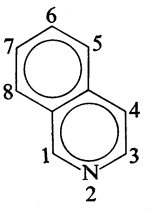


sulfapyridine

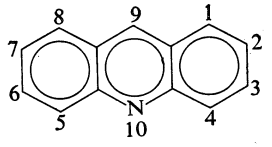
The structures of two isomeric benzopyridines and of two isomeric dibenzopyridines, with their common names and accepted numberings, are shown in the diagrams below:



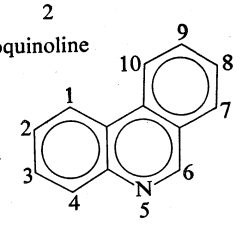
quinoline



isoquinoline

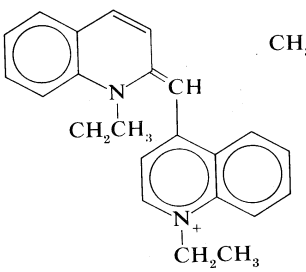


acridine

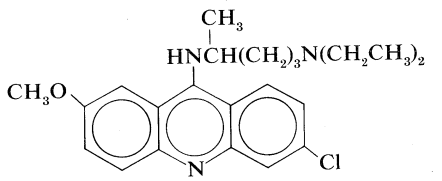


phenanthridine

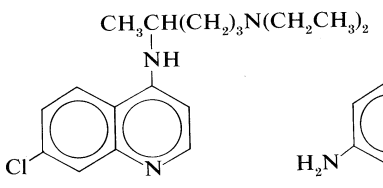
All four of these compounds and some of their alkyl derivatives have been obtained from coal tar. Each of them also is the parent substance of a class of alkaloids; of these, the quinoline (*e.g.*, quinine) and the isoquinoline (*e.g.*, morphine) groups are particularly well known. (For detailed coverage of this family of compounds, see the article ALKALOIDS.) Important synthetic derivatives of the benzo- and dibenzopyridines include members of the family of cyanine dyes, some of which—*e.g.*, ethyl red—are photographic sensitizing agents and others are useful antiseptics for staphylococcal organisms; antimalarials such as mepacrine (also called quinacrine or Atabrine) and chloroquine; antibacterials such as acriflavine (Trypaflavine); trypanocides such as phenidium chloride; and the reagent oxine (8-hydroxyquinoline or 8-quinolinol), used in analytical chemistry. The structures of these compounds are as follows:



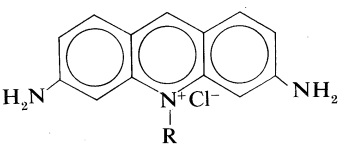
ethyl red



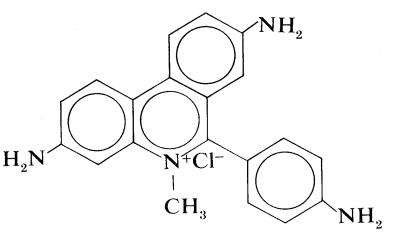
mepacrine



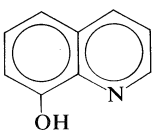
chloroquine



acriflavine (R = H or CH_3)

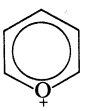


phenidium chloride

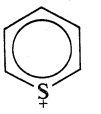


oxine

The parent six-membered, aromatic, monocyclic, oxygen and sulfur compounds are the pyrylium and thiopyrylium compounds bearing positive charges (cations) with the following structures:

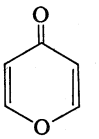


pyrylium cation

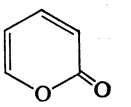


thiopyrylium cation

An uncharged aromatic (completely conjugated) six-membered ring containing an oxygen or sulfur atom is possible only if the ring contains a carbonyl group as in the pyrones:

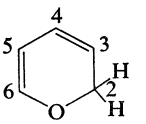


4-pyrone



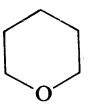
2-pyrone

The pyrans contain extra hydrogen atoms, the position of which is indicated by a number followed by an *H*, as shown:



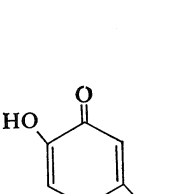
2[*H*]-pyran

Certain sugars are called pyranoses because they contain six-membered tetrahydropyran rings, with the structure:

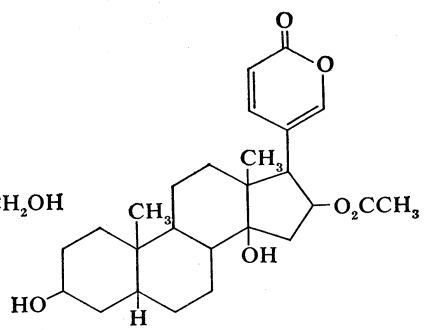


tetrahydropyran

Pyrone derivatives are present in natural products; *e.g.*, kojic acid, an antibiotic derived by the action of certain molds on starches or sugars, and bufotoxin—a poisonous ester of the steroid bufotalin—obtainable from the skin glands of toads. The structures of kojic acid and bufotalin are shown below:

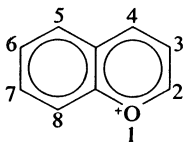


kojic acid

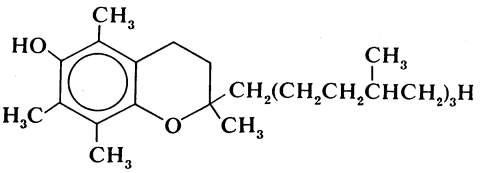


bufotalin

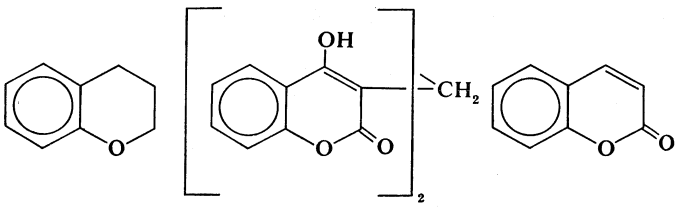
The benzopyrylium cation is the parent of a large number of natural products. Vitamin E is a substituted chroman, whereas dicoumarol, a blood anticoagulant, is a derivative of coumarin. The structures of these compounds are given below.



benzopyrylium ion

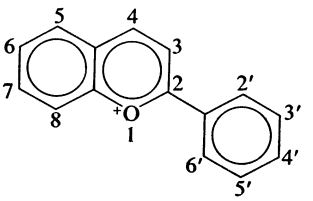


vitamin E



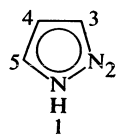
chroman dicoumarol coumarin

The flavylium cation is the parent of the anthocyanidines, substances that are combined with sugars to form the anthocyanin pigments, the common red and blue colouring matters of flowers and fruits. The flavylium ion has the following structure:

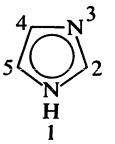


flavylium ion

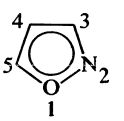
Five- and six-membered rings with two or more hetero-atoms. The names and numbering systems for the five-membered hetero-aromatic rings with two hetero-atoms are shown below:



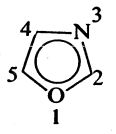
pyrazole



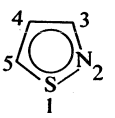
imidazole



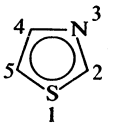
isoxazole



oxazole

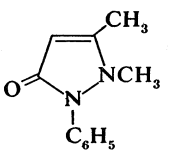


isothiazole

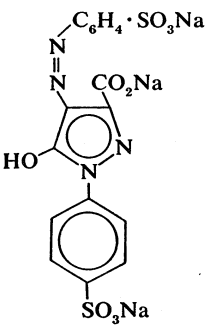


thiazole

Few pyrazoles occur naturally; the antipyretic (fever-reducing) and analgesic (pain-killing) antipyrine and the food and photographic dye tartrazine are important synthetic pyrazoles, with the following structures:



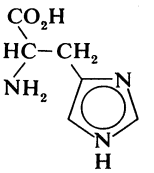
antipyrine



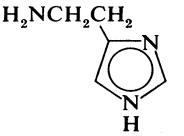
tartrazine

Imidazoles

Imidazoles are most important biologically; histidine, for example, is an essential amino acid, known to be of particular importance in enzyme reactions. A breakdown product of histidine, called histamine, is thought to be responsible for the development of allergies, hence the importance of antihistamine drugs. Histidine and histamine have the structures:

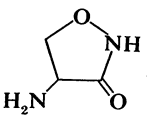


histidine



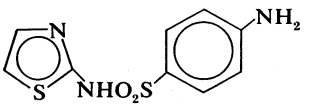
histamine

The antibiotic cycloserine, with structure shown, is one of the few naturally occurring isoxazoles:



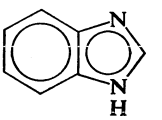
cycloserine

Thiazoles are of great biological importance. This ring system occurs in thiamine (vitamin B_1) and penicillin (see above). Sulfathiazole, an important sulfa drug, has the structure:



sulfathiazole

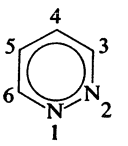
Most bicyclic systems derived from these five-membered rings are named systematically; *i.e.*, by use of the prefix "benz-" to indicate the presence of the benzene ring. Benzimidazole, for example, is the name for the compound:



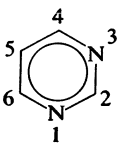
benzimidazole

A benzimidazole unit occurs in vitamin B_{12}.

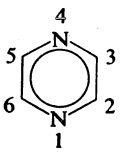
The three monocyclic diazines (six-membered-ring compounds with two nitrogen atoms) are named and numbered as follows:



pyridazine

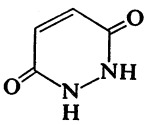


pyrimidine

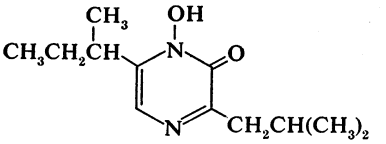


pyrazine

The pyridazine derivative maleic hydrazide is a herbicide, and some pyrazines occur naturally; *e.g.*, the antibiotic aspergillic acid. The structures of these compounds are:

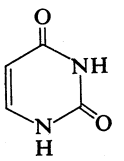


maleic hydrazide

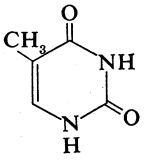


aspergillic acid

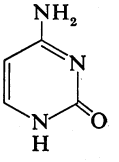
The pyrimidines, however, are the compounds of this group that are most important. Uracil, thymine, and cytosine, for example, with the structures shown, are constituents of nucleic acids.



uracil

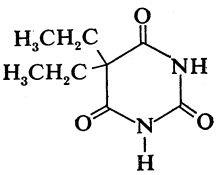


thymine

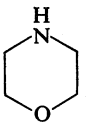


cytosine

Vitamin B_1 contains a pyrimidine ring, and synthetic barbiturates, such as veronal, with the structure given here, are widely used drugs. Various oxazines and thiazines are known but are of little importance, except for the solvent morpholine.

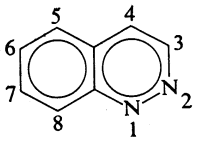


veronal

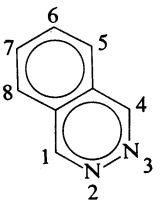


morpholine

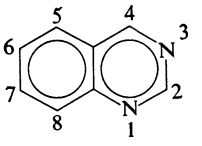
Many of the benzodiazines have common names; the structures and numbering of these compounds are as follows:



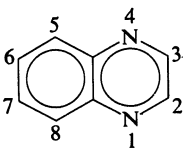
cinnoline



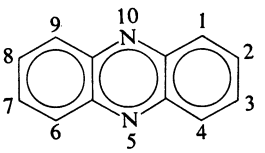
phthalazine



quinazoline

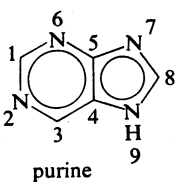


quinoxaline

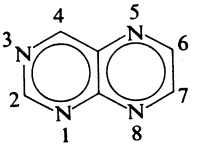


phenazine

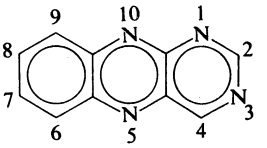
Five other polycyclic systems in this general family are of some significance; their structures and numbering system also are given:



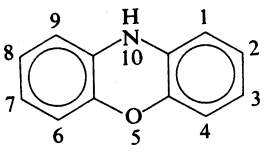
purine



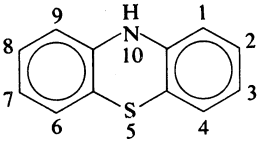
pteridine



alloxazine

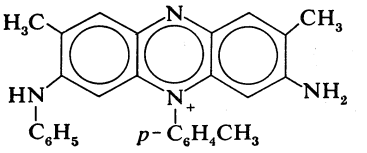


phenoxazine



phenothiazine

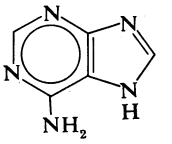
A few quinoxaline alkaloids exist; there are some phenazine natural products, and phenazine dyes are used for fabrics, the first synthetic dyestuff, Mauve, being historically important—its structure is as follows:



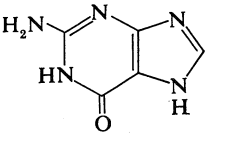
mauve

Purines and pteridines

Biologically, however, the purines and pteridines are of more importance. Two purines, adenine and guanine, occur, together with the pyrimidine bases mentioned previously, in all nucleic acids.

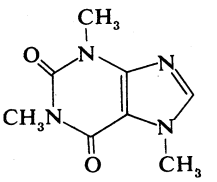


adenine

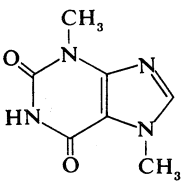


guanine

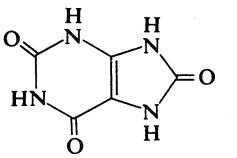
Other natural purines include caffeine (found in tea and coffee), theobromine (found in cocoa), and uric acid (a metabolic product). The structures of these compounds are as follows:



caffeine

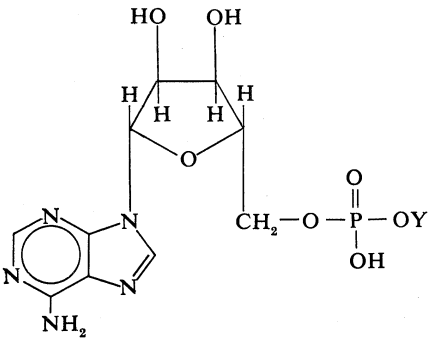


theobromine



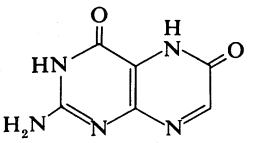
uric acid

Adenosine mono-, di-, and tri-phosphate (AMP, ADP, ATP, respectively) with the structures shown, are important in biological energy processes.



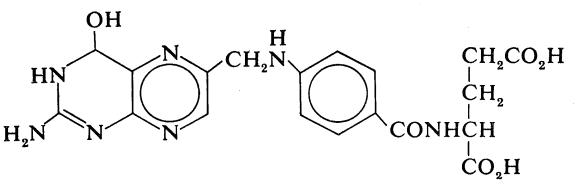
adenosine monophosphate: Y = H
adenosine diphosphate: $Y = PO_3H_2$
adenosine triphosphate: $Y = P_2O_6H_3$

The biological significance of pteridine compounds has become apparent. The first pteridines were discovered as pigments of butterfly wings; *e.g.*, one with the structure:



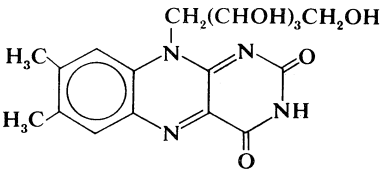
2-amino-4,6-pteridinedione

Folic acid, also a pteridine, is an important growth factor.



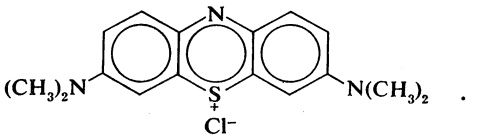
folic acid

Riboflavin, a derivative of alloxazine, is a B-vitamin:



riboflavin

Phenothiazine is an anthelmintic (worm-killing agent) as well as the parent of a family of dyes, such as methylene blue, a substance widely used as a biological stain and as an oxidation-reduction indicator. The structure of methylene blue is:

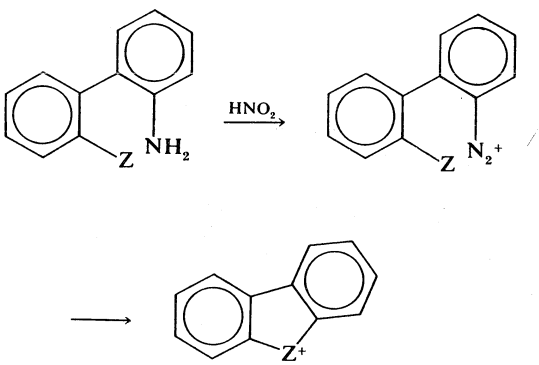


methylene blue

Rings with seven or more members. As the size of the ring increases, the range of compounds obtainable by variation of the number, type, and location of the hetero-atoms increases enormously. Nevertheless, the chemistry of heterocyclic compounds with seven-membered rings, or larger, is still relatively undeveloped, chiefly because the formation of such rings is generally difficult, just as is the formation of large carbocyclic rings. These larger rings, however, can be very stable, once they are formed, and knowledge of their chemistry is likely to increase rapidly in the near future. Some larger ring heterocycles exist in natural products; the porphyrins are widely distributed as pigments; *e.g.*, the chlorophylls and hemoglobins mentioned above.

Rings with uncommon hetero-atoms. Although nitrogen, oxygen, and sulfur are the common atoms found together with carbon in heterocyclic rings, a large number of other elements also form such rings—of greater or lesser stability. Such compounds are as yet of little practical importance. The principal ones known to date contain elements of the classes described below.

Halogens. Certain cyclic chloronium, bromonium, and iodonium ions have been prepared by the following reaction:



Z = Cl, Br, or I

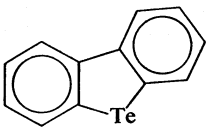
Of these, only the iodine derivative has much stability.

Selenium and tellurium. Many selenium heterocycles are known. Selenium shows much similarity in its behaviour to sulfur; hence selenophene, with the structure shown,



selenophene

resembles thiophene quite closely. Tellurium heterocycles are more rare, and less stable; an example is dibenzotellurophene:

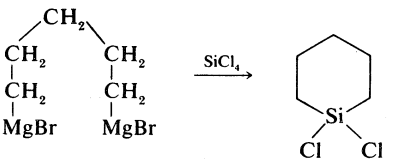


dibenzotellurophene

Phosphorus, arsenic, and antimony. Phosphorus, arsenic, and antimony, all members of Group V of the periodic table of elements, form a closely related group of heterocycles. There is, however, little similarity between their properties and those of the corresponding derivatives of nitrogen, a member of the same group. With few exceptions, no hetero-aromatic compounds are formed by these elements, and the cyclic derivatives resemble their open-chain analogues.

Silicon, germanium, and tin. Although silicon, germanium, and tin fall in the same periodic group as carbon, their atoms form neither stable chains nor double bonds; hence no hetero-aromatic derivatives of these compounds are known. Many saturated heterocyclic derivatives are known, however, including cyclic silanes, which may be prepared as shown by the following example:

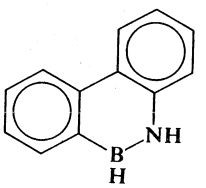
Silanes



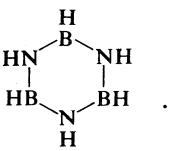
1,1-dichlorosilacyclohexane

All these elements, particularly silicon, readily form rings in which the atoms of these elements alternate with oxygen atoms. Considering this fact, many silicates are, strictly speaking, heterocyclic compounds, but their chemistry is always considered with that of other inorganic compounds, which they closely resemble.

Boron. Although boron heterocyclic chemistry is not yet extensively developed, it has a large potential. A boron and a nitrogen atom together contain the same number of electrons as two carbon atoms. Not surprisingly, it has been found that a boron-nitrogen unit can replace a carbon-carbon unit in benzenoid compounds to give stable hetero-aromatics; a good example is 9-aza-10-boraphenanthrene; its structure is as follows:



Borazine, or borazole, which has been referred to as inorganic benzene, has the following structure:



BIBLIOGRAPHY. Modern textbooks of heterocyclic chemistry include: A.R. KATRITZKY and J.M. LAGOWSKI, *The Principles of Heterocyclic Chemistry* (1967), also available in French, German, Spanish, Italian, Japanese, Russian, and Polish translation, which offers a concise comparative account; R.M. ACHESON, *An Introduction to the Chemistry of Heterocyclic Compounds*, 2nd ed. (1967), which places more emphasis on the division of the subject into compound classes; and A. ALBERT, *Heterocyclic Chemistry*, 2nd ed. (1968), a more detailed treatment. Of the numerous more specialized works available, K. SCHOFIELD, *Hetero-Aromatic Nitrogen Compounds: Pyrroles and Pyridines* (1967), treats the fundamental chemistry of two parent classes of compounds; *Advances in Heterocyclic Chemistry*, 16 vol., ed. by A.R. KATRITZKY and A.J. BOULTON (1963–73); and the *Chemistry of Heterocyclic Compounds*, ed. by A. WEISSBERGER and E.C. TAYLOR (1950–73), are continuing series of reviews.

(A.R.K.)

Heteroptera

Heteropterans, the "bugs" in the strictest sense, are considered a suborder of the Hemiptera by some entomologists, but an order (known either as Hemiptera or Heteroptera) by others. This large group of insects (about 30,000 species) can be recognized by an X-shaped design on the back, formed by the wings at rest. A combination of features—sucking mouthparts adapted to pierce plant or animal tissues and a hardened gula (underside of the head)—separate the heteropterans from all other insect orders. Although most species are terrestrial, a few are aquatic. Some species, which feed on plant juices, are serious pests of cultivated crops; other species are predacious and benefit man by destroying pests. There also are heteropterans that act as carriers of disease.

Heteropterans can be divided into three large groups on the basis of general habitat: the water-dwelling Hydrocorisae (water boatmen, back swimmers, water scorpions, giant water bugs, and creeping water bugs); the surface-swimming and shore-dwelling Amphibicorisae (water striders, marsh and water treaders, shore bugs, and velvet water bugs); and the Geocorisae, a large group of land bugs (plant bugs, bedbugs, assassin bugs, anthocorid bugs, lace bugs, ambush bugs, stinkbugs, burrower bugs, stilt bugs, and fire bugs).

Courtesy of G. Ferris and R. Usinger
Microentomology, vol. 4 (1939); Stanford University

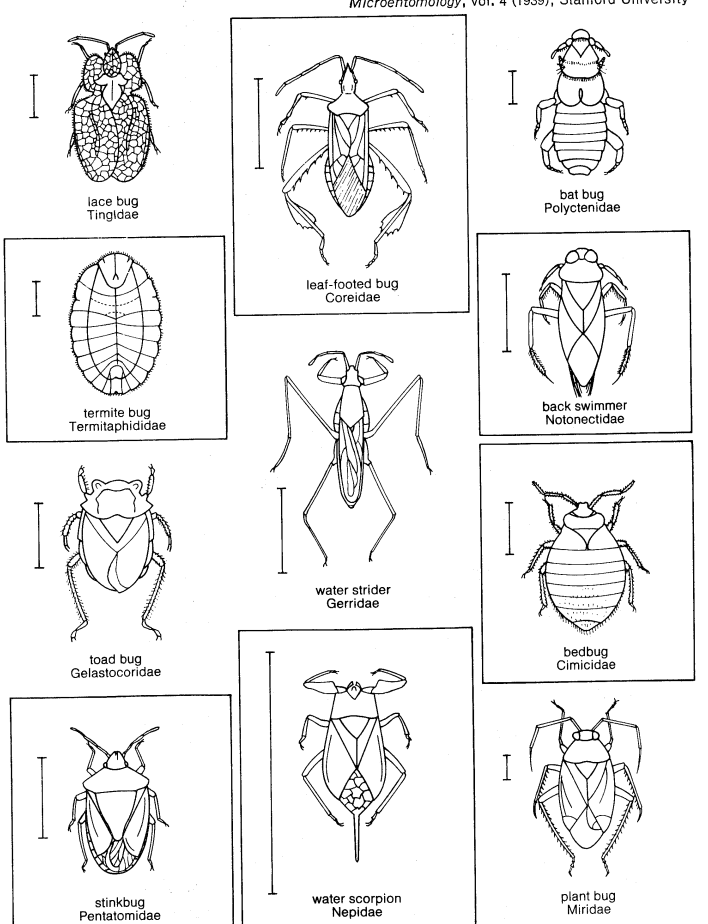


Figure 1: Diversity among heteropterans. Line scales indicate the approximate size of each insect.

GENERAL FEATURES

Size range and diversity. Like any other biologically successful group of organisms, the heteropterans are prolific and diverse and have adapted to a variety of habitats. They range in length from more than 100 millimetres to less than 1 millimetre and have invaded habitats from dry land to water. One of the few insect groups with aquatic adults capable of obtaining adequate oxygen from the water, heteropterans include the only insect species that spend their lives on oceans far from land. Some heteropterans are nourished by blood of animals ranging from mites to man; others feed from the sap of plants as diverse as giant redwoods and algae. Some suck surface fluids (*e.g.*, nectar); others pierce tissues to suck sap, blood, or even to obtain the nourishment provided by dried seeds. Many forms live on open surfaces and escape enemies by running, flying, or remaining inconspicuously motionless. Some seek food and shelter in natural crevices; other actively burrow into the soil or seek out the nests of animals.

Distribution and abundance. The number of individuals in a species depends in part on the availability of food. Growing crops provide concentrated food sources, but large populations, which often occur in areas of no apparent economic importance to man, may result from normal fluctuations in the balance of nature.

Heteropterans are most abundant in the tropics, decreasing in both individuals and species to limits northward beyond the Arctic Circle and southward almost to the Antarctic Circle. Every major land mass harbours different species; migrations to new habitats may be aided by natural agents (*e.g.*, wind, birds, floating debris) or by man. Although heteropterans have been carried throughout the world, only a few species have become established in many lands. Unique among insects are some water striders (Gerridae), which are at home on the open ocean between approximately 40° north latitude and 40° south latitude and may not approach land for several generations.

Approximately 30,000 species of heteropterans are known. Depending upon the availability of favourable habitats, the number of species on a major land mass is directly proportional to its size; the number of species inhabiting small islands is inversely proportional to their distance from large land masses. Age and origin of islands also influence the number of species. Most families containing 150 or more species are represented in every zoogeographic region; two exceptions are the Phymatidae (ambush bugs), with no species known in Australia or the Pacific Islands, and the Plataspidae, with no New World species. Families with the fewest species have the most restricted ranges: Aphylidae, Hyocephalidae, and Lestoniidae are restricted to Australia and contain among them only five species. The Velocipedidae of the Orient and the Vianaididae of the American tropics number four species each. In general, zoogeographic restriction is most evident at the subfamily, tribal, and generic levels.

IMPORTANCE

Heteropterans have complex and important roles in the balance of nature. The majority of them occupy an intermediate position in the ecological food chain; they use food producers (plants) and serve as food sources for parasites and other animals. A few species utilize plant-feeding insects as food.

Beneficial aspects. Heteropterans often serve as food for man and other animals. In some Latin American countries eggs of certain aquatic bugs are collected by providing submerged mats as egg laying sites; the eggs are then dried and made into cakes. Chickens, turkeys, hogs, and other domestic stock consume available heteropterans. Aquatic bugs are an important food source for fish. Wild birds and mammals, from the wrens and shrews to the turkey and grizzly bear, utilize available insects in their diets.

Species that attack unwanted plants and feed on the eggs, immature stages, and adults of injurious insects include most assassin bugs, nabid bugs, anthocorid bugs, stinkbugs, and plant bugs. These heteropterans are used in biological control of noxious weeds and injurious insects (see WEED CONTROL; PEST CONTROL).

Harmful aspects. In contrast, some heteropterans (*e.g.*, plant bugs, chinch bug, lace bugs, stinkbugs) damage both wild and domestic plants by sucking sap or injecting tissue-killing fluids. Insertion of a beak or eggs into a plant tissue harms the protective layers and allows bacteria and other disease-causing organisms to infect the

plant. Although heteropterans are not the major agents for transmission of plant viral diseases, a few species (Piesmatidae) can carry viral diseases of sugar beets and related crops in North America and Europe.

Role in disease

Heteropterans can affect man in several ways. They are common household pests and can spoil the taste of some fruits, (*e.g.*, raspberries contaminated by stinkbugs). But more importantly, some can attack man directly and inflict painful bites as well as introduce disease-causing organisms. The injection of saliva or poison may cause allergic reactions in susceptible persons. Sometimes when they cannot find plants, phytophagous insects probe moist surfaces (*e.g.*, perspiring skin) in search of appropriate food fluids. Transmission of trypanosomes, which cause Chagas' disease in the American tropics, occurs through cone nose bugs (Reduviidae), so-called because of the shape of their head. The insect receives trypanosomes when it feeds on the blood of an infected human. The trypanosome passes part of its life cycle in the insect and again becomes infective to man. Instead of injecting them into a new victim, however, the insect deposits trypanosomes (in its excrement) on the skin of a potential victim. The trypanosomes can enter the bloodstream only through mucous membranes or breaks in the skin (*e.g.*, those that may result from scratching the site of the bite).

NATURAL HISTORY

Life cycle. *General features.* Heteropterans are influenced by seasonal temperature fluctuations in temperate regions. Although the tropical regions, in which they are most abundant, impose no seasonal pattern, heteropterans are influenced by other factors—rainfall, host plant development, internal rhythms.

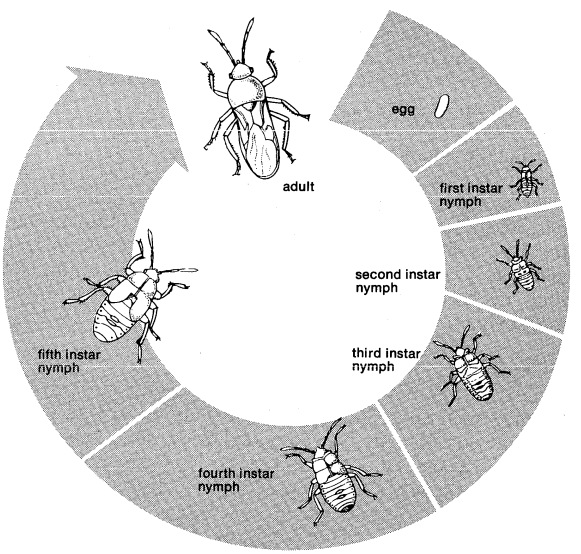


Figure 2: Life cycle of chinch bug.

Dormant stage

Heteropterans generally produce one or two generations a year in temperate regions. The dormant stage, which varies from species to species, is dependent in part on the food source of the species; for example, a heteropteran that feeds on perennial plants often lays eggs that remain dormant through the winter in or on the host plant. The nymphs, which hatch in the spring, then have appropriate leaves and succulent stems to furnish the sap necessary for growth. In contrast, heteropterans that feed on annual plants have a dormant winter adult stage that flies to a new host plant each spring after seeds germinate. Species that have short nymphal and adult stages, which coincide with certain plant characteristics (*e.g.*, short-lived flowers), may pass 10 or 11 months as eggs. Among predatory heteropterans, those whose prey is found on plants or in flowers winter as adults; those whose prey is found on the ground may be dormant as nymphs or adults. Aquatic heteropterans may spend the winter as adults, nymphs, or both; semiaquatic forms are dormant as adults.

Metamorphosis. Heteropterans undergo gradual metamorphosis (hemimetabola). Primary feeding and energy storage, as well as development of adult structures, take place in the nymphal stages. The adults seek mates and potential food sources and are responsible for initiating future generations. These functional specializations are reflected in morphology. The developing embryo, incapable of obtaining food or defending itself, is encapsulated within a protective egg shell that contains adequate food. The nymph, which lacks the locomotor and reproductive ability of the adult, also lacks the hard and rigid extensive exoskeleton necessary for strong muscle action. As a result, however, its body can distend to accommodate large quantities of food. In addition, loss of moisture through the thinner, more membranous body cover is offset by the high water content of the nymph's food.

Reproduction. *Eggs.* Reproduction occurs when eggs are fertilized as they pass through the female ducts by spermatozoa stored in special receptacles in the female. A notable exception to this method occurs in the bat bugs (Polyctenidae). In this case shell-less, yolkless eggs are fertilized in the female ovariole where they remain; the developing embryo is nourished by special maternal cells through a placenta-like structure consisting of tissues from both parent and offspring. The bat bug nymph leaves the mother's body only after reaching an advanced stage of development. Heteropteran eggs generally are moderate to large in size relative to adults and have a smooth to finely sculptured surface, sometimes with a colour pattern and often with slender projections. Eggs contain sufficient nutrients to permit the embryo to develop into a free-living, sexually immature, wingless nymph. Just before an egg hatches, the shell becomes translucent and reveals the segmentation and strongly coloured structures (*e.g.*, eyes) of the nymph inside.

Egg-laying sites

Eggs may be laid singly or in clusters. Eggs of plant-eating species generally are glued to a surface or inserted into the tissues of a selected host plant. Eggs of predatory forms often are laid near prey. Aquatic bugs may lay their eggs above or below the water's surface. Some bugs have special habits; for example, eggs may be glued in a large cluster to the back of a heteropteran male (Belostomatidae) or to the body of a crayfish as in Corixidae.

Nymphs. A newly developed nymph uses one or both of the following mechanisms to escape the egg: a cuticular spine on the head, sometimes known as an egg burster; or internal hydrostatic pressure created by forcing fluids (sometimes in the head) against the site of egg rupture. The pattern of rupture may be controlled by a line of weakness in the egg; in some cases a flaplike operculum lifts back to allow the nymph to escape. After withdrawing itself from the egg, the newly hatched nymph seeks the organic food essential for its development. Some nymphs leave the egg with a sufficient store of nutrients to allow them to pass through their first molt (or ecdysis). This is especially important to predatory species whose newly hatched nymphs might have to spend considerable time searching for prey.

The wingless nymph must rely on mimicry or disguise to prevent detection by enemies. Relatively inactive nymphs often assume colours and shapes of objects in their immediate environment; others disguise their bodies with a layer of small particles collected on a coat of sticky body hairs; and still others exhibit contrasting colours that break up the body outline. Some active nymphs confuse predators by alternating short quick dashes with periods of no motion; others deceive predators by simulating body shape, colour, or erratic gait of ants, with which nymphs sometimes are associated (see COLORATION, BIOLOGICAL; MIMICRY).

Lacking a pupal stage in which to change into an adult, the flightless nymph must sustain itself while it gradually transforms into an adult. As the nymph feeds and grows, it discards through five successive molts (ecdyses) the growth-restricting outer layers of its exoskeleton. Wings

Transformation to adults

appear first as slight outpocketings at posterior regions of the mesothorax and metathorax during the third instar (interval between molts) and enlarge during the fourth and fifth instars; however, the wings are not fully functional until the last ecdysis, when the nymph becomes an adult. At the time of the last molt, other changes occur: the head and thorax assume new shapes; the number of segments in antennae and tarsi may change; scent glands on the top of the abdomen cease functioning and are replaced by metathoracic scent glands; and the external genitalia and the internal reproductive organs become functional.

Adults. Flight is common among adult heteropterans; although they are capable fliers, no species has developed the aerial efficiency of some other insect groups (*e.g.*, dragonflies, flies). Predatory heteropterans lie in wait for a potential victim to wander close enough to be captured, instead of overtaking their prey in flight. Winged adults may rely on flight or inaction and the optical deceptions of colour arrangement and body form utilized by the nymphs to escape enemies. Flight also enables the two sexes to come together and provides an effective means for seeking favourable egg-laying sites.

Behaviour. *Sound production and reception.* Heteropterans cannot produce the conspicuous sounds typical of katydids, crickets, and cicadas. The whirring or buzzing noise common to heteropteran flight is caused by the rapidly moving flight mechanism. Sound production by specialized body parts (also called sonification or stridulation) is common among heteropterans but seldom loud enough to attract human attention. Heteropterans produce sounds by moving one roughened member over a roughened area of body surface. Sometimes both roughened parts consist of a series of minute, closely-set, parallel furrows or ridges called a strigose area, strigil, file, or rasp. Sometimes the second part of the mechanism consists of a series (called a plectrum) of minute pegs, setigerous tubercles, or an upturned edge of a sclerite (hard body plate). The plectrum may be on the movable body member or on the fixed body part. One or both sexes, and sometimes even the nymphs, of a species may stridulate. Some sounds appear to have courtship significance. The presence and form of sound-producing mechanisms are useful in the classification of Heteroptera.

Mechanisms for detecting airborne and possibly waterborne vibrations (or sounds) are known in Heteroptera: auditory organs known as Johnston's organs are found in the antennae of all species, and the tympanal (stretched membrane) organs are known in several families of aquatic bugs. A variety of receptors, especially for detecting substrate vibrations, have been described for members of this order.

Feeding. Heteropterans are essentially nonsocial insects. Newly hatched nymphs occasionally remain together for a short time, clinging to the shells of the eggs from which they have hatched; for example, plant feeding (phytophagous) nymphs may remain together because they hatch on their food source and need not search for it. In contrast, however, most predatory heteropterans encounter their prey by chance; in this case, dispersion increases individual opportunities for finding prey and for escaping other predatory heteropterans. Apparent coordinated migrations reported for chinch bugs generally result from a sudden disappearance or failure of a food source.

Ecology. Heteropteran behaviour, governed by natural instincts, includes patterns that enable individual species to live within specific environments. The patterns are so similar for all members of a family that each family can be characterized as essentially phytophagous or predatory, and terrestrial, semiaquatic, or aquatic. Most heteropterans rely on atmospheric oxygen for respiration.

Habitats. Aquatic Heteroptera vary in their independence of atmospheric oxygen. Members of one of the nonswimming families (Nepidae) submerge just below the water's surface and obtain oxygen through a slender, terminal breathing tube that breaks the surface film. Some swimming families (Corixidae, Notonectidae) dive deeply; however, they must surface periodically to replenish air supplies either stored in chambers formed against body surfaces by the folded wings or entrapped by rows of dense hairs folded against the body. These insects swim to overcome the buoyancy of the renewed air supply, grasping underwater objects to anchor themselves. Heteropterans with air storage chambers are independent of oxygen dissolved in the water; therefore, they survive in warm springs or polluted waters almost devoid of oxygen. Another family (Naucoridae) contains species that spend their entire lives below the water's surface; these heteropterans can obtain oxygen from that dissolved in the surrounding water. Water insulates its inhabitants from cold temperature extremes; thus aquatic species exhibit considerable activity the year around. Chilled water holds more dissolved oxygen than warm water, and air is available also between the water's surface and the ice layer during winter months in temperate regions.

Aquatic species

Semiaquatic bugs literally walk on water; they are supported solely by the water's surface film. Their small, slender bodies have little bulk to support. Long, thin middle and hind legs spread the points of contact with the water; the tarsus of each leg causes only a slight dent in the surface film. The shorter front legs may rest on the water or capture and hold prey. Wetting of the body during submersion is prevented by a thin layer of air trapped by the fine, dense, water-repellent (hydrofuge) pile that covers the insect. Many surface dwellers live in quiet water; others prefer flowing water, even rapids, where considerable strength is needed to overcome the current. In some cases, a bug may anchor the legs on one side of its body to some fixed object in a stream and wait for food carried to it by the rushing water. Adults survive the winter beneath objects on land either close to or some distance removed from a body of water.

Although shore bugs normally frequent shore lines, they occasionally make short forays onto the surface of adjacent bodies of water. Shore bugs generally are heavier bodied than surface dwellers; the strong, and often long, legs of some shore bugs enable them to run quickly across ground, rocks, and other exposed surfaces. Short rapid flights are not uncommon. All shore bug families are predatory. Their front legs often are thickened distinctly and sometimes spined for grasping prey. Adults generally overwinter beneath convenient objects.

Terrestrial species

Terrestrial heteropteran species are more abundant than either aquatic or semiaquatic forms and have a variety of habits and life histories. Some species wander over leaves, flowers, and other surfaces exposed to sunlight; others prefer the shadows (*e.g.*, underside of foliage). Many forms hide under rocks, logs, loose bark, or in crevices. Some species, which are active during the night, hide only during the day; other groups (*e.g.*, Aradidae) live in obscure habitats and leave only if the environmental conditions become unsatisfactory. Members of one family (Cydnidae) burrow by cutting soil in front of them and shoving it behind; these insects live in moving underground chambers. Species in at least two families (Nabidae, Reduviidae) live in spider webs and skillfully avoid entanglement as they eat trapped insects or, in some cases, the spider itself.

Relationships of heteropterans to plants vary. Plants may serve simply as a site on which to stand or move, or as a lure for plant feeding insects on which predatory bugs feed; plants are most important as a source of nourishing sap for phytophagous heteropterans. A few species can cause modifications in plant growth; for example, a feeding insect may cause leaves to roll up around it. In Europe certain lace bugs (Tingidae) insert their eggs into flower buds, which enlarge and house developing lace bug nymphs. A new adult lace bug needs no jaws to gnaw its way out; it leaves when the bud begins to open.

Food habits. Predation is common throughout the Heteroptera, sometimes as a family-wide feeding pattern, sometimes as a deviation within an otherwise phytopha-

gous family. In the suborder Hydrocorisae all families except one (Corixidae) are predatory on small aquatic animals, including small fish, tadpoles, and frogs; the exception has been known to add mosquito larvae to its usual algal diet. The suborder Amphibicorisae, without reported exception, contains forms that prey on plankton, small arthropods, or worms. In the suborder Geocorisae, many families of the series Cimicomorpha (*e.g.*, Reduviidae, Nabidae, Phymatidae), are predatory, generally on other insects; however, a few assassin bugs (Reduviidae), bedbugs (Cimicidae), and bat bugs (Polyctenidae) seek the warm blood of birds and mammals. In addition, some members of plant feeding families (including the Miridae) attack other arthropods; a few families (Pentatomidae, Lygaeidae) of the series Pentatomomorpha, fundamentally plant feeders, contain insect-eating species. The original food habit of this order is not yet clear, but even in some avidly predatory species, the first and second nymphal instars may be plant feeders. In capturing its prey, a heteropteran may use front legs modified for grasping or insert feeding stylets. In either case the victim is paralyzed quickly by an injected salivary secretion.

Predators

Any insectan stage may fall to the hungry predator. Insect eggs are a favourite food for many small (Anthocoridae) or very young heteropteran predators. Caterpillars and other larvae commonly are attacked by predatory Heteroptera, and it is estimated that one stinkbug (Pentatomidae) may destroy more than 200 caterpillars during its lifetime. One stinkbug, long known to feed upon bedbugs, was recommended for use in control as early as 1776.

Other arthropods also may be attacked. One anthocorid bug (Anthocoridae) may destroy an average of 33 mites per day. In some instances size of prey is no barrier. In Sri Lanka a small assassin bug (Reduviidae) only 4 millimetres long can paralyze a giant millipede 150 millimetres long.

Associations. Individual heteropterans may associate accidently on a food source. On quiet pools of water, water striders (Gerridae, Veliidae) often assemble in milling "schools" of a few to hundreds of individuals; they quickly reassemble if dispersed. Sometimes a female will reassemble disturbed batches of eggs. In cases where the female lays her eggs on the back of a male, the parental responsibility is his.

Members of one family (Termitaphididae), unable to live alone, share the nests of certain termites and feed on the fluids of the fungi growing there. In exchange, a substance secreted from pores on the backs of these bugs is eaten by the termites. Reports of heteropteran associations are difficult to assess and may reflect coincidental use of the same environment.

FORM AND FUNCTION

General features. Heteropterans have features typical of all insects and special adaptations that differentiate them from other insect groups. The head tests new environments through light perceptions by compound eyes and simple eyes (ocelli) and through both scent and sound perceptions by the antennae. The mouthparts are located conveniently for quick exploitation of newly discovered food sources. The thorax bears support and transportation mechanisms, legs, and wings. An important function for the third body region, the abdomen, is reproduction. Sutures divide the body surface into sclerites (hard plates), which may be fused or separated by flexible membranes. Hairs, scales, or a variety of glandular secretions may adorn at least some areas of the body. The body surface may be raised into spines of a variety of shapes, knobs, or tubercles; tubercles often are setigerous; *i.e.*, glandular with a straight or curved filament or bristle at the tip. Some of these projecting structures, which have bases associated with specialized cells, probably have a sensory function and respond to vibrations of air, substrate, or touch. Certain long hairs, each arising from a ringed socket, are called trichobothria, and their presence and arrangement on various parts of the body have taxonomic significance.

Special adaptations

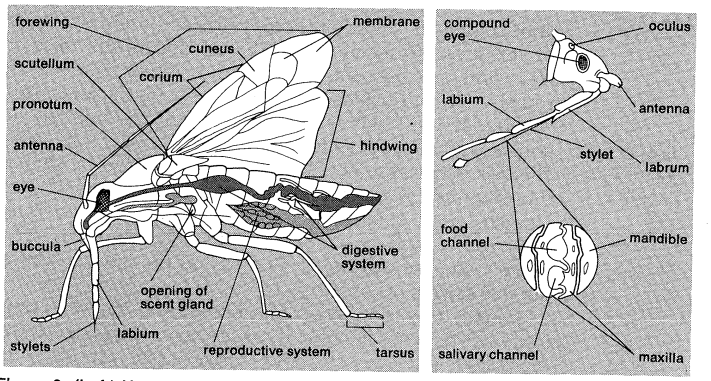


Figure 3: (Left) Heteropteran body plan. (Right) Section through stylet showing relationships of mandibles and maxillae.

Adapted from (left) H. Weber, *Grundriss der Insektenkunde*, Gustav Fischer Verlag, Stuttgart; (right) *Introduction to Comparative Entomology* by R. Fox and J. Fox. Copyright © 1964 by Litton Educational Publishing, Inc.

Distinguishing features. The covering of the head is a sclerotized ring tapering from the flexible neck forward or downward and ending in three lobes: a jugum on each side and a central clypeus (or tylus). The area above and between the eyes (called the vertex) may bear a transverse pair of simple eyes (a median one is never present in Heteroptera). The hard lower surface of the head is called the gula (hypostomal bridge).

Mouthparts. The piercing-sucking mouthparts are composed of a troughlike, four-segmented labium in which lie four stylets; these are modified mandibles and maxillae. Each of the hairlike maxillae has two major grooves plus minor grooves and ridges along its median surface. When brought together and locked by the minor grooves and ridges, the two major grooves form the left and right halves of two separate tubular canals that extend the length of the maxillae. The anterior canal is the food tube through which fluids are drawn by a sucking pump (described below); the posterior tube is the salivary canal through which digestive and certain other fluids may pass into a food source. These fluids stun or kill the prey, prevent coagulation of blood, and initiate the process of digestion. Along the sides of the maxillary stylets are slender mandibles with serrated tips. The mandibles and maxillae alternately advance into the food tissue until the appropriate fluid is reached by the maxillae; then ingestion begins. Since the labium does not penetrate the food tissue, the stylets must extend beyond the protection of the labial groove. This extension may be accomplished in one of two ways. The labium may be shortened by jackknifing some of its segments, leaving only the apical and sometimes the basal segments to guide the exposed stylets. In some species the extensible stylets are much longer than the labium; when not extended, the stylets either are coiled or looped in a chamber located in the labium or the head, or are found in a long sheath that extends backward into the heteropteran's body cavity.

Sucking pump

The sucking pump is a chamber whose front wall is folded lengthwise into the mouth cavity; the fold is pulled out and springs back to change the size of the cavity. Contraction of certain muscles increases the size of the cavity, decreases the pressure in it, and allows the liquids to flow through the food canal into the pump. When the muscles relax, the size of the chamber decreases gradually and pushes the food toward the true mouth, which opens into the esophagus.

Thoracic features. Each of the three somites of the thorax bears a pair of legs; both meso- and metathorax bear a pair of wings. The large conspicuous part of the thorax between the head and the wings is called the pronotum (*i.e.*, the expanded top part of the prothorax). Projecting from beneath the hind margin of the pronotum is the triangular, sometimes U-shaped, scutellum. Along the side margins of the scutellum may be a groove or fold (the frenum). In repose, the inner edges of the front wings hook onto the frenum. On each side of the

metathorax may be a pore (ostiole) that provides an outlet for a highly volatile, repellent fluid secreted by internal scent glands. The ostiole may be surrounded by an elevated ridge (peritreme), and then by an extensive, roughened, evaporative area.

Legs. The legs may be modified in various ways to serve specific functions. Legs may be lengthened or expanded to aid in camouflage. Sometimes the hind femurs and their accompanying leg muscles are adapted for jumping. The middle and hind legs may be flattened for swimming, or their effective width increased by a marginal fringe of long hairs that spread during a power stroke and flatten against the leg on the recovery stroke. The front legs, when specialized for grasping prey, have tibia and tarsus adapted to fold tightly against a thickened femur; stout hairs or spines along one or both of the opposing surfaces of this grasping leg prevent a victim from slipping free. The males of many species have hind legs that are thickened or spined to hold females during mating. Lacking mandibulate mouthparts with which to clean their antennae, many Heteroptera have a comblike series of hairs on the end of the front tibia, across which the antennae may be dragged. Certain genera (Nabidae, Reduviidae) that live suspended slothlike from the underside of spider webs have long, foldable claws.

Wings. The texture of each of the two pairs of wings is distinct in heteropterans. The mesothoracic or forewings (called hemelytra, elytra, or tegmina) are stiff and have an oblique line that abruptly separates the leathery basal half from the membranous apical half; the metathoracic or hindwings are thin, delicate membranes. At rest the forewings are folded over the hindwings; the membranous tips of the forewings overlap, so that the fragile hindwings lie beneath them. In macropterous individuals with wings of normal size, the wings approach or extend beyond the apex of the abdomen; some species have wings that are brachypterous (slightly to moderately reduced), micropterous (very small), or absent. Both pairs of wings are absent in one family (Termitaphididae), apparently an adaptation to the termite tunnels in which they live. In the bat parasite family (Polyctenidae) the budlike character of the nymphal wings is carried into the adult stage with no further development; in effect these insects also are wingless. The front pair of wings of the bedbug family (Cimicidae) are reduced to short, nonfunctioning pads; the hind pair is lost completely. Although these heteropteran families are the only examples of a general pattern of wing loss or reduction, some individual species show a variety of wing developments (alary polymorphism) or aptery (winglessness) in one or both sexes. Species with sexual dimorphism in wing length have females with reduced wings or none at all; apparently the retention of flight in the male assures his ability to move to the female.

Diversity of wings

Heteropteran wings, especially the front pair, may be modified for specific functions. Their presence above the abdomen provides protection for the soft parts beneath. More effective protection is provided in forms with fused forewings. Some forms can couple their wings by means of a fingerlike projection (jugum) on each forewing, thus enhancing their flying ability. The colour and texture of the forewings may be modified so that an insect resembles bark; bright or contrasting wing colours announce unpalatability to an enemy. The lace bugs (Tingidae), so-called because they have a lacelike pattern of many fine, interconnecting wing veins, resemble the fine veins on the undersides of the leaves on which they feed.

Wing modifications

Abdomen. The abdomen contains 11 segments; the tenth and eleventh, and sometimes others, are fused. Sensory appendages (cerci) are never present in this order; the sensory function may be performed in part, however, by several apparently tactile hairs along the underside of the abdomen. The terminal abdominal segments of the male are adapted primarily for transfer of sperm, those of the female for reception of sperm and egg laying. Females that insert eggs into plant tissue have an ovipositor under the tip of the abdomen; the ovipositor consists of blades that slice plant tissue and guide the eggs into the slit. In females that glue eggs to a surface, the ovipositor usually is reduced to a noncutting egg guide.

Internal features. *Digestive system.* The three embryological regions of the alimentary canal become modified by constrictions, dilations, and outpocketings. The stomodaeum becomes a short esophagus that connects to the midgut. The midgut frequently differentiates into four distinct regions, two dilations connected by a short, narrow tube and a second tubular section that gives rise to slender, terminally closed, tubular outpocketings, called mycetomes. The cavities of the mycetomes are filled with bacteria, whose role may be either to supply essential nutrients or to inhibit development of other bacteria. The bacteria, hereditarily transferred from mother to offspring, migrate to the ovarian follicle of the female and invade the eggs. As soon as an embryo begins to develop a gut, the bacteria appear in it.

Role of bacteria in mycetomes

Respiratory system. The respiratory system consists of longitudinal tracheal trunks that branch internally and commnunicate with the external air through ten pairs of holes called spiracles. Respiration under water presents special problems. Young aquatic nymphs may respire exclusively through the thin body wall. Adult aquatic insects, with their hard body walls, must rely on tracheal respiration; modifications of the spiracles and external skeleton surrounding them serve to exclude water while admitting oxygen. In some species the spiracles are sunken and protected by a circlet of hydrofuge (water repellent) hairs or by a fine-pored membrane; in other aquatic heteropterans the spiracles are closed and only one or two pairs function in association with either diving air storage chambers or the base of a snorkel-like breathing tube that breaks the surface film of the water. Although air storage chambers absorb some dissolved oxygen from the surrounding water, they gradually diminish in size as the nitrogen contained in them dissolves outward into the surrounding water. Unless they are replaced by frequent trips to the surface, air supplies may be lost completely, thus allowing water to reach and enter the spiracle; drowning would result. A few forms (Naucoridae) have "plastron" respiration. The plastron is actually a modified air storage chamber that consists of air in a series of cuticular grooves radiating from a spiracle. This air, continuous with tracheal air, is held in each groove by hydrofuge hairs, so that its volume does not change. Sustained absorption of dissolved oxygen by the plastron eliminates the need for contact with atmospheric air as long as the surrounding water is well oxygenated.

Other systems. The skeletal, muscular, circulatory, and excretory systems are structurally and functionally typical of a winged insect with the exception of some adaptive variations. The piercing-sucking mouthparts require protractor and retractor muscles for insertion and withdrawal of each stylet, and dilator muscles to enlarge the sucking pump cavity. General blood flow throughout the body is provided by the four- or five-chambered heart, but circulation in legs and wings requires special devices (*e.g.*, the pulsating membranes located within the base of each tibia and sometimes in the tarsus). Rhythmical waving of pulsating membranes causes a pattern of blood flow to the tip of the limb and back. The number of excretory tubes (*i.e.*, malpighian tubules) generally is four.

The nervous system consists of three interconnected parts known as central, visceral (sympathetic), and peripheral parts. The ventral nerve cord exhibits considerable concentration near the head end (cephalization) of the ganglia; in all species the abdominal ganglia have migrated into the thorax. Most members of the order have three ganglia on the ventral nerve cord; the subesophageal, the first thoracic, and a mass that results from fusion of the second and third thoracic ganglia with all the abdominal ganglia. Complete cephalization of the ventral nerve cord occurs in a few Heteroptera where all the ventral ganglia, including the subesophageal, unite into a single ganglionic centre.

The heteropteran reproductive system consists of a pair of gonads whose ducts unite to form a single tube leading to the exterior near the posterior end of the abdomen. In

Reproductive system

most species spermatozoa are transferred directly into the female reproductive tract and stored in a spermatheca (outpocketing of female tract) until fertilization occurs (as the egg is laid). In two families (Cimicidae, Anthocoridae) the female has an organ separate from the reproductive tract to receive the spermatozoa. This organ is a rounded internal pouch associated with a slit on the underside of the abdomen and is called the organ of Ribaga. During mating the spermatozoa are deposited in this pouch. They then penetrate the pouch wall, travel through the body cavity, and burrow into the spermatheca, remaining there until needed to fertilize the eggs. Excess spermatozoa are absorbed as nutrients by special cells in the female.

EVOLUTION AND PALEONTOLOGY

The fossil record is too poor to offer significant help in reconstructing the phylogeny of the Heteroptera; their small size, the fragile nature of dead individuals, and a preference for habitats that seldom are conducive to fossilization have resulted in few fossils. Often only wings or parts of wings are represented. Known fossils do reveal, however, that during the Permian Period (more than 225,000,000 years ago), two major lines existed. The order Protohemiptera, which revealed a number of features common to homopterans and heteropterans, including stylet-like mouthparts assumed to be piercing-sucking, actually represents an unrelated parallel development.

The appearance of the Heteroptera as a line distinct from the Homoptera also is represented in Permian fossils. The oldest known member of the order Heteroptera (from the Permian in Australia) represents the only family (Paraknightiidae) with no living species. Fossils from the next geological period, the Triassic, reveal that the major forking of the heteropterans into aquatic and terrestrial branches had already occurred near the beginning of the Mesozoic Era. Several modern aquatic family types are recognizable in the Mesozoic. The bulk of modern terrestrial heteropterans appears in the fossil record by the Tertiary Period of the Cenozoic Era.

CLASSIFICATION

Distinguishing taxonomic features. Modifications of mouthparts and wings are generally accepted as important characteristics by which to divide the class Insecta into orders. One peculiar type of mouthpart, possessed by more than 50,000 insect species, is the piercing-sucking type in which the labium forms a trough containing two pairs of threadlike stylets (highly modified mandibles and maxillae). All modern insects with mouthparts of this type are closely related to each other but not to other insects, none of which possess truly comparable modifications. Some entomologists consider all insects with piercing-sucking mouthparts as members of a single order, the Hemiptera (sometimes called Rhyngota or Rhynchota), with two suborders Homoptera and Heteroptera separated traditionally by texture and resting position of the forewings and by the apparent origin of the beak; other entomologists, while recognizing the proximity of relationship between these two groups, consider that the relationship is of superorder value and that features separating the two groups are of sufficient magnitude to warrant full ordinal status for each group.

Separation of the two groups is reflected by differences in several important functioning body parts. The gula of the Heteroptera is a hard bridge that separates the mouthparts from the prothorax. In the homopterans the gula is small and membranous or absent, allowing the base of the mouthparts to abut or fuse to the prothorax. The forewings of heteropterans are generally divided into basal and apical halves of different texture. Homopteran forewings, on the other hand, are generally of similar texture for their full length. Scent glands are present in all terrestrial and some aquatic nymphs and adults of the Heteroptera. Homopterans lack scent glands.

Reduction and loss of the gula in homopterans, coupled with the migration of the mouthparts to the prothorax, were significant changes in functional specialization. In some homopteran families of the suborder Sternorrhyncha, the mouthparts have fused to the prothorax and remain attached to it if the cranium is removed. These and other basic characters that began to differentiate in the geologically remote Permian time (when many other insect orders were evolving) indicate an important branching of the phylogenetic line of homopterans and heteropterans and substantiate their separation at the ordinal level.

The heteropterans are divided into suborders on the basis of habitat, antennal structure, ability to produce sound (stridulate), and presence and arrangement of long hairs (called trichobothria) on various parts of the body. Division into families is based on structural differences and adaptations for specialized ways of life.

Annotated classification.

ORDER HETEROPTERA (true bugs)

Wings, when present, usually number four that may vary in size; basal portion of forewings thick and leathery, apical portion membranous; hindwings membranous, slightly shorter than forewings; wings at rest held flat over abdomen; piercing-sucking mouthparts arise from front part of head; compound eyes usually well developed; two simple eyes, or ocelli, present or absent; scent glands usually present; gradual metamorphosis; widely distributed, terrestrial and aquatic species number about 30,000.

Suborder Hydrocorisae (or Cryptocerata)

Neither cephalic nor abdominal trichobothria; antennae 4-segmented, shorter than head, usually in grooves on underside of head; semiaquatic (Gelastocoridae, Ochteridae) or aquatic (all other families); swimming members with fringe of swimming hairs on hind legs; aquatic members lay eggs in or on submerged exposed objects, are relatively active under water during winter; probably all males, and some females, stridulate; important food for fish, some water birds.

Family Corixidae (water boatmen). Head overlaps front of prothorax; scutellum small, usually concealed by pronotum; front legs short, 1-segmented tarsus with fringe of long hairs beneath for gathering food; middle legs long, slender, for submerged anchoring; males with stridular pegs on front femurs and tibiae of legs, a polished black strigil above on asymmetrical abdomen; females nonstridulatory, abdomen symmetrical; colour pattern commonly alternating light and dark transverse stripes; inhabit fresh, brackish, or salt (Great Salt Lake) waters; use diving air stores; phytophagous on algae and diatoms, but may eat insect larvae; more than 300 species; in all major faunal regions.

Family Notonectidae (back swimmers). Head not overlapping prothorax; scutellum large, exposed; all legs long and slender, hind pair longest; male stridulating mechanism involving front legs and labium; inhabits fresh and mineral waters; diving air stores trapped under hairs on underside of abdomen, hence inverted position in swimming; bite painful to man; predatory on insects, small crustaceans, fish, and tadpoles; almost 200 species; in all zoogeographical regions.

Family Pleidae. Very small, stout, convex above; scutellum large, exposed; hindwings reduced or absent; head and pronotum partially or wholly fused; inhabit fresh waters; about 20 species; in all zoogeographic regions.

Family Helotrephidae. Like Pleidae; about 15 species; only in Asia, Africa, and South America.

Family Nepidae (water scorpions). Body stout to very slender; legs long, generally slender, front pair adapted for grasping; abdomen of most nymphs and adults with snorkel-like breathing tube at tip; nymphs and adults stridulate by opposable files on anterior coxae and coxal cavity plates; inhabit fresh water; swim poorly or not at all; predatory; principal food, other insects; nearly 200 species; in all zoogeographical regions.

Family Belostomatidae (giant water bugs and toe biters). Large, some species exceeding 109 millimetres (4 inches) in length; broad, flattened body; middle and hind legs markedly flattened, with swimming hairs; inhabits fresh water; may lie in mud or debris on bottom; can inflict severe bites; predatory; aggressive, strong swimmers; grasping front legs capture insects, fishes, frogs; approximately 100 species; in New World, Asia, and Africa (including Madagascar).

Family Naucoridae (creeping water bugs). Head large, inserted into thorax; scutellum large, exposed; front femurs enlarged, tibiae opposable; middle and hind legs slender, sometimes with fringing hairs; inhabit fresh water, often warm

springs; generally use diving air storage chambers or "plastron" respiration; predatory mostly on mollusks and insects; about 150 species; most common in tropical continents, ranging north into Europe and North America; subfamily Aphelocheirinae uses plastron respiration, sometimes elevated to family rank.

Family Gelastocoridae (toad bugs). Broadly oval in shape, roughened surface, elevated protuberant eyes; resemble tiny toads; beak reaching only to coxae of front legs; broadened front femurs and opposable tibiae form effective grasping organ; middle and hind legs slender for running, lack fringing hairs; inhabit shores, burrows into mud or sand to lay eggs; predatory principally on small insects; about 100 species; in all zoogeographic regions.

Family Ochteridae (velvety shore bugs). Similar to Gelastocoridae in general appearance; body surface smooth, with velvety pile; beak reaches abdomen; found among plants growing along margins of ponds and streams; predatory upon other insects; about 25 species; in all zoogeographic regions.

Suborder Amphibicorisae

Trichobothria on head but not on abdomen; members generally live on the water's surface film or on mud or sand; predatory on tiny animals, including injured or freshly-killed insects; nonstridulating except for both sexes of genus *Stridulovelia* (Veliidae); within one species individuals may vary from winged to wingless.

Family Gerridae (water striders; pond skaters). Slender bodied; legs long, slender, front pair short, used for support or grasping prey; middle pair longest, used for propulsion over water surface; hind femurs extends beyond tip of abdomen; tips of tarsi project beyond insertion of claws; body covered below with hydrofuge hairs; on quiet or flowing fresh or marine water; *Halobates* occurs on open oceans hundreds of miles from land; predatory generally on insects; oceanic forms feed on plankton and floating jellyfishes; about 350 species; in all zoogeographic regions and intervening oceans between 40° north and 40° south latitudes.

Family Veliidae (smaller water striders). Habits and structures similar to Gerridae, but generally smaller, stouter bodied, and with hind femurs not reaching beyond tip of abdomen; two genera with last segment of middle tarsus split and with collapsible, fan-shaped tuft of long hairs used for propulsion over water surface; about 300 species; in all zoographic regions, genus *Trochopus* marine, but not in open sea.

Family Hydrometridae (marsh treaders). Body, head, and legs long and slender; claws arise at tip of tarsus; wings present or absent; crawl weakly and clumsily among emergent plants in fresh water; predatory; more than 100 species; mostly tropical, a few species in North America and Eurasia.

Family Mesoveliidae (water treaders). Small; body and legs slender; live on fresh water, especially among emergent plants or on saturated mud; one New Guinea species lives on the forest floor away from water; about 30 species; in all zoogeographical regions.

Family Saldidae (shore bugs). Head large; eyes prominent; labium apparently 3-segmented, long, first segment very short and stout, second very long; legs long and slender for running or leaping; generally inhabit land surfaces near fresh or salt water; some species live on rocks in the intertidal zone, hide in air trapped in crevices when rocks are covered at high tide; predatory on small or crippled insects; over 200 species, in all zoogeographic regions.

Family Leptopodidae. Like Saldidae; all 24 species native to Old World, one Mediterranean species established in New World (California).

Family Hebridae (velvet water bugs). Tiny, covered with hydrofuge hairs suitable to habitat on surface or along edges of bodies of still, fresh water; sometimes crawl into the water; food habits not known; about 120 species; in all zoogeographic regions; some widely distributed species may have been carried in drinking water on early sailing ships.

Family Leotichiidae. Structure suggestive of Saldidae but distinguished by strong median ridge on pronotum; habits unlike Hebridae; one of the two known species (both from southeastern Asia) occurs in caves.

Suborder Geocorisae

Trichobothria either present on abdomen or absent from both abdomen and head; antennae longer than head; difficult suborder to define; perhaps a convenient ecological (rather than a taxonomic) category better replaced by elevation of the two series to suborders; terrestrial; habits and foods varied.

Series Cimicomorpha

Trichobothria absent from head and abdomen; front wings with or without a cuneus (special region on hardened basal portion known as the corium).

Family Miridae (plant bugs). Small, fragile; ocelli absent; labium 4-segmented; front wing with cuneus, membranous area with 1 or 2 closed cells; legs slender to stout, for walking, running, or jumping; predatory, phytophagous, or both; often host specific; includes major crop pests (*e.g.*, "lygus bugs" of the Northern Hemisphere, the "tea blight" [*Helopeltis*] in southeastern Asia), also several important predators of small insects; about 8,000 species; in all zoogeographic regions; found north of the Arctic Circle.

Family Cimicidae (bedbugs). Oval to elongate oval, very flat; eyes small; forewings reduced to small, transversely oval pads exposing most of abdomen; hindwings absent; blood-sucking parasites of mammals and birds, usually present on host only while feeding, otherwise hiding in crevices in nest or lair; eggs glued to walls of crevices; contains the two species of bedbugs that attack man, *Cimex lectularius*, (originally in Europe and North America, now in all parts of the world) and *Cimex hemipterus* (common in southern Asia and Africa); *C. lectularius* probably encountered cave-dwelling men; not yet conclusively proved as carriers of any human diseases even though the bug can be successfully infected with some human parasites; about 75 species; found throughout the world but show marked parallels with ranges of their hosts.

Family Reduviidae (assassin bugs). Head usually protruding; beak short, 3-segmented, strongly curved away from underside of head, with tip in deep stridular groove between front coxae; front legs adapted for grasping; prey on insects and other small animals; genus *Triatoma* and some close relatives suck blood of mammals (including man) and are carriers of the tropical American Chagas' disease; can inflict painful bites; nearly 4,000 species; on all significant land masses.

Family Anthocoridae (anthocorid bugs). Ocelli and cuneus present; membrane of wing with or without veins; male external genitalia asymmetrical; predatory, mostly on small arthropods, including their eggs; some eat ants; others inhabit bird nests; may bite man; about 400 species; in all zoogeographic regions.

Family Polyctenidae (bat bugs). Oval; flattened; eyeless; front forewings, rudimentary, unhinged; hindwings absent; head and certain other parts with transverse rows of spines forming ctenidia (combs); eggs hatch within female, young nymphs retained in body and nourished by mother through a pseudoplacenta; blood suckers live entire life in fur of certain bats; 18 species; found from southern North America to South America, in Africa, and from southern Asia to Java.

Family Aradidae (flat bugs). Extremely flat; surface usually rough; head horizontal; clypeus prominently convex; ocelli absent; legs short; tarsi two-segmented; abdomen much wider than wings when present; found in crevices, as under loose bark or debris on the ground; feed on molds or fungi; nearly 1,000 species in all zoogeographic regions; sometimes part of group separated as family Dysodiidae.

Family Tingidae (lace bugs). Head small, without ocelli; antennae slender to stout, generally with first two segments very short, third long; pronotum and forewings often finely to coarsely reticulate (lacy); tarsi two-segmented; phytophagous, on anything from mosses to flowering plants, often host specific; several species are common pests on maize, cotton, cacao, and other cultivated plants; almost 2,000 species; in all zoogeographic regions.

Family Nabidae (nabid bugs). Ocelli present; beak four-segmented, long, generally curved away from underside of head; no striated groove between front coxae of legs; front legs more or less thickened for grasping; forewings without a cuneus; predatory on softbodied insects; frequently investigated as a possible biological control agent for insect pests of field crops; about 300 species; in all regions of the world.

Family Phymatidae (ambush bugs). Stout; surface roughened by tubercles; beak short, three-segmented, curved away from underside of head, tip resting in stridular groove between front coxae and legs; fourth antennal segment conspicuously thickened; front femurs of legs strongly thickened; scutellum small and triangular or large, broad, and reaching tip of abdomen; predatory, generally, lying among flower petals awaiting prey; more than 200 species; found on all major continents, except Australia, not found on oceanic islands.

Family Enicocephalidae (unique-headed bugs). Predatory; about 130 species; in all zoogeographic regions.

Family Dipsocoridae (or Cryptostemmatidae or Ceratocombidae). Biology poorly known; about 200 species; in all zoogeographic regions, more than half of species found in tropical America.

Family Isometopidae. Phytophagous; less than 50 species; sometimes reduced to a subfamily in the Miridae.

Family Joppeicidae. The sole species occurs in the eastern Mediterranean region.

Family Microphysidae. Predatory; about 25 species; found in Eurasia and North America; one African genus questionably placed here.

Family Schizopteridae. About 100 species; sometimes united with the Dipsocoridae (see above).

Family Termitaphididae. Lives with termites; includes 9 species; found in tropical America, Africa, India, and Australia.

Family Thaumastocoridae (or Thaumastotheriidae). Phytophagous; about 12 species; found in Australia, India, North and South America, and the Greater Antilles.

Family Velocipedidae. Predatory; four species; found in Orient.

Family Vianaididae. Lives with ants (myrmecophilous); four species; found in tropical America.

Series Pentatomomorpha (or Trichophora)

Trichobothria on abdomen; cuneus always absent; terrestrial; various habits and foods.

Family Pentatomidae (stinkbugs). Head with side margins generally expanded over bases of antennae; ocelli usually present; antennae usually five-segmented; scutellum large to very large, triangular or U-shaped; without claval commissure; tibiae without rows of spines; phytophagous or predatory, sometimes within same species; several important pests, the "southern green stinkbug" (*Nezara viridula*), a serious pest of fruit and vegetable crops throughout the world, the "harlequin cabbage bug" (*Murgantia histrionica*) on North American cruciferous crops, the senn bugs (*Eurygaster* species) on small grains in the Old World; several genera (*Podisus, Zicrona*) feed on caterpillars and other larvae; about 5,000 species; in all zoogeographic regions; taxonomy unsettled, especially assignment of taxonomic levels; definition and relationships of subgroups not in dispute.

Family Plataspidae. Oval to transversely oval; head small, inserted deeply into prothorax; scutellum large, broadly U-shaped, reaching apex of abdomen; forewings much longer than body; wing membrane transversely folded at base and inserted under scutellum; tibiae not spined; phytophagous; about 500 species; found only in Old World.

Family Cydnidae (burrower bugs). Scutellum triangular or broadly U-shaped; forewings without claval commissure (except genus *Amnestus*); tibiae with rows of spines; colour generally brown or black, sometimes with creamy markings; phytophagous, most species in soil feed on roots of host; more than 600 species; in all zoogeographic regions; subfamily Corimelaeninae (or Thyreocorinae) sometimes given family status.

Family Coreidae. Antennae four-segmented, inserted on or above imaginary line connecting middle of eye to front end of buccula (cheek plates); ocelli present; forewings with claval commissure; wing membrane with numerous interconnected veins; appendages often with leaflike expansions; phytophagous, some as pests of cultivated plants, for example, the "squash bug" (*Anasa tristis*) on cucurbit crops in North America, a rice bug (*Leptocorisa varicornis*) in the Orient; over 2,000 species; in all zoogeographic regions; subfamilies Alydinae and Rhopalinae (or Corizinae) sometimes given family status.

Family Berytidae, or Neididae (stilt bugs). Body slender, narrowly elongate; legs extremely long, linear; labium four-segmented; usually phytophagous, sometimes attack soft-bodied plant-feeding insects; more than 100 species; in all zoogeographic regions.

Family Piesmatidae. Juga (two outer lobes of head) extend as a pair of horizontal, hornlike projections; prothorax with a deep pit below expanded sides; forewings, except apical part of membrane, with numerous coarse punctures; phytophagous, most frequently on plants of the goosefoot family (Chenopodiaceae); proven vectors of viral diseases of sugar beets in Europe and North America; about 20 species; in all zoogeographic regions.

Family Lygaeidae. Antennae four-segmented, inserted on side of head below an imaginary line connecting middle of eye to front end of buccula; ocelli present except in brachypterous forms; forewings with a claval commissure; wing membrane with four or five simple, longitudinal veins; phytophagous (on sap or seeds) or predatory (on mites and small insects); pests include the "chinch bug" (*Blissus leucopterus*) of North America, the Old World "cotton stainer" (*Oxycarenus hyalinipennis*), and several species of "false chinch bugs" in the widely distributed genus *Nysius;* in North America *Geocoris,* considered beneficial, feed upon mites and small insects attacking crop plants; almost 3,000 species found in all parts of the world, including many small islands.

Family Largidae. Head triangular, without ocelli; forewings with claval commissure; female with sixth visible abdominal segment below cleft medially; phytophagous; habits not well known; of no economic importance; about 100 special; in all zoogeographic regions.

Family Pyrrhocoridae (fire bugs or cotton stainers). Head triangular, without ocelli; forewings with long claval commissure; female with sixth visible abdominal segment below entire; phytophagous or predatory; "cotton stainers" (*Dysdercus*), serious pests of cotton crops; some species (*e.g.,* genus *Dindymus*) feed on flies as adults, nymphs prefer termites; more than 300 species; in all zoogeographic regions.

Family Aphylidae. Two species in Australia.

Family Colobathristidae. Phytophagous; about 90 species; inhabit Neotropical, Oriental, and Australian regions; unknow in North America, Europe, and Africa.

Family Hyocephalidae. One species in Australia.

Family Lestoniidae. Two species in Australia.

Family Phloeidae. Three species in tropical America, one in Borneo.

Family Stenocephalidae. About 40 species; except for one species on the Galapagos Islands, all known species live in Eurasia and Africa.

Family Urostylidae. Approximately 50 species confined to Asia and Australia.

Critical appraisal. The classification of the Heteroptera is in a state of flux due to continuous investigations in ontogeny, genetics, ecology, and morphology. Not all results are in agreement; therefore, important changes probably will occur. Wide disparity in superfamily concepts has resulted in dropping them from the above classification of the heteropterans.

The division of the order into three suborders is generally accepted, even if only for convenience; the prefixes reflect the broad ecological habits of the families included. Hydrocorisae reflects the adaptation of most of its families (Gelastocoridae and Ochteridae excepted) for life under water. Amphibicorisae suggests the land and water habits of these insects that live on the surface of the water or along shores. Geocorisae reflects the preferred terrestrial habitats of members of this group.

BIBLIOGRAPHY. Heteroptera literature is voluminous but scattered. Recent comprehensive works are too few and too limited in scope to supersede all the older works. The reader will find additional important, useful, and interesting information on this order in textbooks on insects listed in the article INSECTA.

W.S. BLATCHLEY, *Heteroptera of Eastern North America* (1926), introduction to Heteroptera, especially those of eastern North America; E.A. BUTLER, *A Biology of the British Hemiptera-Heteroptera* (1923), classic work with biological data; W.E. CHINA and N.C.E. MILLER, "Checklist and Keys to the Families and Subfamilies of the Hemiptera-Heteroptera," *Bull. Br. Mus. (Nat. Hist.) Ent.*, 8:1–45 (1959), phylogenetic dendrogram and assignment of names, includes a bibliography of source studies; R.H. COBBEN, *Evolutionary Trends in Heteroptera, part 1, Eggs: Architecture of the Shell, Gross Embryology and Eclosion* (1968), structural and developmental details with interpretive comments; W.L. DISTANT, *Rhynchota,* in *Fauna of British India* (1902–18), five-volume study of Asian Heteroptera, *Biologia Centrali Americana Insecta,* vol. 1, *Rhynchota Heteroptera* (1881–1909), classic study of tropical American fauna; P.T. HASKELL, *Insect Sounds* (1961), readable, though in places technical, introduction to a fascinating topic; H.B. HUNGERFORD, "The Biology and Ecology of Aquatic and Semiaquatic Hemiptera," *Kans. Univ. Sci. Bull.*, 11:1–265 (1919), classic and basic study; N.C.E. MILLER, *The Biology of the Heteroptera* (1956); H.M. PARSHLEY, *A Bibliography of North American Hemiptera-Heteroptera* (1927); T.R.E. SOUTHWOOD and D. LESTON, *Land and Water Bugs of the British Isles* (1959), keys and notes through species; W. STICHEL, *Il-*

lustrierte Bestimmungstabellen der Wanzen, 2, Europa (1955–62), four volumes with keys to European species and a checklist for Palearctic Region; R.L. USINGER, *Monograph of Cimicidae (Hemiptera-Heteroptera)*, vol. 7 (1966), incorporates modern methodology in monographic study of important bedbugs; E.P. VAN DUZEE, *Catalogue of the Hemiptera of America North of Mexico* (1917), important list and literature summary of North American Heteroptera; H. WEBER, *Biologie der Hemipteren* (1930), important source for original and compiled information.

(R.C.F.)

Hieroglyphic Writing

Hieroglyphic writing is a system that employs characters in the form of pictures. These individual signs, called hieroglyphs, may be read either as pictures, as symbols for pictures, or as symbols for sounds. The name hieroglyphic (from the Greek word for "sacred carving") is first encountered in the writings of Diodorus Siculus (1st century BC). Earlier, other Greeks had spoken of sacred signs when referring to Egyptian writing. Among the Egyptian scripts, the Greeks labelled as hieroglyphic the script that they found on temple walls and public monuments, in which the characters were pictures sculptured in stone. The Greeks distinguished this script from two other forms of Egyptian writing that were written with ink on papyrus or on other smooth surfaces. These were known as the hieratic, which was still employed during the time of the ancient Greeks for religious texts, and the demotic, the cursive script used for ordinary documents.

Non-Egyptian hieroglyphic scripts

Hieroglyphic, in the strict meaning of the word, designates only the writing on Egyptian monuments. The word has, however, been applied for about 100 years to the writing of other peoples, insofar as it consists of picture signs used as writing characters. The name hieroglyphics is, for example, always used to designate the scripts of the Indus civilization and of the Hittites, who also possessed other scripts, in addition to the Mayan, the Incan, and Easter Island writing forms, and also the signs on the Phaistos Disk on Crete. Colloquially, the word hieroglyphics has been extended to mean any sort of illegible or barely legible writing.

Because of their pictorial form, hieroglyphs were difficult to write and were used only for monument inscriptions. They were usually supplemented in the writing of a people by other, more convenient scripts. Among living writing systems, hieroglyphic scripts are no longer used.

The rest of this article is concerned only with Egyptian hieroglyphic writing.

Development of Egyptian hieroglyphic writing. The most ancient hieroglyphs date from the end of the 4th millennium BC and comprise annotations to the scenes cut in relief—found on slabs of slate in chapels or tombs—that had been donated as votive offerings. Although by no means all of these earliest signs can be read today, it is nonetheless probable that these forms are based on the same system as the later classical hieroglyphs. In individual cases, it can be said with certainty that it is not the copied object that is designated but rather another word phonetically similar to it. This circumstance means that hieroglyphs were from the very beginning phonetic symbols. An earlier stage consisting exclusively of picture writing using actual illustrations of the intended words cannot be shown to have existed in Egypt; indeed, such a stage can with great probability be ruled out. No development from pictures to letters took place; hieroglyphic writing was never solely a system of picture writing. It can also be said with certainty that the jar marks (signs on the bottom of clay vessels) that occur at roughly the same period do not represent a primitive form of the script. Rather, these designs developed in parallel fashion to hieroglyphic writing and were influenced by it.

Improbability of relationship to cuneiform

It is not possible to prove the connection of hieroglyphs to the slightly older cuneiform characters used by the Sumerians in southern Mesopotamia. Such a relationship is improbable because the two scripts are based on entirely different systems. What is conceivable is a general tendency toward words being fixed by the use of signs, without transmission of particular systems.

Invention and uses of hieroglyphic writing. The need to identify a pictorial representation with a specific, unique event, such as a hunt or a particular battle, led to the invention of hieroglyphic writing. Hieroglyphs added to a scene signified that this illustration represented a particular war rather than an unspecified one or war in general. This new attitude toward time and toward history as unique events in time led to the invention of hieroglyphic writing. The system first appeared only in connection with relief depictions, which they explained by means of place names. Beginning in the 1st dynasty (3100 BC), images of persons were also annotated with their names or titles, a further step toward expressing individuality and uniqueness. The so-called annalistic tablets of the first two dynasties were pictorial representations of the events of a year with specifically designated personal names, places, and incidents. For example, accompanying a scene of the pharaoh's triumph over his enemies is the annotation "the first occasion of the defeat of the Libyans." Simultaneously, the writing of the Egyptians began to appear unaccompanied by pictorial representations, especially on cylindrical seals. These roller-shaped incised stones were rolled over the moist clay of jar stoppers. Their inscription prevented the sealed jar from being covertly opened and at the same time described its contents and designated the official responsible for it. In the case of wine, its origin from a specific vineyard and often also the destination of the shipment were designated, and, as a rule, so was the name of the reigning king.

From the stone inscriptions of the 1st dynasty, only individual names are known, these being mainly the names of kings. In the 2nd dynasty, titles and names of offerings appear, and, at the end of this dynasty, sentences occur for the first time. The discovery of a blank papyrus scroll in the grave of a high official, however, shows that longer texts could have been written much earlier; *i.e.*, since the early part of the 1st dynasty.

Relationship of writing and art. The form of these hieroglyphs of the archaic period (the 1st to 2nd dynasty) corresponds exactly to the art style of this age. Although definite traditions or conventions were quickly formed with respect to the choice of perspective—*e.g.*, a hand was depicted only as a palm, an eye or a mouth inscribed only in front view—the proportions remained flexible. The prerequisite of every writing system is a basic standardization, but such a standardization is not equivalent to a canon (an established body of rules and principles) in the degree of stylistic conformity that it requires. A recognized canon of Egyptian hieroglyphic writing arose in the 3rd dynasty and was maintained until the end of the use of the script.

Formation of the standard and canon of Egyptian writing

In that hieroglyphic signs represented pictures of living beings or inanimate objects, they retained a close connection to the fine arts. The same models formed the basis of both writing and art, and the style of the writing symbols usually changed with the art style. This correspondence occurred above all because the same craftsmen painted or incised both the writing symbols and the pictures. Deviations from the fine arts occurred when the writing, which was more closely bound to convention, retained patterns that the fine arts had eliminated. The face in front view is an example of this. This representation, apart from very special instances, was eventually rejected as an artistic form, the human face being shown only in profile. The front view of the face was, however, retained as a hieroglyph from the archaic period to the end of the use of hieroglyphic writing. Similar cases involve the depiction of various tools and implements. Although the objects themselves fell out of use in the course of history—*e.g.*, clubs used as weapons—their representations, mainly misunderstood, were preserved in the hieroglyphic script. The hieroglyphs corresponding to such objects that had disappeared from daily life were therefore no longer well known and were often distorted beyond recognition. But the style of representation in the hieroglyphs still remained closely bound to the art of the respective epoch. Thus there appeared taut, slender forms or sensuous, fleshy ones, or even completely bloated characters, according to the art style of the period.

Media for hieroglyphic writing. In historical times (2800 BC–AD 300), hieroglyphic writing was used for inscribing stone monuments and appeared in Egyptian relief techniques, both high relief and bas-relief; in painted form;

on metal, sometimes in cast form and sometimes incised; and on wood. In addition, hieroglyphs appear in the most varied kinds of metal and wood inlay work. All of these applications correspond exactly with the techniques used in fine art, and the same craftsmen who produced the works of art painted or incised the hieroglyphic inscriptions.

Subject matter of inscriptions

Hieroglyphic texts are found primarily on the walls of temples and tombs, but they also appear on memorials and gravestones, on statues, on coffins, and on all sorts of vessels and implements. Hieroglyphic writing was used as much for secular texts—historical inscriptions, songs, legal documents, scientific documents—as for religious subject matter—cult rituals, myths, hymns, grave inscriptions of all kinds, and prayers. These inscriptions were, of course, only a decorative monumental writing, unsuitable for everyday purposes. For popular use, hieratic script was developed, an abbreviated form of the picture symbols such as would naturally develop in writing with brush and ink on smooth surfaces like papyrus, wood, and limestone.

Writing and religion. The influence of religious concepts upon hieroglyphic writing was confined to two cases. In the 3rd millennium, certain signs were avoided or used in garbled form in grave inscriptions for fear that the living beings represented by these signs could harm the deceased who lay helpless in the grave. Among these taboo symbols were human figures and dangerous animals, such as scorpions and snakes. Secondly, in all periods and for all uses of the writing, symbols to which a positive religious significance was attached were regularly placed in front of other signs, even if they were to be read after them. Among these were hieroglyphs for God or individual gods, as well as those for the king or the palace. Thus, for example, the two signs,  denoting the word combination "servant of God" (priest), are written so that the symbol for God,  , stands in front of that for servant,  , although the former is to be read last. Moreover, theology traced the invention of hieroglyphic writing back to the god Thoth, although this myth of its divine origin did not have an effect on the development of the script. In the late period, Egyptian texts referred to hieroglyphic inscriptions as "writing of God's words"; earlier, in contrast, they were simply called pictures.

Literacy and knowledge of hieroglyphic writing. At all periods only a limited circle understood the script. Only those who needed the knowledge in their professions acquired the arts of writing and reading. These people were, for example, officials, doctors, and priests (insofar as they had to be able to read rituals and other sacred texts), as well as craftsmen whose work included the making of inscriptions. Under Greek and especially under Roman rule, the knowledge declined and was entirely confined to temples, where priests instructed their pupils in the study of hieroglyphic writing. From the time of the rule of the Ptolemies (305 to 30 BC), national consciousness became more and more narrowly bound up with religion, and for both the national consciousness and religion alike the tradition-filled hieroglyphic writing was an outward sign—in the fullest sense, a symbol. There was no lack of attempts to replace the hieroglyphic writing, cumbersome and ever more divergent from the spoken language, with the simpler and more convenient Greek script. Such experiments, however, remained ineffective precisely because of the emotional value that the old writing system had when the country was under the foreign domination of the Macedonian Greeks and the Romans.

Decline of hieroglyphic writing

Christianity and the Greek alphabet. The situation was altered with the conversion of the country to Christianity in the 2nd and 3rd centuries AD. The new religion fought against the Egyptian polytheism and traditions, and with its victory, the Greek script triumphed. From the beginning, Egyptian Christians used the Greek alphabet for writing their spoken Egyptian language. This practice involved enlarging the Greek alphabet with five supplementary letters for Egyptian sounds not present in Greek. As a consequence, the knowledge of hieroglyphic writing quickly declined. The last evidence of the writing system is a rock inscription from the island of Philae, dating from August 24, 394, from the reign of the emperor Theodosius I. The language as well as the writing system of the Egyptian Christians is called Coptic.

Characteristics of hieroglyphic writing. The system of hieroglyphic writing has two basic features: first, representable objects are portrayed as pictures (ideograms), and second, the picture signs are given the phonetic value of the words for these represented objects (phonograms). At the same time, these signs are also written to designate homonyms, similar-sounding words. The writing disregards vowels, and, also, in earlier times, the semivowels *i*, *y*, and *w*, thus offering more possibilities for the transference of signs to words with identical consonant combinations. For example, the sign for "wood" is written as a branch,  , which is pronounced with the consonants *ḫ* and *t*, which occur in the Egyptian word for wood. Other words with the same series of consonants can also be written with the same basic sign; *e.g.*, *ḫt* "after," *ḫtì* "to retreat," or *ḫtì* "to carve." Words that consisted of only one consonant, plus one or more vowels, supplied single consonant signs. The Egyptians, however, never reduced their writing to an alphabet by discarding the multiconsonant signs; rather, they retained clearly the form of the original words. When doubts occurred, as in the case of the three signs for the frequent consonant series *m* + *r* (the hoe,  , the chisel,  , and the pyramid, ), the plurality was used to make clear distinctions between words: all derivations from the stem *mr* "to love" were written with the hoe; those from the stem *mr* "to be ill," with the chisel; and those words related to pyramids with the sign for pyramid. Thus, two or more existing signs for the same sound or combination of sounds were retained and used in conscious distinction to promote easier readability. Although each sign originally had only one reading, occasional ambiguities did develop through the convergence of two symbols of similar form, such as those for the thighbone and the shankbone of an animal. A few signs, therefore, had two or, less commonly, three readings in classical Egyptian writing.

Ideograms and phonograms

Reading aids: spelling, phonetic complements, determinatives. By means of this rebus system in which letters and pictures were combined the Egyptians could write a large number of the words of their language. But there remained a residue for which no drawable word with the same consonant framework presented itself; *e.g.*, *nḫt* "strong." Here the Egyptians spelled out the word: for *n*, they had a sign, the water symbol (from the word *nwy* "flood"), and for *ḫt* they had the above-mentioned sign for wood, , so that they could now write *nḫt* as  .

Two additional reading aids that were quickly added to this system promoted distinctness and readability. For multiconsonant signs, one or more consonants, or in some cases all of them, were also written to serve as a phonetic complement. Thus there is for *mr* "to be ill," in which the owl (top, left) possesses the phonetic value *m* and the mouth ( ) that of an *r*. In cases like this, the consonants, according to the conception of modern scholars, were written twice but read only once. For the Egyptians, the single consonant signs were there simply as reading aids for clarification of the word sign, the logogram. Accordingly, they wrote , inwhich is complemented by the two signs , *ḫ*, and , *t*, which appear after it.

Phonetic complements

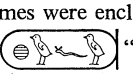
In addition, determinatives—signs that do not represent a phonetic value but serve only to inform the reader as to the family of meanings to which the designated word belongs—were quickly formed out of these. The consonant combination *ḫtì* "to engrave" has a knife written after it; on the other hand, *ḫtì* "to retreat" has legs striding backward. Thus, these two words, otherwise written identically, are differentiated graphically as and . In this

manner, each Egyptian word possessed its own writing picture with which it was strictly associated. Grammatical endings were attached to this word picture and stood after the determinative. From the outset, therefore, Egyptian writing was a complete script; that is, it could unequivocally fix any word, including all derivations and all grammatical forms.

Summary of the types of signs. In summation, hieroglyphs can be separated into three groups, of which the first two render a phonetic value, and the third represents mute reading aids: (1) ideograms, or signs that should be read as the word they represent—*e.g.*,  , "branch"; (2) phonograms, or signs that do not refer back to the objects they represent but stand simply for one or more consonants—*e.g.*,  as *n* and as *ḥt* in , *nḥt* "strong"; and (3) determinatives, which possess no phonetic value but which aid the reader by leading him to the correct interpretation of the meaning—*e.g.*,  in  , *ḥtì* "to retreat." Egyptian writing lacked punctuation in our meaning of the term. Line and stanza signs appeared only in certain literary texts.

Number of symbols. In the classical period, the number of hieroglyphs totalled approximately 700. Their number multiplied considerably in the late period (about 600 BC), because scholars constantly invented new forms or signs. The additions were, however, always in accordance with the principles that had governed Egyptian writing from its beginnings. The hieroglyphic system remained flexible throughout all periods, always open to innovation, even though, as with every writing system, convention played a preponderant role.

Invention of new signs

Direction of the writing. The lines were written from right to left or, less frequently, from left to right. Vertical rows could be placed next to horizontal rows, according to the demands of the architectural setting. The direction of the writing is immediately ascertainable because the signs always face the beginning of the row. Occasionally, some signs are turned around in the row, presumably so that two human figures can face one another, to avoid standing with their backs toward each other. These rotations are infrequent, however, and are found almost exclusively in the names of kings. Royal names were enclosed in a ring, the so-called cartouche; *e.g.*,  "Khufu," in Egyptian *Ḫwfw*. This ring, originally a rope, was supposed to protect the bearer of the enclosed name from injury and, in particular, from harmful magic.

Egyptian pedagogical traditions. To understand hieroglyphic writing, one must know about its tradition within Egypt. The Egyptian student of writing, who brought with him a knowledge of the spoken language as his mother tongue, began by learning the script picture corresponding to each word without having isolated its elements; *i.e.*, its individual signs. Through centuries this pedagogical tradition in the schools helped Egyptian words retain the original established spelling, with only minor—usually stylistic—changes, even when the phonetic form had radically changed. Hieroglyphic writing thus conceals historical sound changes.

Sound changes shown by misspellings

The mistakes in hearing made by pupils in the writing schools have helped scholars to understand the phonetic changes that occurred in the development of the Egyptian language. When the pupil who was learning to write the hieroglyphic script did not recognize a word dictated to him, he wrote it badly—that is, just as he heard it. Because he had not yet learned to spell in the orthodox manner, what appeared on his papyrus was usually a word that sounded similar to the dictated but misunderstood term and whose word picture was familiar. Thus, although Egyptian writing was originally composed of symbols that represented a phonetic value, the system was transmitted in the form of word pictures; that is, closed or indivisible groups, generally of several signs per word.

Cryptographic hieroglyphic writing. That knowledge of the hieroglyphic system and the principles upon which it was devised had not become lost is attested by two phenomena: cryptography and the development of the hieroglyphic writing during the last millennium of its existence. From the middle of the 3rd millennium but more frequently in the New Kingdom (from 1500 BC), hieroglyphic texts are encountered that have a very strange appearance. The absence of familiar word groups and the presence of many signs not found in the canon characterize these texts at first glance as cryptographic, or secret, writing. This kind of hieroglyphic writing was probably intended as an eye-catcher, to entice people to seek the pleasure of deciphering it. Composed according to the original principles of the script, these inscriptions differed only in that certain features excluded when the original canon was formulated were now exploited. The new possibilities involved not only the forms of the signs but also their selection. For example, the mouth was not drawn in front view (), as in the classical script, but in profile (), although it had the same phonetic value. An example of a change in the choice of signs is the case in which a man carrying a basket on his head (), a determinative without phonetic value in the classical script, was later to be read as *f* and was used in lieu of the familiar sign having this phonetic value, that of the horned viper. In the new selection of the sign, the phonetic value is obtained from the word *f3ì* "to carry" (neglecting its two weak consonants), in accordance with a principle that the inventors of the writing had applied in 3000 BC. These cryptographic inscriptions prove that alongside the method of instruction in the schools, which was based on memorization or recognition, not upon analytical understanding, there was another tradition that transmitted knowledge of the basic principles of the hieroglyphic script. A command of the principles of hieroglyphics similar to that which the composers of the cryptic inscriptions had was presupposed for the puzzle-happy decipherers. Because the encoded texts often consisted of a petition by the inventor of the text to say a prayer on his behalf, the number of these decipherers must surely not have been small.

Transmission of the hieroglyphic principles

Growth of hieroglyphic writing during the last millennium BC. At about the middle of the last millennium BC, Egyptian writing experienced new developments and a revival of interest. Again the inscriptions abounded with new signs and sign groups unknown in the classical period, all generated according to the same principles as the classical Egyptian script and the cryptographic texts. The writing of this late period was distinguished from the cryptograms in that this script, like every normal system of writing, developed a fixed tradition, being intended not to conceal but to be read easily, whereas the cryptography strove for originality.

Stages of hieroglyphic writing. The development of hieroglyphic writing thus proceeded approximately as follows: at first only the absolutely necessary symbols were invented, without a canonization of their artistic form. In a second stage, easier readability (*i.e.*, increased rapidity of reading) was achieved by increasing the number of signs (thereby eliminating some doubts) and by employing determinatives. Finally, after the second stage had endured, essentially unaltered, for about 2,000 years, the number of symbols increased to several thousand in about 500 BC. This rampant growth process occurred through the application of hitherto unused possibilities of the system. With the triumph of Christianity, the knowledge of hieroglyphic writing was extinguished along with the ancient Egyptian religion.

Tools. The tools used by the craftsmen for writing hieroglyphic symbols consisted of chisels and hammers for stone inscriptions and brushes and colours for wood and other smooth surfaces. Only for the cursive scripts, hieratic and demotic, were special materials developed. Leather and papyrus became writing surfaces, and the stems of rushes in lengths of 16–32 centimetres, cut obliquely at the writing end and chewed to separate the fibres into a brushlike tip, functioned as writing implements. The split calamus reed was introduced into Egypt by the Greeks in the 3rd century BC.

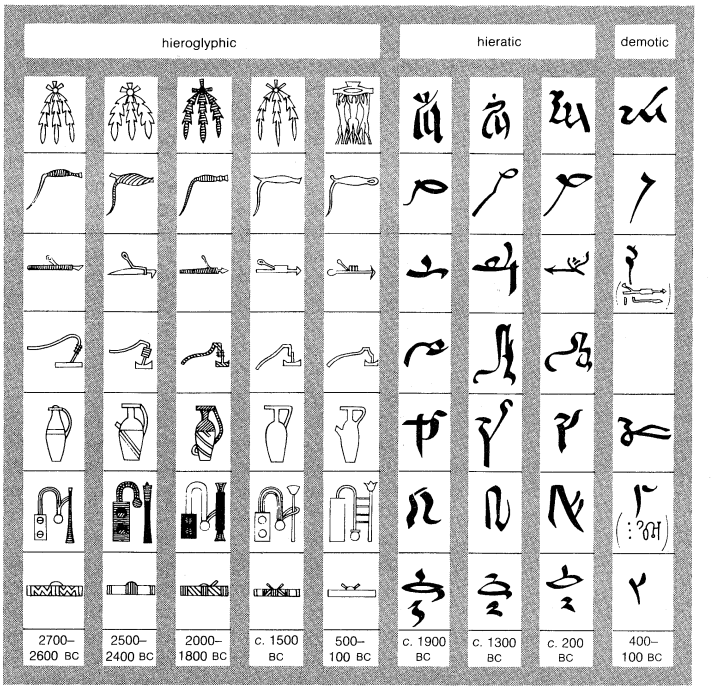
Development of special writing materials

Hieratic script. The Egyptian cursive script, called hieratic writing,. received its name from the Greek *hieratikos* (meaning priestly) at a time when the script was used only

for sacred texts. Everyday secular documents were written in another style, the demotic (Greek *dēmotikos* "for the people" or "in common use") script.

Relation of hieratic to hieroglyphic script. The structure of the hieratic script corresponds with that of hieroglyphic writing. Changes occurred in the characters of hieratic simply because they could be written rapidly with brush or rush and ink on papyrus. In general, the picture form is not, or not easily, recognizable. Because their models were well known and in current use throughout Egyptian history, the hieratic symbols never strayed too far from them. Nevertheless, the system differs from the hieroglyphic script in some important respects:

1. Hieratic was written in one direction only, from right to left. In earlier times, the lines had run vertically, and later, about 2000 BC, horizontally. Subsequently, the papyrus scrolls were written in columns of changing widths.
2. There were ligatures in hieratic, so that two, but no more than two, signs could be written in one stroke.
3. As a consequence of its decreased legibility, the spelling of the hieratic script was more rigid than that of hieroglyphic writing. Variations from uniformity at a given time were minor; but, during the course of the various periods, the spelling developed and changed. As a result, hieratic texts do not correspond exactly to contemporary hieroglyphic texts, either in the placing of signs or in the spelling of words.
4. Hieratic used diacritical additions to distinguish between two signs that had grown similar to one another because of cursive writing. For example, the cow's leg received a supplementary distinguishing cross, because in hieratic it had come to resemble the sign for the leg of a man. Certain hieratic signs were taken into the hieroglyphic script.



Egyptian hieroglyphs and their cursive equivalents. The hieroglyphs depict (top to bottom): three fox skins tied together; a whip; a single-barbed harpoon; an adz at work on a block of wood; a stone jug with handle; a scribe's outfit; a roll of papyrus tied with a cord.
Courtesy of the Oriental Institute, University of Chicago

Role of hieratic in Egypt

All commonplace documents—*e.g.*, letters, catalogues, and official writs—were written in hieratic script, as were literary and religious texts. In the life of the Egyptians, hieratic script played a larger role than hieroglyphic writing and was also taught earlier in the schools. In offices, hieratic was replaced by demotic in the 7th century BC, but it remained in fashion until much later for religious texts of all sorts. The latest hieratic texts stem from the end of the 1st century or the beginning of the 2nd century AD.

Demotic script. Demotic script is first encountered at the beginning of the 26th dynasty, in about 660 BC. The writing signs plainly demonstrate its connection with the hieratic script, although the exact relationship is not yet clear. The demotic characters are more cursive (flowing and joined), and thus more similar to one another, with the result that they are more difficult to read than are the hieratic forms. Countering this difficulty, there is less freedom for the writer's individual variations. It appears that demotic was originally developed expressly for government office use; that is, for documents in which the language was extensively formalized and thus well suited for the use of a standardized cursive script. Only some time after its introduction was it used for literary texts in addition to documents and letters; much later it was employed for religious texts also. The latest dated demotic text, from December 2, AD 425, consists of a rock inscription at Philae. In contrast to hieratic, which is, almost without exception, written in ink on papyrus or other flat surfaces, demotic inscriptions are not infrequently found engraved in stone or carved in wood.

Alternative demotic spelling. The demotic system corresponds to the hieratic and hence also to the hieroglyphic system. Alongside the traditional spelling, however, there was another spelling that took account of the markedly altered phonetic form of the words by appropriate respelling. This characteristic applied especially to a large number of words that did not occur in the older language and for which no written form had consequently been passed down. The nontraditional spelling could also be used for old, familiar words.

Decipherment of hieroglyphic writing. With the possible exception of Pythagoras, no Greek understood the nature of hieroglyphic writing. The Greeks did not obtain guidance from their Egyptian contemporaries, some of whom even lived on Italian soil and wrote proper hieroglyphic inscriptions on Roman obelisks. Rather, the Greek tradition taught that hieroglyphs were symbolic signs or allegories. The Egyptian-born Greek philosopher Plotinus interpreted hieroglyphic writing entirely from the viewpoint of his esoteric philosophy. Only one of the numerous works on the hieroglyphic script written in late antiquity has been preserved: the *Hieroglyphica* of Horapollon, a Greek Egyptian who probably lived in the 5th century AD. Horapollon made use of a good source, but he himself certainly could not read hieroglyphic writing and began with the false hypothesis of the Greek tradition, namely, that hieroglyphs were symbols and allegories, not phonetic signs.

Greek beliefs concerning hieroglyphic writing

The Middle Ages neither possessed any knowledge of hieroglyphic writing nor took any interest in it. But a manuscript of Horapollon brought to Florence in 1422 stirred great interest among the Humanists. Without giving a thought to thepossibility that ancient Egyptian originals might be available in Rome, Renaissance artists designed hieroglyphs after Horapollon's descriptions, as well as from their own imaginations. They used hieroglyphs as wisdom-laden symbols in architecture and also in drawings and paintings.

Kircher's attempts at decipherment. The great German scholar Athanasius Kircher (1602–80) began his attempts at decipherment with the Coptic language and with the correct hypothesis that the hieroglyphs recorded an earlier stage of this language. He also believed, again correctly, that the signs recorded phonetic values. In spite of this, he did not arrive at correct results—with the exception of a single character. This failure can be attributed not only to Kircher's erroneous assumption that the hieroglyphs must correspond phonetically to an alphabet but primarily to the fact that he was most interested in the Renaissance conception of a supposed symbolic meaning constituting the deeper significance of hieroglyphs. In his view, the phonetic value of the hieroglyphs was merely the commonplace, superficial part of the sign.

Discovery of the Rosetta Stone. Both the intellectual and the physical prerequisites for the deciphering of the hieroglyphic script first presented themselves at the end of the

18th century. By accident, a stone that exhibited three different scripts—hieroglyphic, demotic, and Greek—was discovered by members of Napoleon's expedition to Egypt in 1799 near Rashid (French Rosette; English Rosetta) on the Mediterranean coast. The Greek text stated clearly that the document set forth the same text in the sacred script, the folk or popular script, and Greek. The stone was promptly made known to all interested scholars. Important partial successes in the effort of deciphering the scripts were achieved by the Swede Johan David Åkerblad and by the great English physicist Thomas Young, who mainly studied the demotic text, again beginning with the false hypothesis that the hieroglyphs were symbols. Young succeeded in proving that they were not symbols—at least that the proper names were not—and that the demotic and hieratic signs had come from the hieroglyphs. (He first published this result in the supplement to the 4th, 5th, and 6th editions of the *Encyclopædia Britannica*.) He was the first to isolate correctly some single consonant hieroglyphic signs. But a wrong turn in the course of his investigations then prevented him from fully deciphering the writing.

Thomas Young's work

Champollion's decipherment. This task of complete decipherment was first accomplished by the Frenchman Jean-Francois Champollion (1790–1832) in 1822, after long years of intensive work and many setbacks. His success was due to the recognition that hieroglyphic writing, exactly like the hieratic and demotic scripts derived from it, did not constitute a writing system of symbols but rather a phonetic script. He arrived at this breakthrough by an exact comparison of the three Egyptian forms of writing, as well as by reference to Coptic, the late phase of the Egyptian language that was written with the Greek alphabet and was thus directly readable. The Coptic language was also understood at that time. Starting, as had his predecessors, from Ptolemy and Cleopatra, both ring-enclosed royal names, and adding the hieroglyphic spelling of Ramses' name, Champollion determined, essentially correctly, the phonetic values of the signs. Soon after, he also learned to read and translate a large number of Egyptian words. Since then, precise research has confirmed and refined Champollion's approach and most of his results.

BIBLIOGRAPHY. Books in English on hieroglyphic writing include: A.H. GARDINER, *Egyptian Grammar*, 3rd ed., rev. (1957), with a list of hieroglyphs from the Middle Kingdom; H. PETRIE, *Egyptian Hieroglyphs of the First and Second Dynasties* (1927); F.L. GRIFFITH, *A Collection of Hieroglyphs* (1898); and N.M. DAVIES, *Picture Writing in Ancient Egypt* (1958), a collection of artistically valuable hieroglyphs.

For those with a knowledge of German, the following texts may be consulted: H. BRUNNER in the *Handbuch der Orientalistik*, pt. 1, vol. 1, ch. 1 (1959), and "Die Schrift der Ägypter," *Handbuch der Archäologie: Allgemeine Grundlagen der Archäologie* (1969); S. SCHOTT, *Hieroglyphen* (1951), the origin and early development of hieroglyphic writing; P. KAPLONY, *Die Inschriften der Ägyptischen Frühzeit*, 3 vol. (1963, suppl. 1964), a collection of inscriptions of the archaic period; G. MOLLER, *Hieratische Paläographie*, 3 vol. (1909–12); W. ERICHSEN, *Demotische Lesestücke I. Literarische Texte*, pt. 3 (1937), and *Auswahl frühdemotischer Texte*, pt. 3 (1950). The last two sources concern demotic texts.

(H.Br.)

Higher Education

The best short definition of higher education may be the one devised and accepted in 1962 by 44 nations participating in a UNESCO conference on higher education in Africa:

> Higher education is defined as all types of education (academic, professional, technological, or teacher education) provided in institutions such as universities, liberal arts colleges, technological institutes and teachers' colleges for which: (a) the basic entrance requirement is completion of secondary education . . . ; (b) the usual entrance age is about 18 years; and (c) in which the courses lead to the giving of a named award (degree, diploma, or certificate of higher studies).

The definition has unavoidable limitations since each nation uses its own nomencature for its various educational institutions and programs. Many countries, for example, use the term school to describe an institution of higher education, particularly a professional institution such as medical school or law school, and the word college, on the other hand, is frequently used to describe an institution of secondary education. Moreover, although all institutions of higher education provide courses concluded by the giving of a named award, many also offer some courses that do not conclude in this way.

In addition to setting up institutions of higher education, all countries provide in different ways for the education of those who, at 18 years of age, are described as adults and who do not wish to proceed to an institution of higher education, but who do wish to continue their education. Courses of study in such cases are generally shorter and less rigorous than those pursued in institutions of higher education and rarely lead to a named award. They are best described as courses in "adult" education. It is not always easy, however, to draw the line between institutions of higher and adult education.

The article is divided into the following sections:

I. Major models of higher education
  Characteristics of education in major countries
  Overseas influence of the major models
II. Organization of higher education
  Curriculum in higher education
  Social and administrative patterns of higher education
  Research and postgraduate studies
III. General trends of change
  Recent changes
  Regional trends of change

I. Major models of higher education

The essential characteristics of higher education in western Europe may be said to have been firmly established in its institutions between the 11th and 16th centuries; after that, new historical forces worked to produce different modifications of the medieval pattern in different countries. It is possible to recognize five major models of higher education that developed in the five countries of France, Germany, England, the United States, and the Soviet Union; these, more than others, have powerfully influenced the development of higher education elsewhere in the modern period.

CHARACTERISTICS OF EDUCATION IN MAJOR COUNTRIES

France. The medieval universities of France, abolished during the Revolution in 1793, were re-established by Napoleon in 1808, and because by that time he had made France supreme in Europe, their influence was powerful over a very large area of Europe and for a very long time. The Napoleonic pattern of government was a metropolitan one, and the centralization of political power was reflected in a system of higher education characterized by the primacy of a single university, the University of Paris. Only in very recent years has a policy of decentralization been initiated to offset the weakness of provincial universities due to a drain of superior talent to the capital. Even so, today, a third of all French students go to Paris for their higher education.

Centralization in the French system

The clarity of the organization of French education is shown in the geographical division of all France and its overseas possessions into a number of *académies*, each composed of from three to eight civil *départements*. The academic head of each *académie* is the rector of its major university, who is not only the head of the university but also the head of the whole educational system in the *académie* and the representative there of the central Ministry of National Education. The rector is appointed by the president of the Republic himself and is both chairman of his own university council and president of a regional council for higher education.

Central control over higher education is exercised by a Directorate acting on the advice of a Council of Higher Education, which is responsible for policy in matters of curriculum, examinations, awards, and administration. Parallel with the council is a Universities Consultative Committee organized in five divisions corresponding to the fivefold Faculty organization that is typical of the French pattern of higher education: law, science, letters,

medicine, and pharmacy. (At the University of Strasbourg there is also a faculty of theology.) The consultative committee deals with faculty matters and is responsible for the nomination and promotion of professors. Each faculty is under the charge of a dean, an officer appointed for three years by the Minister of National Education on the joint recommendation of faculty and the university council. The Dean is in charge of faculty finance and administration and is provided with a secretariat. The office of dean symbolizes the degree to which university teachers are saved from many of the burdens of administration, but in a way that also relieves them of some of the advantages of self-government.

Higher education in France is free and open to all students who have passed the *baccalauréat* examination with which the cycle of secondary studies is concluded. The first year of studies in a university is called the preparatory, or *propédeutique*, year. It is terminated by an examination the severity of which may be judged from the fact that only 30 to 40 percent of the students are selected to continue their degree studies. Those excluded can, however, proceed with a further year of higher studies, and many of them do so in a two-year *collège universitaire*, where certificates or diplomas of higher studies are awarded at the end of the two-year period. Those who are successful in the *propédeutique* examination embark on a course lasting for further three or four years and leading to the first university degree of *licence*.

Grandes écoles

In France there are important institutions of higher education that have grown up outside the universities. Called the *grandes écoles*, they are concerned with advanced professional and technological training. Though some of these schools are affiliated with universities, the majority are not. Their students, whose number is limited, are recruited by competitive examinations held all over France for candidates who already possess the *baccalauréat* and who, customarily, have followed it by a year or two of special coaching. In these schools the work of students is closely supervised so that their diplomas have a standing in France that is higher than that of most *licences*. These high standards have ensured a supply of intelligent, well-educated men into all branches of applied science and technology in France, and their organization has been copied wherever the French model has had influence.

The degree of financial aid given to French students is less than might be expected in a state system. Only some 30 percent of students receive state grants, and the average value of these is no more than £100 a year. In addition, there are subsidized meals, welfare facilities, and student loans, repayment of which need not begin until ten years after graduation.

Following the example of Paris, many universities have built a *cité universitaire*, or group of residence halls, but these are mainly for the benefit of foreign students or, to a lesser degree, students from distant parts of France, since most students are obliged, for financial reasons, to live at home. The effect of this is to give to universities outside Paris a strongly regional character. There is therefore nothing like the movement between universities that is common in Germany.

An aspect of higher education that follows different lines of development in different countries is that of institutions concerned with the training of teachers. In France, as in a number of other countries, training institutions for teachers in primary schools are not regarded as part of the system of higher education. Graduates who wish to teach may study for their *agrégation* either at an institute of education attached to the faculty of letters or science at a university of at one of the four *écoles normales supérieures*. If less ambitious, a *licencié* may undertake a short period of study for the professional qualification of the *certificat d'aptitude* and then follow this by a longer period of practical training at one of the regional centres set up by the Ministry of Education under a decree of 1957.

After the student disturbances of 1968, the Ministry of Education initiated far-reaching reforms that were expected eventually to bring French universities much closer to the American and English models. In place of central control there was to be institutional autonomy. The rigid faculty structure was to be replaced by a flexible grouping of Unités d'Anseignement et de Recherche (UAR). One-third of the places on the elected UAR councils were to go to students. Formal lectures and infrequent examinations were to be replaced by work in small groups with continuous assessment. A new award, the *maîtrise*, was to follow the *licence*. New university construction was to be modelled more nearly on the campus pattern. Twenty new university Institutes of Technology began to invade the monopoly of the *grandes écoles*.

Germany. As in France, the institutions of higher education established in Germany in the 19th century were, from the outset, operated as agencies of the state. But the separate German principalities aimed at strong federal unity, and so the control of universities in a federal Germany came from the separate *Länder*, and there developed no single centre of power, as in the French model. There did, however, develop a view of the university as the embodiment of the national mind, and this was combined with the ideal of a university as primarily a research institute. These two seminal ideas soon influenced both European and American institutions of higher education.

Federalism and centralization in the German system

The relations between universities and the state are close in both the French and German models of higher education, but a sharp difference results from the status of two officials in the German university—the *Rektor* and the *Kurator*. The *Rektor*, the curricular and faculty chairman, is elected by secret ballot from among the titular professors and holds office for one year only. The *Kurator* is appointed by the state minister of education as a permanent officer to be the head of the university's administrative staff and to control its finances. The internal organization of universities has developed in relation to the nature of these two offices. Because of the annual change of the *Rektor*, the university senate—which includes not only the titular professors but also representatives of all grades of the academic staff and also student members elected by a student parliament—becomes the main initiator of policy.

A feature of staffing in German institutions of higher education is the method of recruiting most teachers from among the large groups of assistants who have found professors willing to recommend them to the faculty for a period of study beyond the first degree (which is a doctorate), leading to the postgraduate award of *Habilitation*, which can confer on its possessor the rank of *Privatdozent* (corresponding roughly to the English reader or American assistant or associate professor). The *Habilitation* award requires the student to present a research thesis to the faculty and to conduct a colloquium and lecture before the faculty board. The *Privatdozenten* are by far the largest group among university teachers, and titular professors are drawn from them. Titular professors holding chairs in established departments and "extraordinary professors" holding chairs in new or very specialized disciplines are civil servants for life. In recent years a new grade of teacher has been established—that of academic councillor, conferred upon *Privatdozenten* who, at the recommendation of the university, are appointed civil servants for life. Another group are the *Lektoren* or lecturers appointed to teach a special subject, some of whom come from outside the institutions of higher education.

Admission to German universities follows the same general plan as admission to the French, but in matters of curriculum, examinations, and degree awards there are marked differences, springing from the fact that in Germany both students and teachers are traditionally compulsive wanderers. It is the custom for students to attend two, three, or even four different universities in the course of their undergraduate studies; and the majority of professors at a particular university may have taught in four or five others. As a result, the personal loyalty to his university that marks a Harvard man in the United States or an Oxford man in England is almost absent in Germany. Able students go where their subject is best taught or where research is of repute. This inevitably

German scholastic wandering

means that schemes of study and examination are marked by a freedom unknown in France. Minimum requirements are laid down, after which students are free to choose the component parts and the lengths of their degree courses. There are no annual examinations, but, as a course proceeds, written papers are prepared, test essays are set, and the quality of these must be adequate before authorization to the final examinations is given. This movement and freedom in undergraduate studies means that universities have to respond to a nationwide competition, and the enfeebling effect of metropolitan institutions upon the more provincial is thus avoided.

The first degree award in Germany is called the doctorate, and afterward, as in France, students can sit for further examinations for awards, some of which are set by the university and some by the state.

One of the results of the tradition of drawing academic staff from the large group of candidates for the *Habilitation* is that postgraduate courses rarely have a formal organization; a period of service as an assistant to a professor, combined with research and postgraduate study, is seen as serving the same purpose.

The academic year in Germany is divided into two semesters, rather than three as in the English model, and for the early semesters of a degree course a large majority of students receive state grants that meet about two-thirds of the cost of fees and living expenses. For the final two semesters of the course, this grant is supplemented by an interest-free loan. The custom of student wandering has led to the provision of residential accommodation on a much wider scale than in France. Some of this takes the form of buildings on the university campus where 30 to 40 members of a student corporation live together; there are also larger halls of residence built by religious bodies or by the welfare service of a student parliament or by the university itself. Many halls of residence attempt to be more than simply hotels and develop educational, social, and cultural activities.

England. The pattern of higher education in England is different from that in either France or Germany mainly because until 1969, when the Open University was constituted by royal charter, the state had never founded a university; universities were responsible to no ministry; no government representatives sat on their governing bodies; and universities themselves appointed their own staff. Universities are also independent of any control by local education authorities that control elementary and secondary education in a region. English universities, despite the fact that they receive 90 percent of their capital and recurrent finance from the state, enjoy almost complete autonomy. In addition to rich endowments, about £100 million have been contributed by private individuals or corporations for the purchase of land, the erection of buildings, the endowment of chairs and scholarships, and the building of halls of residence.

Self-government and autonomy in the English system

In recent years, as the size of the state subvention has increased, an administrative invention called the University Grants Committee has been used to associate the Department of Education and Science with the work and expansion of universities while yet allowing them to retain all their ancient freedoms. The members of this committee are appointed by the government, but the majority are individuals actively engaged in university teaching or research and the minority are drawn from industry or nonuniversity research establishments. Universities prepare their budgets on a five-year basis, and once every five years the grants committee visits each university to discuss the present and future needs of the university with its council, senate, academic staff, administrative staff, and students. The committee then advises the government of the total grant needed by the universities, and by tradition the government accepts the advice of the committee without question. Once a block grant for a quinquennium has been given to a university, it is free to use it in any way it thinks best.

The existence of the University Grants Committee has placed upon the academic members of universities the burdens of their own administration, but it has also produced a firmer and more elaborate organization of self-government than in other models. This self-government takes one form in the ancient universities of Oxford and Cambridge and a different form in almost all the others. Oxford and Cambridge are federations of colleges, and the centre of teaching and research is in each of the separate colleges and only to a minor degree in the university itself. The university is the creature of the colleges and possesses only such powers as the colleges have delegated to it. The head of the university is a vice chancellor, who is elected from among the heads of colleges for a period of three years, but who nevertheless retains all his powers and duties as head of a college during his term of office. Colleges are autonomous, each governed by a group of fellows engaged in teaching or research, many of whom live in rooms in college. Fellows are elected by the college in most cases for life, and they are presided over by a master, whom they elect also for life. Fellows teach only the undergraduates of their own college, except when two or more colleges combine to provide teaching in subjects that draw few students. This college teaching is supplemented by the university, which appoints its own professors, readers, lecturers, and research fellows, and maintains libraries, laboratories, institutes, museums, and other facilities open to the students of all colleges.

In other universities in England the pattern is quite different. The main organs of self-government are the senate and council—the senate being an exclusively academic body and the council a mixed lay and academic body. The majority of the senate membership is composed of professors or nonprofessional heads of academic departments, together with a minority representation of the nonprofessorial staff. Under the chairmanship of a vice chancellor (the administrative and academic head of the university, who is appointed by the council for life), the senate is the central governing body of the university in all academic matters. The council exercises control in matters of finance and in matters in which university policy involves issues of regional or national importance; by tradition, it also confirms without amendment the academic decisions of senate. The academic members of the council are elected by the senate, and the lay members by an electorate composed of all the interests in the region concerned with higher education.

The conditions of entry into English institutions of higher education are more complicated than those in France or Germany. The school-leaving examination that secures the General Certificate of Education corresponding to the *baccalauréat* in France or the *Hochschulreife* in Germany is a subject examination, in which each subject may be passed separately either at "advanced" or at "ordinary" level. Possession of a General Certificate does not, as in France or Germany, automatically entitle a school leaver to university entrance, however. Although the minimum standards of entrance to universities are generally the possession of at least two advanced level passes with appropriate supporting passes at ordinary level, faculties and departments require higher qualifications than these, and entry is very competitive. University-entrance procedures have been simplified by the organization of a central admissions bureau, to which candidates for admission are able to give their choice of universities in an order of preference. Minimum standards of admission to colleges of technology or colleges of education are generally the possession of five ordinary level passes, but for these institutions the entry is also competitive, and higher standards than the minimum are necessary to secure a place.

Selective admission and the tutorial system

The selective admission to universities, combined with the close supervision of students through a tutorial system of teaching, makes it possible for most English undergraduates to complete a first degree course in a period shorter than the normal period in France and Germany. For 85 percent of students the length of course is three years.

Although becoming less so, courses in English universities are in general more highly specialized than those in France and Germany. About 80 percent of undergradu-

ates choose to follow an "honours" course in a single subject or, at most, two allied subjects. The remaining minority of students take "pass" courses that include the study of three, four, or five subjects. (Some pass courses may result in an honours degree, however, if the student achieves distinction in the final examination.) Most degree courses in England are examined at the end of the first year of studies and then not again until the "finals" examination at the conclusion of the course.

A unique feature of the English examination system is the use made of "external examiners." Tests are set and marked by the staff of a particular institution, and the results are passed on to a member or group of members of the staff of a different institution who have been appointed as external examiners for a fixed period of years. The external examiners review the assessments given, propose changes, and finally reach agreements with their colleagues in the examining institution on what the final assessments should be. This pragmatic device for securing some uniformity of degree standards among universities has been followed in colleges of technology and education and also in those countries that have borrowed from the English model.

Colleges of education in England are closer to universities than they are in most other countries because they are all now constituent colleges of Institutes of Education, administered and controlled by universities. These Institutes of Education are peculiarly English in their mode of operation. They supervise and coordinate the work of a group of colleges, conduct examinations, and make recommendations to the Minister of Education and Science for the award of "qualified teacher status" to those who are successful in their examinations. They also engage in research and provide extensive series of courses of inservice training for serving teachers both in the university and at regional centres.

Colleges of technology, architecture, commerce, music, and art provide both full-time and part-time courses, to meet professional and technical needs; and an increasing proportion of their students study for "external" degrees awarded by the University of London, for degrees and diplomas awarded by the Council for National Academic Awards, or for "higher certificates" awarded by national professional examining bodies.

In 1971 the Open University began a program of teaching by radio and television for students wishing to study while remaining at home or in employment. The teaching is supplemented by correspondence courses sent by post, by assistance given at a network of study centres by tutors and counsellors, and by required attendance at summer schools. In 1971 only foundation courses were taught, but faculties of arts, science, social sciences, mathematics, technology, and educational studies were to be established in succeeding years.

English inresidence tradition

In all institutions of higher education in England there is a high priority given to expenditure on halls of residence for students. At Oxford and Cambridge almost all students spend a period of residence in their college and another period in "recognized lodgings." At all other universities, halls of residence, generally four or five in number, provide study-bedrooms for students who are not living at home; and some universities also provide special accommodations for married students. In the halls of residence there are rooms for members of the academic staff, as well as common rooms, dining rooms, and facilities for social and athletic activities.

In England, students are extensively assisted from public funds. A standard grant adequate to cover the cost of fees, equipment, and living expenses is available to all students. The grant is reduced if parental income exceeds a given sum, but a minimum of £50 is available to all regardless of parents' income. More generous grants are available for postgraduates' studies, although, unlike undergraduate grants, these are limited in number.

United States. The Laws, Liberties, and Orders of Harvard College, drawn up in 1636, are a fairly faithful copy of the Elizabethan Statute for the English University of Cambridge. Nevertheless, the New England colleges, of which Harvard was simply the first, soon began to adapt old traditions to a new way of life. There were two early adaptations that influenced the whole pattern of university development: the first was the transfer of the government of a university from faculty members to a lay board of trustees; the second was the democratization of the curriculum.

Democratization in the American system

The first adaptation did not live on in its original form. Over the years there has been a gradual and sometimes painful redistribution of power between boards of trustees, faculty members, and university administrations. The democratization of the curriculum began in the 19th century when farmers believed that they needed less of the education suitable for a scholarly elite and more of a broad, liberal, and practical education for an expanding and changing rural population. Social pressures produced a new type of "land-grant college," in which agriculture and veterinary science were taught alongside some traditional subjects and in which there were also such subjects as home economics (or domestic science), journalism, and engineering. The land-grant colleges also developed another idea that has remained as a permanent feature of American higher education—the idea of the university as a service agency for the whole community. Hence, university teachers carried lectures, specimens, books, and demonstrations to the farms and homesteads within range; and university bureaus of welfare answered thousands of questions about everything from sanitation to the testing of milk. Today, American universities feel themselves under an obligation to develop a whole range of services for the general public, including television programs and highly organized home-study courses that may earn a person credits toward a degree.

At the present time, a high proportion of youth in America complete a full cycle of secondary school studies with sufficient high school work to their credit to admit them to a university. As a result there appears to have emerged a nationwide assumption that American children should enjoy at least two years of university education. This has led to an increasing provision of two years of initial undergraduate study at a "junior college" or "community college" rather than at a university, where a majority of students intend to complete four years of study for a degree and where substantial numbers go on for one to three years of postgraduate study in a "graduate school."

Universities that provide four-year degree courses are either private foundations receiving little state or federal assistance or state or city foundations depending heavily or almost completely on public financial support. Because of the freedom enjoyed by private foundations, the publicly supported also enjoy great freedom from public control; and their boards of trustees, on which there is generally a minority of representatives from the state or city, carry large responsibilities. In both private and public schools, it is the board of trustees that appoints the university president, and it is to him that they delegate large powers. He is the main agent in securing funds for a private institution or in securing appropriations in a public one. He is responsible for maintaining good relations between the university and leaders in commerce, administration, and industry and between the university and its alumni, who are generally better organized than alumni in other countries. Above all he is the head of a powerful administrative machine.

In Europe, academic standards do not vary much from university to university, but this is not the case in America, where there is also a much greater proliferation of courses than in European universities. The way to a first degree does not lie through the passing of a single "finals" examination as in England, but through the accumulation of course "credits" or hours of classroom study. The quality of work done in each classroom course is assessed periodically and a student's continuous record of credits and grades is recorded in a course transcript. When a student's progress is finally assessed, his transcript is scrutinized for a possible degree award.

Variety in the American curricula

During the first two years of a student's studies, either in a junior college or a university, three-quarters of his studies generally will be made up of prescribed courses

in the humanities, the natural sciences, the social sciences, and the fine arts. The remaining quarter will be "elective" courses selected by the student himself from a very wide range of options. In the third and fourth years, a student will specialize in one or perhaps two subject fields, with the equivalent of a full year of study in a major field (and sometimes half a year in a minor field) and a full year (or half a year) in a number of supporting studies.

Postgraduate students who wish to pursue either advanced studies or research, or both, do so in one of the many graduate schools. At these schools some students are engaged in working for "master" or "doctor" degrees, while others, who already have their doctorates, will be engaged in their own research or in contract research.

University schools or departments of education have made contributions to higher education that are more notable than those made by similar institutions elsewhere. Most of these schools also have graduate schools.

Financial aid to students is not as uniform as it is in England or Germany, but the number of scholarships sponsored by private individuals, foundations, religious bodies, and industry is increasing rapidly. The federal government grants a large number of scholarships to the more able students, particularly in the fields of science, engineering, and health. Student loans are available, with little or no interest and liberal repayment privileges. Many students pay their fees with earnings gained from working in college cafeterias or student centres or working as student assistants in a library or laboratory.

Most American universities provide housing in dormitories, in houses provided by student societies, or in officially approved off-campus housing.

Soviet Union. No university arose in Russia until 1755, when the empress Elizabeth set up the University of Moscow. Vilna, Kharkov, Karan, and St. Petersburg followed in the next half century. From the outset, all were developed as agencies of the state with no memory of a medieval life as autonomous institutions devoted to the search for truth. German influences, however, brought a growing number of Russian institutions into the main stream of European higher education, until the abrupt reversals of policy that followed in the wake of the Revolution of 1917.

A striking feature of higher education in all the republics of the U.S.S.R. today is the high proportion, (nearly 60 percent) of students who are reading for degrees by part-time study while still engaged in productive enterprise. Such students follow very carefully prepared correspondence courses, are aided by radio broadcasts and television, and meet regularly at centres of study scattered all over the U.S.S.R. Worker-students, they are given leave to attend the study centre, to attend for examinations, and to carry out laboratory work. It is a measure of the efficiency of these part-time studies that the majority of such students require only one more year of study for a degree than do students working in institutions for full-time study. The close relation between the economic needs of the people and their higher education is seen not simply in the system of part-time study; it is also seen in the strong encouragement given to all aspirants for university education to complete two years of productive work before beginning any full-time studies in an institution of higher education.

Institutions of higher education can be grouped under three headings: universities, institutes, and polytechnical institutes. Universities are concerned with the teaching of the humanities and the pure sciences, although in the smaller republics, universities often add the study of agriculture, technology, and medicine. Numerically, the largest group is comprised of the institutes, which train specialists in some single field, such as agriculture, economics, law, medicine, education, technology, art, drama, or music. The polytechnics offer the same subjects as the institutes but with the difference that several studies will be grouped together on a broader scientific base.

Higher education is financed and controlled by the Ministry of Higher and Specialized Secondary Education, and its policy is formulated in association with the State Institute of Planning for Higher Education. The large majority of the institutions of higher education have been set up by the government, and a minority by trade unions, cooperatives, and the Communist Party.

Centralization in the Soviet system

The head of each institution is a rector, appointed by the state for life. He is assisted by three prorectors responsible respectively for teaching, research, and administration. The rector is chairman of an academic board composed of the prorectors, the faculty deans, heads of departments, representatives of the academic staff elected by faculty boards, representatives of public bodies, and specialists appointed by economic or cultural organizations whose interests are involved with the education of specialists in institutions of higher education.

All candidates wishing to enroll in an institution must sit for a competitive examination. Industrial firms, collective farms, and state enterprises are entitled to send forward their best workers for admission; and the workers most successful in the entrance examination are entitled to a financial grant that is supplementary to the grants available to all entrants. In granting admission, the state gives a degree of preference to these worker candidates and to others who have completed two years of productive work after leaving school.

The duration of studies for a first degree varies from four to six years, the average being five years. The curriculum consists of compulsory, alternative, and optional subjects. A student chooses his specialist subject, and the compulsory courses will be in the basic and theoretical areas of that subject. To these are added alternative courses consisting of related disciplines or narrower aspects of the major specialty. To these two divisions a student must add courses chosen from a wide range of options, in which the emphasis is on recent scientific and technological research.

Examinations are the special responsibility of a state examinations commission, and the examinations at the end of a course include those in two or three basic disciplines related to a chosen specialty, one of which must be a social science. Candidates for a degree must also defend a thesis embodying the results of individual research or group research or a practical project. The scientific societies that exist in all institutions organize student research, which, under the guidance of professors or senior research workers, is regarded as an important aspect of undergraduate training.

At the conclusion of a first degree course, all students receive the same diploma—an award conferring identical rights as a specialist—but students with the best results are awarded a "distinction." Universities, institutes, and state research establishments all organize graduate schools with three-year courses for full-time students and four-year courses for part-time students. Postgraduate studies are concluded by a set of examinations for the degree of *kandidat nauk*, which include three written examinations and the defense of an original research thesis. The highest degree, the doctorate, is awarded for a thesis that records the solution to a major scientific problem on original lines.

Virtually all students are supported by government grants, a special feature of which is that the amount of grant is related to progress in studies. Students who achieve a "distinction" standard in their work receive a bonus award of as much as 25 percent of the basic grant. Students who are unable to live at home find accommodation in hostels at costs that they are able to meet from their grants.

OVERSEAS INFLUENCE OF THE MAJOR MODELS

The French model overseas. The various avenues of influence offered by a well-established model of higher education can be seen in the example of Tunisia. Even before the French went to Tunis, Muslim missionaries had established a religious university at El-Zitouna, which has persisted to modern times (its programs of philosophy, theology, law, language, and literature were recognized and incorporated as a Faculty of Theology and Religious Science in the modern University of Tunis). The agricultural and vinicultural potential of Tuni-

sia attracted French settlers, and it was for their sons and the able sons of natives that a French colonial government set up the École Coloniale d'Agriculture in 1898. Later in the colonial period, an Institut des Hautes Études was founded in Tunis to provide the early years of studies required by the University of Paris. Just prior to the grant of political independence this institute had grown to the point of providing full degree courses for a limited student body in science, letters, and law. Finally, in order to strengthen the force of French administrators and teachers, the government in Paris set up an École Nationale d'Administration, École Normale Supérieure, and Institut Bourguiba des Langues Vivantes in Tunis. At first, and for a number of years, all these institutions prepared students for the *licences* of the University of Paris, but by 1958 a new *licence èn lettres arabes tunisiennes* had been established, and by 1960 all existing institutions had been combined in a single University of Tunis awarding its own degrees. Even so, today after years of independence, higher education in Tunisia follows the French pattern very faithfully.

French missionary influences

The number of *instituts des hautes études* or even universities on the French model that had been established overseas by the middle of the 19th century was large, partly because even in such ex-colonial territories as Quebec, French institutions had survived, and partly because in the early years of the century the religious orders were restricted in France but were able to establish strong educational movements in Indochina, China, Japan, the islands of the Pacific, and Latin America. At that time, for example, 4,000 members of a single teaching order, the Christian Brothers, went to Latin America and with the Jesuits influenced first secondary and then higher education. The Federal University of Rio de Janeiro, when it was created out of a union of schools of law and medicine and a large polytechnic, emerged as an institution of French design. Later still, as the University of Brazil, it was taken as the model for all the universities of Brazil. In the case of Japan, in 1872, by a stroke of the imperial writing brush, all Japanese universities were remodelled on the pattern of the University of Paris. (In the following decade, though, Japan's Germanophile minister of education tried to incorporate German organizational and pedagogical patterns.) French influence remains strong in the former colonial territories of France, because, when such territories were part of France, their institutions were governed by French law, which included a detailed decree governing every aspect of higher education. During this period the academic staffs of overseas institutions were part of the cadre of French universities and were eligible for all their privileges—each faculty of every institution having its own administrative autonomy within the overriding governance of the Ministry of Education in Paris.

Germanic educational hegemony in the 19th century

The German model overseas. As French cultural influences declined in Europe, German influences grew strong. In the second half of the 19th century, Prussia, in particular, produced educational philosophers whose works were widely read. Even earlier, the German tradition of the wandering scholar had taken German ideas northward into Scandinavia, where their influence has remained strong ever since, and eastward up the Baltic coast into Russia. The Russian University of Dorpat (Tartu), for example, founded in 1802, used German as the language of teaching for most of the 19th century, and most of its professors were either Germans or had been educated in German universities.

In the second half of the 19th century emigrants from Germany to America, and American students returning from study in Germany, carried with them German ideas of higher education, particularly the emphasis on research and postgraduate studies. These indeed were the most important educational influences of that era. The Ph.D. degree, the first degree in a German university, from that time forward became the criterion of ability for the serious university student in America, and in the 20th century the popularity of this degree has influenced even countries that in the 19th century were not greatly influenced by German thought.

The English model overseas. The large groups of migrants from the British Isles who settled in Canada, Australia, New Zealand, and South Africa established universities almost as soon as they had solved their pressing economic problems. In Canada, colleges on the English model received royal charters as early as 1802 in Nova Scotia and 1828 in New Brunswick. In 1849 in Australia, a select committee appointed to report on the best means of instituting a university advised that the University of London should be taken as the model to be followed. When the University of Sydney was founded as the first of the Australian universities and the University of Otago as the first university of New Zealand, the English model was copied very closely.

In the 19th century and the first half of the 20th century, wherever the English sent military and trade missions, colonial administrators fostered the development of universities on the English model. The first plans for setting up universities in India grew out of the work of the secretary to the Bengal Council of Education in 1842. As in Australia, the University of London was taken to be the appropriate model to copy. As a mainly examining university, testing the standards of external students or affiliated colleges and free from religious bias, it seemed best suited to the needs of a subcontinent of wide distances and several creeds. In spite of this early start, 15 years of persuasion were required before the government in Westminster would agree to the incorporation of three universities in Calcutta, Bombay, and Madras. From India's point of view these universities and those that followed proved to suffer from the disadvantage that they were designed to implement two aims of the colonial administration, neither of which was proper for a university. The first of these was to transmit an alien culture (namely, English), and the second was to provide a way of measuring eligibility for government employment. It was not until political independence had been achieved and Indian graduates had taken over the control of higher education that Indian institutions began to be more responsive to Indian needs. Fortunately, the mistakes made in India in the 19th century were not repeated in Africa, the Pacific, Malaya, and the Near and Far East, where the story of university development as a cooperative enterprise is a happier one.

Anglicization of colonial schools

In the 20th century the University of London continued to play an important and growing role in the development of higher education overseas, in nurturing potential institutions toward the goal of complete independence. It did so because by a statute of 1858 it was authorized to examine for degrees any students who presented themselves, regardless of how or where they had studied, and to affiliate suitable institutions as colleges of the university with privileges similar to those afforded to its associated group of colleges in England. Over the years, the visits of external examiners overseas, the interchange of academic staff, and the award of scholarships to enable overseas students to study in England have all strengthened the influence of the English model overseas. These links have remained even as, one after the other, the territories in question have secured political independence. And they are kept strong today chiefly through the work of a body called the Inter-University Council for Higher Education Overseas, which is composed mainly of the vice chancellors of English universities. Cohesion and interchange is similarly promoted among the countries of the British Commonwealth by the Association of Universities of the Commonwealth.

The American model overseas. The United States, like France and England, has had charge of the development of colonial territories and has thus influenced their institutions of higher education. The story of this process is not, in principal, different in character from the similar story of change in French and English institutions overseas. To this colonial story must be added that of the influence of American missionaries who, in the 19th century, often went deliberately to territories in the Near and Far East, where they would not be in competition with French or English educational missions. To North Africa, Lebanon, Iraq, the Middle East, and above all

American missionary influences

to China, men and money for the development of education were sent in large amounts. The American University of Beirut and the medical school in Peking were typical examples of energetic developments of this kind. When Americans went to the Philippines, they found there several universities that had been built in the 17th century by the Spanish; today all Philippine universities and colleges broadly follow the American pattern. Military or economic intervention in such countries as Japan and Formosa led to the modification of indigenous institutions of higher education in directions that brought them closer to the American pattern. More recently in Africa, joint Anglo-American advisory missions or American missions alone have influenced higher education in Nigeria and Zambia.

The Soviet pattern abroad. All countries that have adopted a political system based on Marxist ideas have looked to the U.S.S.R. when seeking inspiration for the development of their institutions of higher education, and this applies particularly to the group of countries that lie between western Europe and the U.S.S.R. To this influence must be added the influence of the many specialist and advisory groups invited abroad from the U.S.S.R. for their higher education. Since 1960 three organizations have been particularly concerned with fostering closer relations between the U.S.S.R. and other countries in the field of higher education: the Soviet Committee of Solidarity, the Union of Soviet Societies for Friendship and Cultural Relations with Foreign Countries, and the large Patrice Lumumba People's Friendship University for foreigners in Moscow.

Organized Soviet cultural relations

Mixed influences. In many parts of the world, where institutions of higher education are of relatively recent growth, it is possible to trace the influence of several models of higher education. When a second university was founded in Nigeria, for example, it did not follow the English pattern of the University of Lagos but borrowed much more from the ideas of the American land-grant colleges. New universities in Zambia and Malawi have actually been planned by joint Anglo-American teams drawing freely upon their knowledge of two models of higher education. The developing countries could, in fact, learn much from multiple models: the English and French commitment to high standards of scholarship is needed wherever standards are in danger of erosion; the German emphasis on research is needed in situations in which educational miming is too easy; American flexibility and social relevance could be vital to countries that need more middle-level than high-level specialists; and Soviet methods of planning and part-time study are particularly relevant where financial resources are limited.

In contrast to the broad influence of several models today, there is sometimes a more direct influence as a result of the link between a single institution in one country and a single institution in another. There are many examples of this twinning in the modern world, such as the universities of Rabat and Bordeaux, of Rhodesia and Birmingham, and of Cornell and Djakarta. In this work, the influence may be stronger because it is based on personal links.

II. Organization of higher education

CURRICULUM IN HIGHER EDUCATION

Preparation for higher education. In most countries students rely on the secondary schools to prepare them for the next stage in their education, so that the pattern of higher education in a country is necessarily reflected in the organization and standards of its secondary education. In this matter there are great similarities between France and Germany. The *baccalauréat* and the *Hochschulreife*, which admit students to higher education, are both examinations that require a pass in nine or ten subjects at a high standard. The position is quite different in England, where a secondary school curriculum of increasing specialization has grown up in relation to university requirements for the "honours" degree. In the United States the position is closer to that of Europe since admission to an institution of higher education depends on the possession of about nine good high school credits. The advantages of the European and American patterns are that the first year of higher studies can be wide in scope, leaving the choice of a major subject of study until the student's inclinations and abilities have been disclosed. As against these advantages the English pattern avoids any considerable dropout of students at the end of the first year of higher education, and the wasteful repetition of some elements of secondary school work in the first year of higher studies is also avoided.

Subject patterns. The curricular patterns of higher education vary greatly from country to country, and in countries in which institutions have great autonomy they vary from institution to institution. The greatest contrast is to be seen between the array of optional subjects open to students of a junior college in America and the single "honours" subject in an English university. Between these two extremes there is great variety, and the situation in most countries is also one of great change and experimentation. Only in the traditional professional studies (where there is a need to relate the curriculum to a single vocational purpose) is there reasonable stability, although even in medical schools, for example, revolutionary changes are being tested at such institutions as Northwestern University in the United States and the international medical school on the Greek island of Kos. Because the present period is one of experiment, change is also accompanied by the desire for stability in the production of structured courses and for the return of unity into the untidy proliferation of subjects to which modern research has given birth.

Traditional departmentalization

Despite the pragmatic means taken in a few countries to bring coherence into the curriculum, the calendars of most universities still show subject disciplines organized as departments, in which new courses have been added to old mainly by a process of accretion. Ancient languages and their literature are still there, and to them have been added modern languages and their literatures. Philosophy remains from the distant past, and history from the 18th century. To these basic "humanist" studies others such as geography, music, fine arts, and architecture have been added. A new group of disciplines called the social sciences reflect the interest of scholars of this century in the study of society. They include anthropology, sociology, economics, and psychology. The older professional studies of law, medicine, pharmacy, and theology remain and to these have been added, many new ones, such as agriculture, veterinary science, and commerce. The older sciences of physics, chemistry, botany, zoology, and geology have been so enlarged by research that it has been convenient to break them up into smaller units like nuclear physics or organic chemistry, physical chemistry or geological spectroscopy. And perhaps the largest proliferation of courses has taken place in the rapidly growing areas of applied science and technology. Much thought is therefore being given in every country to the problem of combining depth of study with the need for coherence in the curriculum.

Teaching methods. The problems created by rapid advances in knowledge have led not only to structural reforms of the curriculum but also to new methods of teaching that emphasize principles rather than factual knowledge, and conceptual thought rather than mechanical skills.

Lectures, which still have their old merits in the teaching of large groups, are being supported or even replaced by seminars and tutorials. The seminar is generally a group of no more than 20 students meeting relatively informally under the chairmanship of a teacher to discuss a prescribed text or a paper presented by a student. In the seminar it is easy for a teacher to detect the difficulties of some students, and there is an impact of mind on mind that it is both remedial and stimulating. The tutorial is a meeting of one to six students, frequently based on reading or writing assigned by the tutor.

Audiovisual aids

Audiovisual aids are now widely used in lectures, seminars, tutorials, and small working groups involved in practical activities. Closed-circuit television may be the most flexible of such aids. In some large institutions this enables a lecture or demonstration to be made available

simultaneously to several different groups of students meeting in different places. The production of relatively inexpensive video-type recordings has made it possible to store illustrated lectures and demonstrations for future use. Some institutions not only have closed-circuit television equipment but also operate open-circuit systems to bring university teaching to groups some distance away, or to students following part-time courses of study at home. Open-circuit television can be combined with well-tried methods of teaching by correspondence course; it is already used in this way in the U.S. and the U.S.S.R. and in the recently organized Open University of Great Britain.

Examinations. In many countries examinations in higher education are given at the end of a period of study and determine what kind of award or degree that the student should receive. They also act, however, as an incentive to work, particularly if they take place from time to time during the course of studies rather than only at its conclusion. In some countries the examination at the end of the first year of studies is largely a selective one, to determine those fitted to proceed to a further period of study of two or four years, but in others it is used to determine the kind of specialization that will be best for succeeding years. England is typical of countries in which the final examination at the end of the course provides the main evidence for a degree award; the United States is typical of those in which the final examination carries much less weight and the main emphasis is on continuous assessment.

In professional schools it is almost universal practice for students to sit for examinations as and when sections of the course are successively completed. In science and the applied sciences, practical and laboratory exercises are used widely for examination purposes. A thesis and the defense of a thesis, which to some degree embodies original work or participation in research, form an essential part of the examination process for a first degree in Germany and Russia. The only generalization that can be made about examinations is that there is a general movement toward continuous assessment.

SOCIAL AND ADMINISTRATIVE PATTERNS OF HIGHER EDUCATION

Relations with governments. Everywhere in the world the cost of higher education has increasingly been met from governmental sources. When governments supply funds they must hold the recipients to a standard of public accountability; no government can fail to be interested in the degree to which its educational institutions supply it with the high-level manpower it needs; and some feel that their control must be absolute. If universities are to retain their freedoms, they must pursue them within a framework that brings a number of different and sometimes conflicting principles into the most fruitful balance.

Scholarly responsibilities of higher education

A measure of critical dissent emanating from universities can help to keep national institutions flexible and growing, but if this reaches the point at which university discipline or social order is threatened the fruitful balance will have been broken. The pursuit of truth is usually held to require a certain detachment from immediate political and social problems; yet if institutions are not deeply involved with the life of the region and the nation, they will fail to produce the well-trained talent that both need. Institutions of higher education are involved in the transmission of a cultural heritage, or perhaps in the revival of a failing national culture, and yet they are drawn by other loyalties into an international community of scholarship transcending national and regional differences. If a balance is not held, nationalism easily deteriorates from a desirable dynamic of political unity into the isolation of group hatred. By the same token, when a university teacher uses his position as a base from which to exert pressures toward purely political objectives, or when a government official uses his power to prevent a university from achieving its intrinsic aims, the desirable framework of balance is weakened.

Internal Relations. The charter granted in 1200 to the University of Paris secured for it a large measure of autonomy and the basis of a faculty organization of self-government, and this faculty organization has persisted almost unchanged in European universities until the present century—except curiously enough, at the University of Paris itself. In this type of organization, the teachers in a group of related disciplines meet regularly to organize their business under an elected dean. Deans and all senior teachers in all disciplines meet as a senate to regulate university business. In many universities, the head of the faculty of arts was at first appointed to be the chairman of the senate as well as the rector or academic head of the university. Later the senate more often elected its own rector, or he came to be appointed by an external authority.

Lay influence in academia

A modification of the faculty pattern of internal organization occurred as universities became increasingly involved in public affairs. At first a university chancellor was appointed by the authorities of church or state. With the growth of representative government the universities came increasingly to elect an officer called a vice chancellor or rector, and he came in time to be both the academic and administrative head of the university. As the vice chancellor took over the power and duties of his chancellor, the latter became the chairman of a mixed lay and academic body, generally called the council, which has remained as a most useful link between the activities of a university and the social, political, and economic life by which it is surrounded.

This European model of organization has undergone further modifications elsewhere. In the American approach, the president of a university combines some of the functions of a chancellor and a vice chancellor in the European model. His administrative role is stronger than that of a vice chancellor and his board of control (which approximates to the European council) has greater power over the appointment, dismissal, and conditions of service of university staff than is the case in Europe, where such matters are dealt with either by the university senate or by the officers of a government ministry.

In countries where institutions have been influenced by the Soviet model of higher education the modifications are more pervasive. Institutions of higher education are closely linked with a ministry of higher education and planning, and this gives a more hierarchical structure to their internal organization. Such a structure might easily develop some of the defects of bureaucracy were it not for the fact that individual institutions are increasingly seen as socialist collectivities with intrinsic traditions and objectives that need to be fostered and strengthened by external authorities.

Student initiative

The great expansion of numbers in higher education, coming at a time of acute housing shortage, has led students to organize in support of their welfare needs. In France, for example, nearly all students belong to the National Union of Students of France, which assists students' societies and presses their claims and needs upon all who are likely to listen and help. The Centre Nationale des Oeuvres en faveur de la Jeunesse Scolaire et Universitaire operates low-cost restaurants for students, gives assistance in finding lodgings, and tries to meet other welfare needs.

In West Germany the general movement of students from one institution to another has made the building of halls of residence one of the priority fields for the expenditure of public funds. The ultimate goal is to provide residential accommodation for about half of the student body at all institutes. Most halls of residence are maintained by students' welfare organizations, which also administer grants and loans, operate student restaurants, and provide health and counselling services and vocational guidance. They receive considerable government financial support.

In England there is a National Students Union resembling the French one. In addition, each institution has its own students union; these have become particularly powerful agencies for the furtherance of social life and welfare.

In the United States the concentration of university buildings facilitates the development of a vigorous stu-

dent's social life. A large campus may have student societies and organizations by the hundred, as well as student advisory and governing councils. A peculiarly American feature is the provision of many grants-in-aid to students willing to perform duties about the campus—serving as student assistants in libraries, laboratories, or administrative offices, or working in cafeterias and student centres.

In the U.S.S.R., participation in a wide variety of cultural activities in clubs and hostels is regarded as an integral part of higher education. A particularly Soviet characteristic is the degree to which famous writers, poets, scientists, actors, musicians, and artists actively direct these student activities. There is much emphasis upon the positive development of health, and all students are required to attend classes in physical education.

In 1968 a wave of student protest arose in most countries of the world. Its strongest manifestations were in France, Japan, and the United States, where large-scale clashes with the police took place. One of a number of student demands was for representation on the self-governing bodies of the universities. Student protestors were also critical of the content of courses of study, methods of teaching, examinations, and the ways in which subjects were grouped into courses; but many demonstrations had to do with broader political and national issues, involving armaments or wars or minority rights. In many countries some degree of representation has since been given to students, although the prevailing view is that in academic situations, in which individuals meet as teacher and learner, superior authority should remain with the teacher.

RESEARCH AND POSTGRADUATE STUDIES

Research and teaching. Teaching rises to a high level when it is shot through with the thread of discovery, and in institutions of higher education teachers are most successful when then can convey to their students something of the excitement that has come to them personally as they have worked at the frontiers of knowledge. But not all university teachers are equally good at teaching and research. Thus most institutions maintain some posts wholly for teaching or for research, though these are generally in a small minority of appointments.

Role of research in higher education

In highly industrialized countries most scientific and technological research is now carried on outside of the universities. Consequently there is an increasing degree of cooperation between governmental or industrial research bodies and the institutions of higher education; there are exchanges of students and staff and even common research projects. Sometimes a distinction is drawn between fundamental research and applied research, often with the implication that the main concern of teaching institutions is with fundamental research. But modern scientific work cannot be divided that way. Very little fundamental research is without practical application or social relevance, and often applied research forces its workers back into areas of fundamental research.

Much research is nowadays carried out by teams, sometimes large ones, in which the combined intelligence of a large body of workers is brought to bear on the solution of a problem. Along with this there is an increasing specialization of knowledge and skills. This situation has created a need for the wide and rapid dissemination of research plans and results, leading to a growth of coordinating bodies and their publications.

Postgraduate studies. Postgraduate schools in most countries (other than those in the French system) award a master's degree and a doctor's degree or their equivalent. In the shorter course leading to a master's degree the main element is acquisition of skills or knowledge at a more advanced level than required for a first degree. Work for the doctor's degree also involves participation in research and requires some account to be given of a piece of original research at its conclusion. The research frontier of most subjects advances so swiftly that only in two or three years of postgraduate study can a student be brought to the point at which he is capable of undertaking research himself. During these years he can continue to advance in knowledge and skill while undertaking a minor research project under guidance. Opportunities will also be opened to him by senior research workers who need the help of juniors in team research.

Not all students have the inventive gifts that make good researchers, nor need this be their aim in following a postgraduate course. As society grows more complex, many enterprises call for more specialized knowledge than can be acquired in a first degree course. Some students who proceed directly into employment return later for advanced study, and these now form an important element in postgraduate schools. Thus one finds, all over the world, postgraduate education provided in increasing measure to produce research workers and university teachers, to give refresher courses for older graduates, or to train those who require a greater degree of specialized knowledge than is given in undergraduate studies. Because postgraduate schools draw upon so many sources of recruitment and meet so many pressing needs, they are expanding in some countries at a rate faster than that of higher education in general.

Postgraduate schools always include among their students some who come from countries where higher education is only beginning to develop. Just as, in earlier years, these countries sent secondary school graduates to countries more advanced in the provision of higher education, so now they send graduates from their new universities for higher study abroad.

III. General trends of change

RECENT CHANGES

The world enrollment in higher education doubled during the 1950s. In those countries where a steadily rising standard of life may make possible an ultimate provision of higher studies for all who wish to follow them, growth is likely to continue. The more rapid growth of higher education in developing nations was expected to level off as the needs for trained manpower were increasingly met. At the same time there was an increase in unemployment among graduates. In 1950 unemployment was practically nil in nearly every country, but by 1970 the world percentage had risen to 10 percent, an average that concealed much higher figures in some of the developing nations. This figure did not include a rising degree of unemployment among graduates in such countries.

Differing national requirements

In many developing countries, particularly India and the Latin American countries, higher education has been oriented too much to the needs of the modern sector of the economy. It may be that until it is reoriented to the manpower needs of a mainly rural economy, and until that economy itself is transformed, higher education will not have a social function in such countries. The position is quite otherwise in countries where a balanced rural and industrial economy produces a steady rise in the national product. In these countries the continual emergence of new industries and the expansion of established enterprises call for new skills, more and more specialists, and for new and more extensive courses for graduates employed in industry.

A trend in higher education found everywhere is the steadily increasing number of women students. This has been particularly noticeable in the U.S.S.R., where between 1960 and 1970 the proportion of women in full-time and part-time education rose from 40 percent to 50 percent, and, in the nonscientific and nontechnological institutions, to as high as 80 percent. By comparison, statistics published by the United Arab Republic show that the percentage of women enrolled in institutions of higher education increased from 7 percent in 1950 to 14 percent in 1960. The increase in France over the same period was from 34 to 38 percent, a slower but still significant pace of advance. This trend must be expected to continue as the social, political, and economic disabilities under which women continue to suffer in many parts of the world are removed.

REGIONAL TRENDS OF CHANGE

North America. In spite of the already high proportion of North American youth who attend institutions

of higher education, enrollments in both Canada and the United States continue to rise year by year, and this trend has a compounding effect since all researches show that the educational attainments of parents influence the educational aspirations of their children. Not only are enrollments increasing but an increasing proportion of students now tend to go on to postgraduate studies. Overall, the institutional response has been the expansion of facilities in existing colleges and universities and the multiplication of newly established institutions of higher learning. In California, for example, all the existing branches of the state university in various cities are envisaging satellite campuses geographically and organically related to the branches.

The lack of a powerful central ministry responsible for higher education in either the United States or Canada has meant a diversity of academic standings has been difficult to avoid. There is now a trend for groups of institutions to develop coordinating organizations of various kinds, which are delegated powers to accredit institutions meeting agreed standards and which in different ways work to bring about more cooperation between institutions that previously had worked in isolation.

In Canada, two specific trends are caused by the demand for more specialized degree courses and for more postgraduate schools. In the past the reputation of neighbouring U.S. postgraduate schools was so high that there were always more Canadian graduates studying abroad than at home; this balance is now changing. There are also an increasing number of institutions organizing work-study programs in which a student alternates periods of study with periods of work in industry, government, or the professions.

Latin America. The main trends in Latin America reflect the prevailing desire to raise academic standards while at the same time relating courses of study and the choices of students to the economic development of the country and the region. A past imbalance between the overenrollment in courses in literary, legal, or sociological studies and the underenrollment in departments of science, technology, medicine, and agriculture has led to graduate unemployment. This picture is now changing and many institutions are making use of differential admission standards between departments in order to adjust the balance. This imbalance is also reflected in the shortage of university teachers in departments in which expansion is greatly to be desired. These shortages have often been met by the appointment of part-time teachers whose major employment and interests might be elsewhere than in the university.

Imbalance between humanistic and scientific enrollments

Western Europe. The institutions of higher education in Europe are engaged in the difficult process of transition from providing higher education for a selected elite to doing so for much larger numbers drawn from a wide spectrum of home backgrounds. As the economic and political ties between countries grow closer, there is a tendency to look more closely at different practices over the whole field of higher education and to seek for explanations of the degree of difference. This is beginning to lead to changes even of deeply rooted models like the English, French, and German, thus bringing them into a closer approximation of each other.

Increases in size and the pressure of student opinion are leading to a re-examination of historic forms of self-government and often to change. At the Sorbonne, for example, the senate and titular professors, which governed for so many centuries without change, has reformed its membership so that only 50 percent is now composed of titular professors, the other 50 percent being made up of students and junior teaching staff. In most institutions a democratization of committee procedures, even if less startling, is taking place.

Eastern Europe and the Soviet Union. There is an increasing trend in the countries of eastern Europe for separate countries to ratify bilateral cultural agreements with Western neighbours. These agreements make provision for joint educational conferences, in some cases for joint staff appointments, and for intervisiting by students and staff.

In the U.S.S.R. the effort to make higher education available to all who want it by 1980, is powering a major expansion movement. This has led to the steady development of facilities for students to continue in full-time employment, with periods of release for special study purposes. An increasingly sophisticated system of personal tuition has tended to improve the quality in this large sector of part-time higher education.

Re-emphases on work-study

Asia. The largest territory and the largest student population in Asia is that of the People's Republic of China, but because of the relative political isolation of that great land mass, development there has proceeded in unique ways. The only striking similarity is between the developments in the U.S.S.R. and China, despite their growing political antagonisms. The influence of the Soviet model on higher education in China has been profound and has now quite replaced the earlier work of foreign missionaries, other models, and influences coming across merely geographical barriers. As in the U.S.S.R., institutions have been developing into the three groups of universities, specialized institutes, and polytechnical universities. New universities and institutes have been founded in large numbers, but mainly in the coastal or near-coastal belt of the large cities. In Peking, for example, no less than 50 institutions of higher education have been set up. As in the U.S.S.R., nearly all these institutions promote correspondence and evening classes for students who wish to study for a degree while remaining in productive work. In addition, there are "spare time" universities as well as "work-study" universities on the American model.

All this development in China went on steadily until 1965, when political relations with the U.S.S.R. began to deteriorate. At the same time a "cultural revolution," with the Red Guard and students at its spearhead, began to disrupt all institutions of higher education. There followed a protracted period of violence and disorder in higher education in which thousands of lives were lost, the economy was disrupted, and education brought almost to a standstill. By 1968 the extremes of disorder had produced a reaction, and in July the army began to intervene to suppress the cultural revolution. The army was successful and by 1970 professors and teachers driven out by the young Red Guard were reinstated, and normal academic life was slowly resumed in institution after institution.

Across the Yellow Sea, in Japan, a period of rapid expansion in higher education has also been retarded by episodes of student violence, but the violence has not seriously held back the speed of progress, particularly in the field of technology education.

In the subcontinent of India and over the numerous and widely spread countries of Southeast Asia, the main concern has been to combine expansion with a process of decolonization in higher education. In all these countries the fundamental feature of higher education had been that it was imported from the West with ready-made value systems crystallized in institutions and techniques. Criteria for admission, curricula, teaching methods, and constitutional structure were better related to conditions in Western industrial nations than to the financially poor countries of Asia. Even late reports of visiting commissions recommended development along foreign lines. The Carr-Saunders Commission for Malaya and the Darwin Commission for Thailand recommended British practices, and the Allen Report recommended American practices for higher education in Indonesia. Now the trend is all the other way, and innovations are striking roots deep in native soil. One result of all this has been to destroy the image of a university as a place where students learn to pass examinations in a foreign language and make it instead, one of a place where attitudes are changed, values created, and national traditions made strong and relevant.

Nativism and nationalism

A second trend in all Asian countries arises from linguistic nationalism—that is, from the desire to make the medium of instruction in higher education a native and not a foreign language. Unfortunately, some new universities have tried to use a native language even though it

has led to such seeming absurdities as training Javanese chemists through the medium of Malay, although Malay is neither a language native to the Javanese nor a language with any technical vocabulary or technical literature. In some Asian institutions there is parallel teaching in a native and an international language. In others an international language has been selected (generally English) to be taught in the secondary schools so that university students will have acquired a sufficient reading ability (if not verbal ability) to avoid the need for an expensive translation service. This policy has been followed in Thailand with growing success. The Japanese have shown that it is possible to learn to read a European language very quickly without speaking it at all.

The emphasis upon a regional language does create formidable problems, but it has one advantage in that it may lead to the development of genuine regional research. Thus have Indian and Indonesian institutions conducted village surveys that have provided valuable information to their respective governments.

Africa. The trends of development in Africa differ considerably as one moves from countries north of the Sahara to countries of tropical Africa and then further south to the Republic of South Africa.

In the countries north of the Sahara political independence came relatively early so that the expansion of higher education that so largely determines development in European countries has also begun to affect the countries just south of the Mediterranean in the same way. These countries, like those of the subcontinent of India, have begun to deal with the problem of decolonization and are tackling the problem of the language of instruction in higher education relating higher education more closely to the nation's need of specialists, replacing foreign cultural aims by those of an indigenous culture, and replacing foreign academic staffs by graduates from their own universities.

In the countries of tropical Africa the most important stimulus to development came from a Conference of representatives from 31 African nations and 14 non-African nations (including both the U.S. and the U.S.S.R.), which was held in the Malagasy Republic in 1962. The Conference produced a detailed and comprehensive plan of development for higher education in Africa over a 20-year period. Expansion targets were proposed and the financial implications of the plan worked out by some of the best economists in the world.

Since 1962 tropical Africa has seen the development of new postgraduate schools, the setting up of regional centres for the training of laboratory technicians, the development of library interchange, and the standardization of degrees and diplomas. To carry such work forward, a new permanent body, the Association of African Universities, has been set up.

South of the Limpopo River, the Republic of South Africa pursues a political policy of "separate development" for different racial groups, which has led to a pattern of educational advance different from that of any other part of Africa. In the general expansion, separate, new, ethnic universities have been created for the Bantu and Coloured populations of the Republic. The effects of racial discrimination in higher education, like those of religious discrimination in earlier centuries or of political discrimination in the 20th century, can lead inevitably to a debasing of the academic currency. A nation's university degrees, like its money, must be acceptable in other nations, and racial, religious, or political discrimination inevitably leads to the loss of independent-minded teachers so that first the quality of its research, and then that of its teaching, declines. Institutions of higher education are unlike institutions in other sectors of education in their essential need to maintain bonds of loyalty not only with their country of origin but also with the international company of universities.

Necessity for racial, religious, and political tolerance

BIBLIOGRAPHY. E. ASHBY, *Community of Universities* (1963), a study of the universities of the British Commonwealth and of the coordinating work of the Association of Universities of the British Commonwealth; *Technology and the Academics* (1958, reprinted 1963), a study of the impact of the scientific revolution on the universities of Great Britain and of the gulf that exists between the study of technology and all other higher studies in universities, and with M. ANDERSON, *Universities: British, Indian, African* (1966), an historical account and analytical survey of higher education in India and the English-speaking countries of tropical Africa; K. BOEHM (ed.), *University Choice* (1966), a symposium on the spectrum of disciplines in higher education studies in relation to the career choices of students and the needs of the modern world; N.E. FEHL, *The Idea of a University in East and West* (1962), a comparison of eastern and western ideas concerning the concept of higher education; W.D. WEATHERFORD (ed.), *The Goals of Higher Education* (1960), a symposium by American scholars on the aims of higher education in the U.S., particularly in relation to contemporary conflicts of philosophy; M.G. ROSS, *The New University* (1961), a study of the aims and plans of a group of Canadian scholars concerned with the founding of a new university, and (ed.), *New Universities in the Modern World* (1966), the study of ten new universities in Australia, England, Canada, Africa, India, and the U.S.; A.S. NASH, *The University and the Modern World* (1945), a critical analysis of the unconscious philosophy upon which the modern university has developed its life and the elaboration of a new philosophy relevant to the needs of the modern world; M.A. CLAPP (ed.), *The Modern University* (1950, reprinted 1968), a series of essays on the development of 19th-century ideas of the university in Europe, England, and America in relation to modern problems and trends in higher education; B.E. MELAND, *Higher Education and the Human Spirit* (1953), a study of the degree to which the day-to-day work in institutions of higher education develops humane and religious insights; H. SMITH, *The Purposes of Higher Education* (1955), a study of the purposes of higher education in a free society; B.A. FLETCHER, *Universities in the Modern World* (1968), a comparative study of higher education and of trends of change in the main regions of the world; J.S. BRUBACHER, *Bases for Policy in Higher Education* (1965), an attempt to formulate a comprehensive statement of the principles on which higher education should be based; R.R. FIELDS, *The Community College Movement* (1962), a historical and contemporary picture of the two-year junior college and its development into the community college; A. KERR, *Universities of Europe* (1962), a comprehensive analysis of universities and of student life in all the countries of Europe; T.H. SILCOCK, *Southeast Asian University* (1964), an account of the development problems of the universities of Southeast Asia and of four Western university models that have influenced this development; A.M. CARR-SAUNDERS, *New Universities Overseas* (1961), a description of the development of university education since 1946 in all the British territories that had colonial status at that date and of their contribution to the solution of problems created by political independence; O.C. CARMICHAEL, *Universities: Commonwealth and American* (1959), a study of the different ways in which Commonwealth and American universities have grown to meet the needs of a society in which science and technology are requisite to stability and progress; V.H.H. GREEN, *The Universities* (1969), a detailed study of the history and character of university life in Great Britain; W.H.G. ARMYTAGE, *Civic Universities* (1955), a historical account of the rise of 15 civic universities in Great Britain since their foundation in the second half of the 19th century; A.W. GRISWOLD, *In the University Tradition* (1957), a series of essays on the aims of higher education; N. SANFORD (ed.), *The American College* (1962), a psychological and sociological study of many varied aspects of life in an American college; C. JENCKS and D. RIESMAN, *The Academic Revolution* (1969), the story of the revolution by which the academic profession in America freed itself from effective lay control; A.G. KOROL, *Soviet Education for Science and Technology* (1957), a survey of the aims and organization of higher education in the Soviet Union in relation to scientific–economic planning.

(B.F.)

High-Pressure Phenomena

High-pressure phenomena are those changes that occur in ordinary matter when subjected to pressures that may range from ten to millions of times greater than normal atmospheric pressure. Under such conditions, matter is compressed, and chemical and physical properties are altered; the material may even become unrecognizable. Laboratory studies conducted at these pressures may also involve temperatures that range from near absolute zero (−273° C [0° K]) to far beyond the melting point of any element under normal conditions.

Definitions. Pressure is measured as a force per unit area. The unit of pressure is the bar, which is approxi-

mately equal to the sea-level pressure of the atmosphere on an average day. More precisely, it is equivalent to 0.98692 atmosphere (standard). Specifically, one bar equals 1×10^6 dynes per square centimetre (one dyne is the force that gives a mass of one gram an acceleration of one centimetre per second). It is also equal to 1.01972×10^3 grams per square centimetre (14.504 pounds per square inch). The range of pressure in nature is extreme: at the deepest ocean depths (about 11 kilometres [seven miles]), the pressure is about one kilobar (1,000 bars); at the base of the earth's crust (about 40 kilometres' [25 miles'] depth), the pressure is approximately ten kilobars; at the base of the earth's mantle (about 2,900 kilometres [1,800 miles]), the pressure is about 1.3 megabars (1,300,000 bars); and at the earth's centre, it is at 3.4 megabars. Pressures at the centre of Saturn and Jupiter are estimated to be 20 and 100 megabars, respectively, and those of some stellar interiors are thought to be 10^{10} bars or greater. While it is exceedingly difficult to understand those laws of physics and chemistry that apply to material properties over the wide range of pressure that occurs in the universe, present technology does make it possible to study materials at all pressures found in the earth. A few studies using explosive techniques have even attained pressures greater than those at the centre of Saturn.

Range of pressures

History. The amount of pressure necessary to create similar effects on different materials is highly dependent on the materials involved. Whereas a pressure change of 0.2 bar on a gas, for example, causes the volume of the gas to change by 20 percent, a similar volume change requires 10,000 bars in a liquid and 200,000 bars in a solid. Hence the selection of materials and effects to be studied has been strongly influenced by the historical development of high-pressure apparatus.

Before the 20th century, when research was concerned primarily with the effects of pressure on gases and liquids because of apparatus limitations, considerable work was done at pressures of three kilobars and over a temperature range of 200° C. Typical of such investigations were those concerned with the volume behaviour of liquids and with the existence of gas–liquid critical points (the combination of temperature and pressure at which the specific volumes of a gas and liquid become identical, thus forming one phase). At the critical point the temperature and pressure are such that the transition from a gas to a liquid takes place with zero change in volume and energy. This transition involves two factors: (1) the critical temperature, which is the highest temperature at which a gas can be liquefied by pressure; and (2) the critical pressure, which is the pressure of a gas at its critical point. The important question of a critical point between liquid and solid matter was also raised during the 19th century, but the pressures available for research at that time were far too small to settle the matter.

Modern approach to research

Modern high-pressure research was made possible primarily by the work of Percy Williams Bridgman, a physicist in the United States who received the Nobel Prize in 1946 for his investigation of high-pressure phenomena. Except for those techniques that involve the use of explosives, all current high-pressure apparatus is derived from equipment that Bridgman designed and used (see below *Principles of apparatus design*).

PHYSICAL AND CHEMICAL EFFECTS OF HIGH PRESSURE

In order to describe high-pressure phenomena over extremely wide ranges of variables, it is necessary to use mathematical statements that relate small increments of the various changes taking place simultaneously during the overall transformation of conditions and reveal graphically what is occurring.

On typical single-component systems. The three classical phases, or states, of matter consisting of atoms and molecules are solid, liquid, and gas. (There is a fourth, plasma, consisting of atoms stripped of their electrons, and several others are recognized.) Most phase changes are familiar phenomena: a gas condensing into a liquid, a liquid evaporating to a gas, a solid melting to a liquid, and a liquid freezing into a solid. Modern high-pressure research in phase changes, particularly the work of Bridgman, has revealed that a myriad of transformations occur as a result of atomic rearrangements, many of which involve gross changes in the physical properties of a substance.

Gibbs free energy and high pressure. Two or more phases can exist in equilibrium, a balanced coexistence that is governed by thermodynamic (heat) relationships. These relationships are formulated into a concept called Gibbs free energy, which is the difference between the heat content of a substance (called enthalpy) and the product of its temperature and entropy. (Entropy, the energy of disorder, is a measure of the unavailable energy in a thermodynamic system; if left to itself, the entropy of any closed system increases, while the available energy decreases.) The equation for this thermodynamic relationship is written in symbols: $G = H - TS$, in which G is the Gibbs free energy of the phases in equilibrium, H is the enthalpy, T is the absolute temperature, and S is the entropy of the phase mixture. As the equation indicates, at any constant pressure and temperature, when the phases are in equilibrium, the mixture has the lowest Gibbs free energy, relative to the energy of the mixture at any other configuration.

By differentiating this equation it may be seen that there is a simple relationship between Gibbs free energy and the volume and entropy of the phase. The partial derivative of the free energy with respect to the pressure (P), the temperature (T) being held constant, is equal to the volume (V). This is written $(\partial G/\partial P)_T = V$. The partial derivative of the free energy with respect to the temperature (T)—the pressure (P) being held constant—is equal to minus the entropy (S). It is written $(\partial G/\partial T)_P = -S$.

If entropy and volume are known as a function of the pressure and temperature for two phases, these equations may be integrated and an equilibrium-phase boundary thus calculated, separating the regions in P–T space (a phase plot of pressure versus temperature) in which the two differing states are stable. Stability does not imply that a phase transformation will take place, however; for example, with care, most liquids may be preserved as liquids for lengthy periods of time below their freezing points, at temperatures where the solid is the normal stable phase. Such supercooling (or any other such sluggish transformation) is known as metastable behaviour, marked by only a slight margin of stability for that phase, because any disturbance will bring about a transformation to the stable phase. A supercooled liquid, for example, becomes solid rapidly if a minute crystal is dropped into it.

Supercooling

Effects on melting. Melting phenomena are best discussed with reference to the above equations. Most melting curves have been studied only up to less than 100 kilobars by static apparatus; melting has been detected, however, at higher pressures in some dynamic shock-wave experiments on copper and on lead. The tangent to the melting curve that results when pressure is plotted against temperature, $\partial P/\partial T$, is given by the ratio of the change in entropy (ΔS) and the change in volume (ΔV) between solid and liquid, or $\partial P/\partial T = \Delta S/\Delta V$. Thermodynamics shows that the liquid, being the high-temperature phase, has the greater entropy. Usually, because the liquid has the larger volume, both the change in entropy and the change in volume are positive. This is not true for water, bismuth, and a few other substances.

In most cases, pressure raises the melting point, and, most frequently, the liquid is more compressible. Therefore, as pressure is increased, the change in volume decreases, and the rate of temperature increase of the melting curve is slowed. In a few cases, however, such as cesium and potassium nitrates, a maximum temperature is reached at which the liquid becomes denser than the solid. Thereupon, any further application of pressure decreases the melting point. If a transition in crystal structure takes place, thereby transforming one type of solid into another (*i.e.*, if there is a solid–solid transition in a solid), then a triple point occurs on the melting curve at which solid (1), solid (2), and liquid are all in equilibrium at one pressure and temperature. After continuing past

this break in the melting curve, the process may be repeated.

Both cesium and potassium nitrates actually have two maximums in their melting curves before a third solid phase intervenes. To illustrate further the occasional complex behaviour of matter, the melting temperature of cesium decreases with pressure until a minimum is reached, after which the melting temperature increases. In this case, even though the solid is less dense than the liquid, it has a lower initial compressibility, hence the volume of solid and liquid approach each other and become equal at the minimum point.

Effects on polymorphism. Polymorphism is the phenomenon in which a substance can exist in more than one form. Diamond and graphite, which are radically different forms of carbon, and calcite and aragonite, two forms of calcium carbonate ($CaCO_3$), are examples of polymorphic substances. Although this aspect of solids had been known for over 200 years, it was not until Bridgman extended the range of pressure studies that it was realized how universal this behaviour of matter is. In polymorphic substances, a thermodynamic rule states that for two-phase equilibrium, at constant temperature, the high-pressure phase must be denser. In a crystalline solid, one obvious way of obtaining a denser phase without chemical change is to increase its coordination number (the number of neighbour atoms surrounding a given atom). The common form of iron (alpha iron, or α-iron), for example, is a body-centred cubic crystal at room temperature; and, hence, it is eight coordinated (each atom has eight atoms around it positioned as though at the eight corners of a cube). One might anticipate, therefore, that, at some unspecified elevated pressure, a closer packed form of iron might become stable. Such a form of iron, which is 12 coordinated, has been found at 130 kilobars in shock-wave studies. X-ray crystallography has shown it to be hexagonal close-packed, not cubic close-packed, as the gamma iron already known to occur at elevated temperatures. It is paramagnetic (a weak form of magnetism), unlike normal ferromagnetic α-iron.

Coordination number. The statement that an increase in a substance's coordination number results in a higher density, however, is true only if the atoms in the higher coordinated phase do not have drastically larger radii. For some elements—*e.g.*, titanium, zirconium, and ytterbium—the converse is true. In these cases, application of pressure to an initially close-packed 12-coordinated element changes it to a more open-body eight-coordinated cubic, the atomic radii of which are smaller. The tetrahedrally coordinated elements germanium and silicon are semiconductors (materials differing from metals in that their electrical conductivity increases with temperature), the electrical conductivity of which becomes metallic when they adopt the structure of metallic white tin at high pressure, which is 20 percent denser. Other atomic arrangements are known for these elements. Indeed, white tin itself changes to a tetragonal form.

Phase diagram. To illustrate further, the complexity of polymorphism, the phase diagram for water is shown in Figure 1. On this scale the gas phase cannot be shown, but ice under different conditions is identified by roman numerals. Ordinary ice (I) is less dense than the liquid; hence pressure lowers the melting point until a triple point is reached for liquid ice (I) and ice (III). Because ice (III) is denser than liquid, its melting point is raised by pressure until a further triple point is reached. Ice (V) is also denser than its liquid, so that "normal" melting behaviour is achieved. From Figure 1 it may be seen that an application of 6.5 kilobars to water at 0° C is sufficient to freeze it to a different form of ice, ice (VI). A phase diagram such as Figure 1 represents only equilibrium behaviour; in the case of water, as in most other systems, reaction does not necessarily occur, because as conditions are changed, metastability may result. Diamond is an example of a metastable form of carbon. With water, ice (II) may be chilled to the temperature of liquid nitrogen, after which the pressure may be released, because metastable ice (II) does not transform to ordinary ice until its temperature is raised.

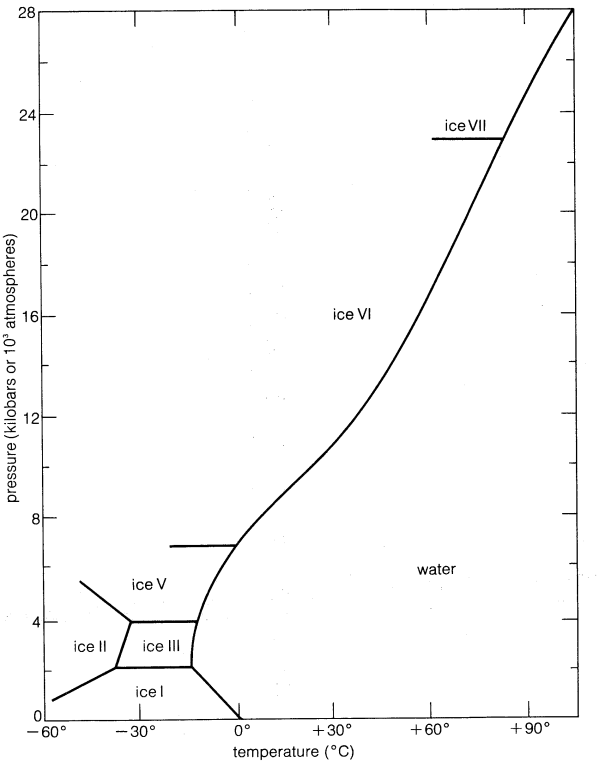


Figure 1: Phase diagram of water (see text).

Shock-wave techniques have provided considerable insight into the volume behaviour of matter. They reveal how cations (positive ions), much more than anions (negative ions), influence the pressure at which the coordination structure changes. Sodium chloride (NaCl), for example, in its normal six-coordinated structure has a density of 2.165 grams per millilitre at one bar, 2.248 at ten kilobars, 2.731 at 100 kilobars, and finally 3.349 at 300 kilobars, at which it transforms to the eight-coordinated modification. In potassium chloride (KCl) the same transition as that in sodium chloride occurs at 18 kilobars, in rubidium chloride (RbCl) at four kilobars, whereas cesium chloride (CsCl) is in this form at one bar. The effect of the chlorine anion on these transformations is trivial by comparison with the cation (*e.g.*, sodium, potassium, or other). Shock-wave techniques have also been used to study the volume behaviour of many elements at a pressure of two megabars. At this pressure the densities of most metals have been increased by 60–100 percent, temperatures of 2,000° C are attained, and all thermodynamic properties of the materials may be obtained for use in calculating chemical equilibrium. By using the Rankine-Hugoniot relations (see below *Principles of apparatus design*) the internal energy (E) and the enthalpy (H), with $H = E + PV$, are obtained directly from the shock-wave data. The entropy (S) may be obtained from an equation relating change in energy to pressure, change in volume, temperature, and change in entropy, as follows: $dE = -PdV + TdS$, but not directly; an approximation has to be made, as must also be done for temperature. From these expressions, Gibbs free energy is derivable. Further thermodynamic considerations permit reduction of the data to an isotherm (constant temperature), so that thermodynamic functions covering a large rectangle in P–T space become known. In liquid oxygen, for example, the upper corner in P–T space is 10,000° C and 1.4 megabars.

Effects on minerals. Although the necessary apparatus had been in existence for some time before World War II, it was not until afterward that there was a major effort to solve high-pressure mineralogical problems in the laboratory. Because the entire range of geological conditions that exist in the earth's crust (namely, pressures up to ten kilobars and temperatures up to 1,500° C) may be duplicated in the laboratory, it is possible to investigate natu-

ral reactions in which pressure is a factor. Moreover, by using such fluids as water and carbon dioxide as pressure mediums, physical properties as well as chemical reactions can be studied. Among the geological conditions that can be measured are the effect of pressure on the electrical conductivity of minerals and on various Curie temperatures (a critical temperature above which a ferromagnetic substance becomes paramagnetic). Much laboratory effort has also been expended on measuring the effect of pressure on the propagation of longitudinal and transverse elastic waves through the earth. It has been found that such measurements, to be meaningful, must be made under pressure; this is because, at a pressure as low as one bar, the initial porosity and lack of compactness of rocks make it impossible to determine values of wave velocity between different samples of nominally the same rock type. In those cases in which the deformation (under confining pressure) of at least a few rock types has been studied in detail in the laboratory, pressure enhances rock flow and helps to explain the tortuous nature of certain folded rocks in nature.

Duplicating nature in geological research

With the development in the mid-1950s of the apparatus illustrated in Figure 2, conditions duplicating those in the earth's upper mantle could be achieved. There then followed a major change in geophysical concepts, such as the previously unrealized importance of polymorphic changes induced by high pressure. An indication of what was to come was the discovery of a denser (2.91 grams per cubic centimetre versus 2.65 grams per cubic centimetre) form of quartz (coesite), first in the laboratory and subsequently at meteorite craters. Later, a still denser (4.29 grams per cubic centimetre) modification, now called stishovite, was found. Thus it was seen that a common constituent (quartz) of surficial rocks could exist in a form that was 62 percent denser at depths of about 400 kilometres. Similar pressure studies have revealed modifications of olivines, pyroxenes, iron oxides, and feldspar in which density increases of 25–50 percent are not uncommon. As a result of such studies, it is safe to say that the mineral components of the earth at a depth of 1,000 kilometres—corresponding to a pressure of 400 kilobars—will contain completely different crystal structures than their chemical counterparts found in the crust. There may be a few exceptions, however, such as magnesium oxide (MgO) and carbon in the form of diamond.

Laboratory studies also destroyed the belief that the rise time of a shock wave is too rapid to allow a polymorphic transition to occur. Many mineral transitions are found to take place—even diamonds have been made—when explosives are used. Stishovite has also been recovered from explosive studies, and its occurrence at meteorite craters is thought to have been caused by natural explosive events when the meteors struck the earth.

Effects on magnetic and electrical properties. Solids are classified as insulators, semiconductors, or metals, depending on the magnitude of their electrical resistivity and its temperature coefficient (*i.e.*, the change of resistivity with temperature). Although such insulators as magnesium oxide and certain alkali halides have been studied while under explosive shock, comparatively few studies have been made of the effect of pressure on the electrical resistance of insulators because of experimental difficulties. A common method of estimating temperatures in the earth is to compare the earth's known electrical conductivity, which varies as a function of depth, with the conductivity at various temperatures of those minerals thought to exist there. Although olivine, which has a wide distribution in the earth, has been studied extensively in the laboratory, no definitive statement concerning the effect of pressure on electrical conductivity can be made. In general, however, one can say that, if the electrical conductivity in a substance is the result of ionic motion, pressure should inhibit the motion and thus increase the resistivity. Similarly, diffusion rates should decrease. If, on the other hand, the process is electronic —*i.e.*, the result of motion of an electron or a hole (a positive charge caused by the absence of an electron in a conductive band)—pressure may either increase or decrease the electrical resistance.

Electrical effects

Semiconductors are much more sensitive to pressure than other metals, because pressure can have gross effects on carrier density or mobilities or both through an alteration of the conductive-band structure. Changes of several orders of magnitude in resistivity can occur; for instance, 30 kilobars at room temperature changes the resistivity of the semiconductor germanium by a factor of 100. Even more startling changes in resistivity happen when a phase change occurs, as, for example, when at about 120 kilobars germanium adopts the white-tin structure and becomes a metallic conductor. Even in metals, the high density of charge carriers, the free electrons, can react in a complicated manner with the nuclear potentials, causing great variability in behaviour when pressure is increased. Strontium, for example, has maximum resistivity at about 40 kilobars, whereas calcium has a maximum at over 300 kilobars. Barium's electrical resistance is at a minimum near ten kilobars and at a maximum near 130 kilobars. Similarly, antimony has a maximum near 25 kilobars, whereas arsenic has a minimum near that pressure but a maximum near 60 kilobars. Phosphorus (the black form), which has the same number of electrons as arsenic in its outer atomic shell, adopts the arsenic structure at an elevated pressure and then is transformed to a simple cubic structure around 110 kilobars. Presumably, both latter forms of phosphorus are metals, although at a pressure of one bar phosphorus is a semiconductor.

Magnetic effects

Few studies have been made of the effect of pressure on the magnetic properties of matter. In the range below 30 kilobars, where hydrostatic mediums may be used, the effect of pressure is slight. In the range above 30 kilobars, where solid mediums must be used, experimental design and interpretation of data are difficult. Nevertheless, the Curie point of iron has been found to be unaffected by pressure up to the point at which paramagnetic γ-iron forms from ferromagnetic α-iron. In a few other elements, pressure causes moderate fluctuations in Curie points.

APPLICATIONS OF HIGH PRESSURE

The most important economic use of "high" pressure, namely in refining and plastic industries, will not be discussed here, because the pressures employed are lower than those with which this article is concerned. The primary industrial application of truly high pressure is in the manufacture of diamonds from other forms of carbon. In the belt apparatus (Figure 2A), carbon can be transformed to the diamond modification at about 55 kilobars and 1,500° C in the presence of a transition metal (one with an incomplete inner shell) such as nickel. The small diamonds made this way are of industrial grade. Although the same technique can be used to manufacture gem-quality diamonds, the time involved makes this process uneconomical at present. A denser form of boron nitride, with diamond-like hardness, has also been manufactured under pressure.

Another use of high pressure is to alter the mechanical properties of materials, such as increasing the ductility of a metal. In one such application, a new high-speed wire-drawing system has been developed in which external hydrostatic pressure is used to reduce the size of the wire in one die, thereby replacing a whole series of reduction dies that had formerly been needed. Attempts are also being made to compress hydrogen or its compounds in order to create a stable form of hydrogen metal in the solid crystalline state.

The comparative paucity of major applications of ultrapressure technology is probably not because of any lack of phenomena but rather because of the newness of the field and the magnitude of the area to be surveyed. To map out the temperature properties of a hypothetical alloy at normal temperature and pressure, for example, might require 20 samples that differ in composition and, thus, 20 experiments. To make the same temperature study for the 20 samples at intervals between ten kilobars and 200 kilobars, 400 experiments would be necessary.

Principles of apparatus design. *Use of hydrostatic pressure.* In 1909 Percy Williams Bridgman published a

description of a leakproof packing that would extend high-pressure research from three kilobars to 30. Although there had been no problem previously in finding theoretical ways to obtain high pressures by hydrostatic means, no one had discovered how to keep a fluid at such pressures from seeping out of the system. To circumvent this problem, Bridgman used the principle of unsupported area in his packing, which is annular; *i.e.*, shaped like a ring. The pressure of fluid is against one side of the disk, which in turn presses on the annulus. Because the annulus has no material in its centre, its area is less than that of the disk. Hence the pressure per unit area on the annulus is greater than on the disk. Thus the fluid, which is confined behind the disk, cannot escape past it. With this packing arrangement, the sealing of gases and liquids was no longer a problem, but then the strength of containers and plungers became an important factor. The problem with plungers was solved by the use of cemented tungsten carbide, which rarely fails for pressures under 50 kilobars.

Containers. As for containers, such as heavy-walled steel cylinders that may not burst until 40 kilobars of internal pressure, they suffer plastic deformation starting at 15–20 kilobars. Therefore, they are inefficient for repeated use. Bridgman devised a scheme for overcoming this weakness, by fashioning a conical container. As pressure is increased in the interior of the container, the conical vessel is forced into a conical sleeve, because this creates a stress on the exterior of the interior vessel that is approximately equal to the pressure on the inside; thus, deformation is avoided. Bridgman modified this principle of external support by immersing an entire small pressure vessel in a fluid at 30 kilobars. This enabled him to perform at pressures as high as 100 kilobars. More recently, by using the same modification principle, pressure generators have been developed, as illustrated in Figure 2. Figure 2A is a diagram of the belt, an apparatus

From H. Drickamer, "Scientific Research at Very High Pressures," *Physics Today* (November 1964)

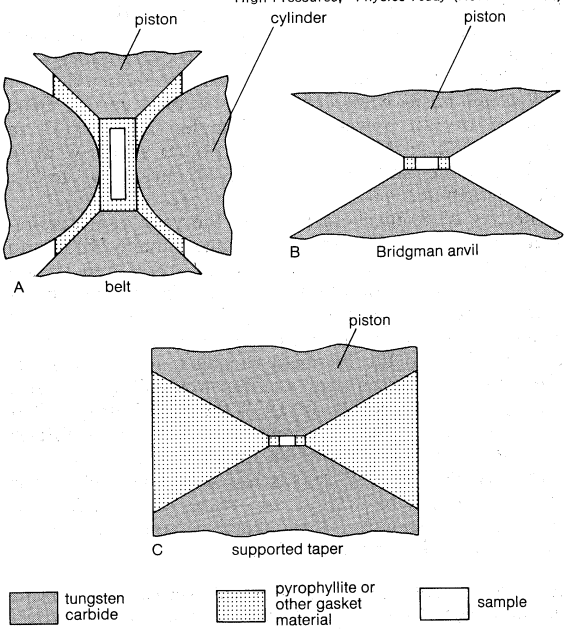


Figure 2: Schematic apparatus designs.

developed in the United States by the General Electric Company for manufacturing the first man-made diamonds. Devices of this general type have been used to achieve pressures of about 200 kilobars and a temperature of about 3,000° C (but not concurrently). Bridgman anvils, which are shown in Figure 2B, were developed to study the electrical resistance of materials up to 100 kilobars. Their chief use has been by geologists in studying mineral reactions in the laboratory; generally, a furnace is added externally to provide elevated temperatures. Figure 2C shows a modification of Bridgman anvils. By supporting the tapered faces, it has been possible to achieve the greatest static (nonexplosive) pressure to date, about 500 kilobars. These and similar designs are successful for two reasons. One is termed "massive support"; that is, a high pressure is exerted on only a small region of a massive object, as in all three designs of Figure 2. The second reason is that a comparatively soft gasket, such as that filling the space between piston and cylinder in Figure 2A, seals against high pressure by its own internal friction as well as by friction against the walls, but only as long as the piston–cylinder gap is not too great.

Shock-wave technique. The greatest extension of pressure and temperature studies began after World War II with the development of controlled explosive events. Even higher pressures and temperatures are now being achieved with nuclear devices. Studies with explosives, termed shock-wave research, are excellent illustrations of how an experiment may be designed to yield valuable information that can be interpreted without mathematical complexity. The basic tool for such studies; displayed in Figure 3, is called a plane-wave generator. When detonated, the velocities of the two explosive charges in the device are such that, due to the conical angle, the velocity at the edge of the slower central explosive wave is exactly matched by the faster wave of the outer explosive. Thus the inner wave is kept exactly plane, and any waves it induces in another pad of explosive and the sample assembly are also plane, at least for a known distance.

Plane-wave generator

From K.A. Gschneidner, M.T. Hepworth, and N.A.D. Parlee (eds.), *Metallurgy at High Pressures and High Temperatures* (1964); Gordon and Breach

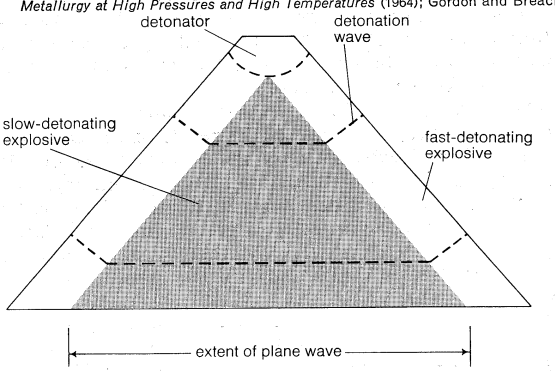


Figure 3: Plane-wave generator.

The three governing equations for a shock wave under the above conditions are called the Rankine-Hugoniot relations. These equations, which are based on the conservation of mass, momentum, and energy, show the relationship between the energy, pressure, and specific volumes (reciprocal of the density) as well as the shock velocity and the partial velocity (the velocity at which the shock material is moving after the shock has passed). When the shock velocity and the partial velocity are measured, the volume behaviour of a substance, which is a function of pressure, can be computed.

The Rankine-Hugoniot equations also apply to projectile impact when it is more convenient to impel a projectile at a high velocity and allow it to impact a sample than it is to use explosives. Whereas explosive research is limited to pressures of about two megabars, high-velocity studies with a gas gun may reach several megabars. Workers in the Soviet Union have reported experimentation at 34 megabars, but they have not specified the technique they employed to reach such pressures.

BIBLIOGRAPHY. P.W. BRIDGMAN, *The Physics of High Pressure* (1949), the major source of information prior to 1949; R.S. BRADLEY (ed.), *High Pressure Physics and Chemistry*, vol. 1– (1963–), a series of review articles started in 1963; *Advances in High Pressure Research*, vol. 1– (1966–), a review series started in 1966; N.B. BRANDT and N.I. GINZBURG, "Superconductivity at High Pressure," *Scient. Am.*, 224:83–88 (1971); F.P. BUNDY, W.R. HIBBARD, and H.M. STRONG (eds.), *Progress in Very High Pressure Research* (1961); W. PAUL and D.M. WARSCHAUER (eds.), *Solids Under Pressure* (1963); K.A. GSCHNEIDNER, M.T. HEPWORTH, and N.A.D. PARLEE (eds.), *Metallurgy at High Pressures and High*

Temperatures (1964); C.T. TOMIZUKA and R.M. EMRICK (eds.), *Physics of Solids at High Pressures* (1965); R.H. WENTORF (ed.), *Modern Very High Pressure Techniques* (1962); A. ZEITLIN, "High Pressure Technology," *Scient. Am.*, 212:38–46 (1965).

(J.C.J.)

Hilbert, David

The German mathematician David Hilbert was a towering figure in world mathematics for an entire generation. He fully grasped the major problems that since his time have occupied the attention of 20th-century research in many branches of mathematics.

Hilbert was born on January 23, 1862, at Königsberg, Germany (now Kaliningrad, Russian Soviet Federated Socialist Republic). Partly because he was able to earn his living by the mathematical work he chose to do—as was seldom the case for his predecessors—his life was reasonably placid. The first steps of his career occurred at the University of Königsberg, at which, in 1884, he finished his *Inaugurel-dissertation* (Ph.D.) and remained at Königsberg as a *Privatdozent* (lecturer, or assistant professor) in 1886–92, as an *Extraordinarius* (associate professor) in 1892–93, and as an *Ordinarius* in 1893–95. In 1892 he married Käthe Jerosch, and they had one child, Franz. In 1895 Hilbert accepted a professorship in mathematics at the University of Göttingen, at which he remained for the rest of his life.

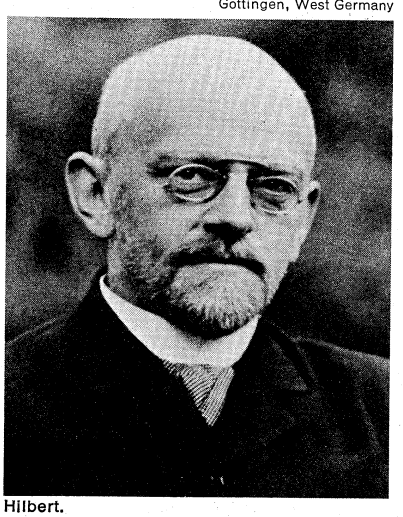
Association with the University of Göttingen

The University of Göttingen had a flourishing tradition in mathematics, primarily as the result of the contributions of Carl Friedrich Gauss, Peter Gustav Lejeune Dirichlet, and Bernhard Riemann in the 19th century. During the first three decades of the 20th century this mathematical tradition achieved even greater eminence, largely because of Hilbert. The Mathematical Institute at Göttingen drew students and visitors from all over the world.

Hilbert's intense interest in mathematical physics also contributed to the university's reputation in physics. His colleague and friend, the mathematician Hermann Minkowski, aided in the new application of mathematics to physics until his untimely death in 1909. Three winners of the Nobel Prize for Physics—Max von Laue in 1914, James Franck in 1925, Werner Heisenberg in 1932—spent significant parts of their careers at the University of Göttingen during Hilbert's lifetime.

In a highly original way, Hilbert extensively modified the mathematics of invariants—the entities that are not altered during such geometric changes as rotation, dilation, and reflection. Hilbert proved the theorem of invariants—that all invariants can be expressed in terms of a finite number. In his *Zahlbericht*, a report on algebraic number theory published in 1897, he consolidated what was known in this subject and pointed the way to the developments that followed. In 1899 he published the *Grundlagen der Geometrie* (*The Foundations of Geometry*, 1902), which contained his definitive set of axioms for Euclidean geometry and a keen analysis of their significance (see GEOMETRY, EUCLIDEAN). This popular book, which appeared in 10 editions, marked a turning point in the axiomatic treatment of geometry.

By courtesy of the University Library, Gottingen, West Germany



Hilbert.

Hilbert's 23 problems

A substantial part of Hilbert's fame rests on a list of 23 research problems he enunciated in 1900 at the International Mathematical Congress in Paris. In his address, "The Problems of Mathematics," he surveyed nearly all the mathematics of his day and endeavoured to set forth the problems he thought would be significant for mathematicians in the 20th century. Many of the problems have since been solved, and each solution was a noted event. Of those that remain, however, one, in part, requires a solution to the Riemann hypothesis, which is usually considered to be the most important unsolved problem in mathematics (see NUMBER THEORY).

In 1905 the first award of the Wolfgang Bolyai prize of the Hungarian Academy of Sciences went to Henri Poincaré, but it was accompanied by a special citation for Hilbert.

In 1905 (and again from 1918) Hilbert attempted to lay a firm foundation for mathematics by proving consistency—that is, that finite steps of reasoning in logic could not lead to a contradiction. But in 1931 the Austrian–U.S. mathematician Kurt Gödel showed this goal to be unattainable: propositions may be formulated that are undecidable; thus, it cannot be known with certainty that mathematical axioms do not lead to contradictions. Nevertheless, the development of logic after Hilbert was different, for he established the formalistic foundations of mathematics (see MATHEMATICS, FOUNDATIONS OF).

Hilbert's work in integral equations about 1909 led directly to 20th-century research in functional analysis (the branch of mathematics in which functions are studied collectively). His work also established the basis for his work on infinite-dimensional space, later called Hilbert space, a concept that is useful in mathematical analysis and quantum mechanics. Making use of his results on integral equations, Hilbert contributed to the development of mathematical physics by his important memoirs on kinetic gas theory and the theory of radiations. In 1909 he proved the conjecture in number theory that for any n, all positive integers are sums of a certain fixed number of nth powers; for example, $5 = 2^2 + 1^2$, in which $n = 2$. In 1910 the second Bolyai award went to Hilbert alone and, appropriately, Poincaré wrote the glowing tribute.

Hilbert space

The city of Königsberg in 1930, the year of his retirement from the University of Göttingen, made Hilbert an honorary citizen. For this occasion he prepared an address entitled "Naturerkennen und Logik" ("The Understanding of Nature and Logic"). The last six words of Hilbert's address sum up his enthusiasm for mathematics and the devoted life he spent raising it to a new level: "Wir müssen wissen, wir werden wissen" ("We must know, we shall know"). In 1939 the first Mittag-Leffler prize of the Swedish Academy went jointly to Hilbert and the French mathematician Émile Picard.

The last decade of Hilbert's life was darkened by the tragedy brought to himself and to so many of his students and colleagues by the Nazi regime. He died at Göttingen on February 14, 1943.

BIBLIOGRAPHY. Hilbert's collected works, *Gesammelte Abhandlungen*, 3 vol. (1932–35, reprinted 1965; 2nd ed., 1971), contain almost all of Hilbert's papers, including the *Zahlbericht;* there are also assessments of his work by other mathematicians. Four of his six books have been translated into English, but the *Zahlbericht* is available only in German and French. CONSTANCE REID, *Hilbert* (1970), is a full-length biography. OTTO BLUMENTHAL, Hilbert's first Ph.D. student and lifelong colleague, wrote two biographical sketches; the first appeared in the 1922 *Naturwissenschaften* and the second at the end of the collected works. HERMANN WEYL's obituary notice in the *Bull. Am. Math. Soc.*, 50:612–654 (1944), is a definitive assessment by Hilbert's leading student. KURT REIDEMEISTER (ed.), *Gedenkband* (1971), contains some previously unpublished papers of Hilbert and also the recording of his 1930 speech.

(I.K.)

Hillel

In his time—the last half of the 1st century BC and approximately the first quarter of the 1st century AD—the Jewish sage Hillel was the foremost master of biblical commentary and interpreter of Jewish tradition.

Hillel was born in Babylonia, where he received both his early and secondary education. As a young man he went to Palestine in order to continue advanced studies under the leading teachers of Scripture and the Oral Law who belonged to the group or party called Pharisees. Although a strictly biographical account of Hillel's life cannot be set forth, for virtually every narrative about him is encrusted with legend, the literary sources do combine coherently to summon up what may be called the first distinct personality of Talmudic Judaism, the branch of Jewish thought and tradition that created the Talmud, a commentative work on the Oral Law. Put another way, it can be said that the life of Hillel is more than a vague recollection of anecdotes or a name with a saying or two attached.

The first distinct Talmudic personality

More than one story underscores his whole-hearted devotion to study; for example, on one occasion when he lacked the necessary funds for admission to the academy, he is said to have climbed to the roof of the study house, and so attentive was he to the discussions he heard through the "skylight" that he once failed to notice that snow had begun to fall and almost froze to death. It is impossible to separate fact from fancy in such narratives, but such stories are illustrative of the character Hillel projected.

As with most of the Talmudic sages, no miracles or supernatural performances are ascribed to Hillel, but he is represented as a person of exemplary, even superlative virtues. He is, in the traditional accounts, the model of patience, and even when repeated attempts are made by some to insult him, his equanimity and civility remain unaffected. He appears as a fervent advocate of peaceful conduct, a lover of all men, a diligent student, a persuasive and ready teacher, a man of thorough and cheerful trust in God, and the most considerate of hosts. In this last quality he influenced his wife as hostess. In short, he appears in all accounts as the model of the ideal Jewish sage.

This idealization is not entirely storyteller's praises. Critical analysis of Hillel's sayings, of his two legal enactments to relieve economic hardships in society, and even of the motifs the legends seek to emphasize leave little doubt that Hillel did indeed affect the texture of Jewish life profoundly.

While he is nowhere described as the originator of rules to guide the student in the legitimate interpretation of Holy Scriptures, Hillel is unquestionably one of the most influential Talmudic sponsors and practitioners of a conscious, carefully applied exegetical discipline necessary for the proper explanation of the contents of the Bible. The "Seven Rules" he employed—some of which are reminiscent of rules prevailing in Hellenistic schools where Homer was studied and interpreted—were to serve as the basis for more elaborate rules in the 2nd century. Homilies or parables ascribed to Hillel reveal him as a superb pedagogue. For example, there is a story of a heathen who, having been dismissed by Shammai, a sage who was Hillel's chief opponent, came to Hillel and asked to be converted so that he might be appointed high priest. Instead of saying to him that his request was preposterous, Hillel said, "If one wishes to greet a king of flesh and blood, is it not right that he learn how to make his entrances and exits?" The heathen agreed. Hillel continued: "You wish to greet the King over kings of kings, the Holy One, blessed be He. Is it not all the more right that you learn how to enter into the Holy of Holies, how to fix the lights, how to approach the altar, how to set the table, how to prepare the row of wood?" To this too the heathen agreed and thereupon began to study. The heathen's biblical studies finally disclosed to him "of his own accord" why his request could not have been granted.

Mastery of parable and epigram

Along with his other gifts, Hillel had an epigrammatic felicity that is apparent in his sayings, in Aramaic no less than in Hebrew ("A name made great is a name destroyed"; "If I am not in my own behalf, who will be in my behalf? And if I am for myself only, what am I? And if not in the present, when?"), and that inevitably contributed to their being long remembered and very likely taught to schoolchildren as examples of good style. Significantly, in the unique treatise of the Mishna (the authoritative collection of Oral Law), *Pirqe Avot* ("Chapters of the Fathers"), Hillel is quoted more than any other Talmudic sage. As head of a school known as the House of Hillel, he succeeded in winning wide acceptance for his approach, which liberated texts and law from slavishly literal and strict interpretation; indeed, without him an uncompromising rigidity and severity might have developed in the inherited traditions.

Hillel's appreciation of the socioeconomic needs of his age, and of the large possibilities that are inherent in both biblical statements and values, plus his preference for persuasiveness to get across his point of view led to the adoption, with few exceptions, of the Hillelite view of Talmudic teaching and to its establishment as the legal norm.

Talmudic sources speak of Hillel's promotion to patriarchal leadership after he had proved his intellectual superiority to the incumbents then in office. In any event, the Jewish patriarchs—the Roman term for the official leaders of the Palestinian Jews—down to about the 5th century, when the patriarchate came to an end, were descendants of Hillel.

Many of the stories about Hillel, especially those in which he is contrasted with his countercolleague Shammai, are among the most popular Talmudic tales in Jewish literature and folklore.

BIBLIOGRAPHY. In addition to the summary description in the general Jewish histories, see W. BACHER, *Die Agada der Tannaiten*, vol. 1, pp. 4–24 (1890); A. BUCHLER, *Types of Jewish-Palestinian Piety* (1922); L. FINKELSTEIN, *Ha-Perushim ve-Anshe Keneset Ha-Gedolah*, pp. 1–16 (1950), English summary, pp. vi–viii; L. GINZBERG, *On Jewish Law and Lore*, pp. 77–124 (1955); N.N. GLATZER, *Hillel the Elder: The Emergence of Classical Judaism* (1956); and J. GOLDIN, "Hillel the Elder," *Journal of Religion*, 26:263–277 (1946). The nature of the material on Hillel is such as to make impossible a solid reconstruction of his life along the lines of scholarship. The better studies on him, listed in this bibliography, are ultimately speculative. The most useful presentation, therefore, remains the chapter in *Die Agada der Tannaiten* (cited above), for it assembles conveniently all that the sources have on Hillel, and Bacher also provides useful notes for the reader.

(J.Gol.)

Hillslopes

Few landscapes are completely flat. Most plains contain isolated hills rising from the main land surface, and most plateaus have been dissected by streams to produce valley-side slopes. The form of hillslopes, using the term to cover the sloping surface of both hills and valleys, imposes a distinctive mark on the scenery of a landscape. The chalk downs of southern England are, for instance, renowned for their gently rolling profiles; the limestone massifs of Europe for their spectacular gorges; and the semi-arid Colorado Plateau area for its steep-sided scarps and hills that rise abruptly from the desert floor. It is not surprising that scientists have long been fascinated by hillslopes and their change in form through geological time.

Concepts and historical background

The history of concepts of hillslope development is a colourful one. A particularly prominent figure in the early days of earth science was Maj. J.W. Powell (1834–1902), a veteran of the American Civil War and pioneer explorer of the Grand Canyon of the Colorado River. Powell was one of the first to accept the idea that the landscape is molded not only by internal earth movements but also by erosion associated with the action of rain and streams. He recognized that steep, narrow valleys cut in upland massifs, and low-lying plains broken only by occasional small hills, are often merely extreme stages in a gradual process of earth sculpture in which

highlands become plains. Many of his ideas were expanded by another American geomorphologist, W.M. Davis (1850–1934), who set forth the concept of the erosion cycle.

According to Davis, the landscape, and its component parts such as hillslopes, undergo a systematic evolution through geologic time. Landmasses are uplifted from beneath the sea, and once uplift is halted, a gradual process of destruction through erosion takes place. The early stages of the cycle presumably are characterized by steep hillside slopes undergoing rapid undercutting by streams. As the cycle progresses, erosion flattens hillslopes, and the gentler slopes result in slower rates of erosion. Eventually very gentle slopes are formed, and these constitute the supposed ultimate planar surface, a peneplain. Davis was clearly influenced by the evolutionary ideas of Charles Darwin a few years earlier and, indeed, in outlining the cycle he compared the various stages of landscape dissection to organic life cycles and gave his stages the anthropomorphic designations "youth," "maturity," and "old age."

Although Davis attracted many disciples in North America, his ideas were strongly challenged in Europe, particularly by Walther Penck (1888–1923), a German geologist. Penck argued that insufficient attention had been paid to the basic premise underlying the whole concept. He maintained that the notion of rapid uplift of the earth's surface followed by stable conditions, during which erosion destroyed the landmass, was a great oversimplification of actual continental events and histories. It completely ignored the possibility of concurrent uplift and erosion and the emergence of time-independent, equilibrium landforms associated with this. Notwithstanding the theoretical significance of Penck's views, evidence accumulated during this century tends to support Davis. Modern rates of mountain upheaval, averaging 7.5 metres per thousand years (24.6 feet per thousand years), are about ten times greater than maximum rates of erosion found in drainage basins today. Uplift, when it occurs, takes place at rates much greater than erosion rates, whereas in stable areas, erosion appears to be the dominant process at work.

It should be emphasized that Penck did recognize that many continental areas of the earth's surface have been tectonically stable for long periods of time, and that, in these areas, a cycle of erosion not unlike that suggested by Davis would occur. For many years this was not realized, and it was thought that Penck's scheme of landscape development during stable conditions was radically different from the Davisian cycle. According to the popular view of the Penckian scheme, hillslopes retreat parallel to themselves, leaving behind a very gently sloping surface at the foot of the slope. The real founders of the school of parallel slope retreat, however, were Kirk Bryan, an American geomorphologist, and Lester King, a South African geomorphologist, who worked in semi-arid areas in the United States and southern Africa, respectively.

Modern focus on processes

Much of the early work in hillslope studies was speculative. It was often based on visual impression rather than field survey; little attention was directed to the mechanics of the processes held responsible for hillslope erosion; and many ideas, usually expressed in words rather than mathematically, were in fact illogical. More recently, however, hillslopes have attracted the attention of engineers concerned with erosional problems from a practical viewpoint. Landslides pose problems in hilly urban areas; clearing of forests on valley slopes has increased flood frequency; and exposure of the soil surface on hillside slopes has resulted in washing away of valuable topsoil and rapid silting of expensive reservoirs. Detailed research by engineers tackling these problems has provided considerable information that may also help to solve the more academic problem of how hillslopes develop through time.

This article treats the basic aspects of hillslope form, processes, and development. (For further information on concepts of landscape development in general, see LANDFORM EVOLUTION; for more detail on the relevant processes involved, see EARTH MOVEMENTS ON SLOPES; FLUVIAL PROCESSES; and SEDIMENT YIELD OF DRAINAGE SYSTEMS.)

HILLSLOPE FORM

Description of slope profiles. One of the basic problems in hillslope studies is to define quantitatively the form or geometry of hillslopes. If slopes were simply tilted planar surfaces of bare rock or soil-mantled rock, this would be fairly easy; two parameters, slope angle and slope length, would be sufficient to describe the slope form. Even if hillside slopes were not planar, but curved like part of the surface of a cone, it would be relatively simple to add an extra parameter to describe the contour curvature, provided that the slope profiles were straight in section.

Few hillslopes are, however, entirely straight. Many, in fact, are predominantly curved, either convex (steepening from hillcrest to slope base) or concave. In such cases, other parameters are required to describe slope profiles quantitatively. At this point it is simplest to assume that contour curvature (the aspect of hillslope form as viewed from above) is zero; later this will be built into slope description. Slopes that are entirely curved in profile may be described in terms of vertical and horizontal distances by a general curve of the form:

$$H = cL^f, \text{ or}$$
$$\log H = \log c + f \log L \qquad (1a)$$

in which H is the vertical fall from the hilltop, L is the horizontal distance from the hilltop, and c and f are constants for a particular hillslope. The constant c is a measure of average steepness, and f is a measure of the amount of curvature—that is, the rate of change of steepness along the slope profile. Whereas a straight slope is defined by the two parameters of slope angle and length, three parameters are needed for curved hillslope profiles. Studies of hillslope profiles in the Appalachian Mountains of the eastern United States have shown that convex hillslope profiles are properly expressed by curves of this type. Typical values of c and f vary between 0.001 and one, and one to two, respectively. These values vary for slopes on different rock types. For concave profiles f would be less than unity but greater than zero. This type of power function is often called a logarithmic curve; if a graphic plot of H against L were made with logarithmic values along the axes, a straight line would result. An alternative form of curved profile, in which the symbols have the same meaning, is the exponential relation:

$$H = f \log L + C. \qquad (1b)$$

If values of L were plotted on a logarithmic axis and values of H on an arithmetic axis, a straight line would result. Such curves have been applied to the longitudinal profiles of rivers, which are often regularly concave. In areas where most of the lowering of hillslope profiles is the result of erosion by water flowing downslope, stream profiles could be regarded as merely an extreme form of hillslope profile.

Few hillslope profiles, however, conform to a single, regular curve. Most profiles combine one or more convex, concave, and straight parts, and this inevitably makes the study of slope profiles rather more difficult. Three distinct responses to this problem can be detected in the work of the last two or three decades. Some authorities have maintained that hillslope profiles are so complex that studies must use very detailed procedures for delineating, describing, and classifying the various hillslope elements. A second approach has ignored the complex geometry of profiles and attempted merely to pick out salient geometric properties such as maximum angle of slope, average slope steepness, and slope length. Third, some workers, while recognizing that the detailed form of slope profiles is complex, have argued that the basic form is fairly simple, and, without any concern for microscale features, have attempted to describe and explain the large-scale geometric properties of profiles.

Much work on the detailed form of hillslopes has been

Slope segments and elements

done by investigators in England. Hillslope profiles are usually surveyed in the field by measuring the steepness and lengths of successive sections of slope along the line of steepest gradient (that is, perpendicular to contours); the problem is then to separate the profile into its various components. A lengthy nomenclature has been provided. The term segment is used for straight sections, and element for curved parts of slopes and slope profiles. Unfortunately, such classifications sometimes pose more problems than they solve. Terms such as approximately rectilinear and approximately smoothly curved must be defined quantitatively, unless they are to be interpreted in different ways by different workers. Moreover, it is doubtful if any single method of subdividing slope profiles would prove suitable for all of the areas; techniques that are satisfactory for small slopes may be meaningless for longer slope profiles. It is questionable whether very detailed methods of slope-profile description are more useful than generalized delineation of slopes into straight, convex, and concave sections. The ultimate test of any classification is the extent to which it enables improved comprehension of the system. Notwithstanding the painstaking thoroughness of this morphometric approach, the amount of light thrown on slope-profile development has been rather limited.

Major characteristics of slope profiles. The second approach, limited to quantitative description of a few basic parameters of slope profiles, has yielded a number of rather interesting points. A.N. Strahler, an American geomorphologist, examined the relationship between the maximum steepness of hillside slopes and the steepness of the stream gradient at the slope base for a large number of hillslopes. A strong correlation was found between the two variables: hillsides with steep maximum-slope sections were undercut by streams with steep gradients, whereas gentler hillsides were associated with gentler streams. Subsequently, other workers have detected additional correlations between maximum hillside slope and such factors as rock type, infiltration capacity of the soil mantle, vegetation cover on the hillside, and related attributes. Maximum hillside slope, thus, appears to be a useful property of slope profiles inasmuch as it seems to reflect processes operating on and at the base of hillslopes. Early hillslope studies were also directed to average hillslope steepness, but this is a less useful parameter because it is essentially dictated by the drainage pattern. It is easily shown that average hillside slope can be expressed as:

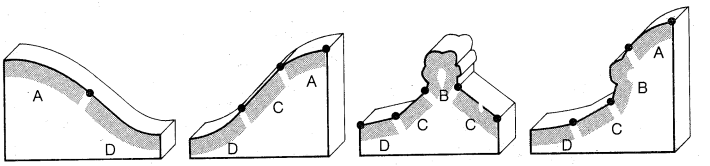
$$\bar{\theta} = 2HD_d \qquad (2)$$

in which $\bar{\theta}$ is average hillslope angle; H is the vertical height of the slope; and D_d is the drainage density (ratio of total stream channel length in a drainage basin to the drainage area) of the area. Studies restricted to basic properties such as length and average and maximum steepness are, however, inevitably rather limited in their approach to slope form, even though they have produced certain interesting data.

The four-section profile

The most successful of the three approaches has been that directed at describing and explaining the generalized form of hillslope profiles. Descriptions of profiles by many workers in different geologic and climatic environments have shown that certain types of profile occur repeatedly in the landscape. It has been noted, for example, that profiles of hillslopes cut in highly fractured rocks, such as granites, shales, and some limestones, typically show a four-section form in a wide variety of climates. This standard profile is convex at the top, exhibits a steep rock cliff below, and then, in succession, a straight slope mantled by debris from the cliff above, which merges into a concave basal slope. Although not all hillslopes contain all of the four components, many profiles (Figure 1) are essentially variants on the basic form. Along with this recognition of the overall form of hillslope profiles, recent work has offered explanations for the basic features of these profiles in terms of the various erosion processes occurring on the slopes.

Before turning to the topic of hillslope processes, mention must be made of the plan form (as viewed from above) of hillslopes. Clearly, the major aspects of the plan geometry of hills are determined by the pattern of the drainage network. Parallel streams produce similar ridges, and tributary streams that cut into those ridges will produce alternate hollows and noses or spurs. Rounded hills in plan are simply the extreme case of a nose. These features are easily explained in terms of the drainage pattern and have important implications for slope profiles. Profiles in hollows and on noses may be expected to differ from those on true valley-side slopes or hillslope ridges. In hollows, for instance, water flow converges as it moves downslope, and, therefore, discharge per unit of surface area will tend to increase downslope; on noses, in contrast, water flow diverges, and discharge will increase more slowly in moving downslope from the divide.



A Upper convex hilltop
B Outcrop of bare rock as cliff
C Straight, debris-mantled slope
D Lower concave basal slope

Figure 1: Variants on the basic hillslope profile.

Similar statements may be made about the downslope movement of soil on slopes. Observations suggest that the hillcrest convexity is more strongly developed on noses, and the concave basal slope is more prominent in hollows than for comparable profiles where hills are straight in plan view. The long profiles of streams are dominantly concave and, in a sense, stream profiles can thus be considered slope profiles in extremely pronounced hollows.

HILLSLOPE PROCESSES

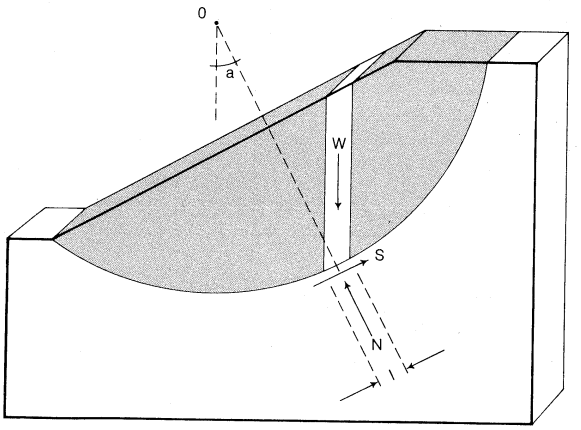
The various processes that act to move material down, and eventually off, hillside slopes may be conveniently grouped into two categories; namely, fluvial processes and mass movements. The former processes occur on all slopes irrespective of the steepness of the hillside. The latter, with the exception of certain very slow mass movements, are most frequent on slopes that are being undercut rapidly; in the absence of undercutting, they occur only on slopes that are greater than a critical angle of stability. This limiting or threshold slope angle depends on the type of material forming the slope. Slopes subject to slow undercutting are, thus, molded primarily by fluvial processes and slow mass movements; slopes subject to rapid undercutting are lowered essentially by instability processes—that is, by rapid mass movement.

Shear force and strength of materials

Mass movements. There are many types of instability processes, but all of them obey the same physical laws. Any point in a mass beneath a sloping surface supports a vertical force due to the weight above. The component of this force acting along any potential slip surface is called the shear force; this increases with depth and with slope angle. The component acting perpendicular to the surface, and, hence, at right angles to the shear force, is called the normal force. These relations are shown in Figure 2. At any depth beneath the surface the shear force is attempting to move the material above that depth downslope, and eventually, to reduce the slope to a level surface. Whether or not movement of the material actually occurs depends on the strength of the material to withstand the shear force. Water, for instance, has no strength and, if it is subjected only to the force of gravity, will always move to produce a level surface. Slopes cut in soil or rock, in contrast, do have strength to withstand shear forces; if they did not, the landscape would be flat. Nonetheless, there are limits to the vertical height of canyon walls and to the steepness of debris slopes; these

limits depend on the amount of shear strength of the material in which the hillslopes are formed.

The strength of earth materials is usually separated into two components. One is frictional resistance between



W weight of slice
S shear resistance along potential slip surface
N force normal to potential slip surface
l thickness of hypothetical slice
a slope angle

Figure 2: Section through a circular landslip (see text).

mineral particles. This includes not only true friction between flat surfaces but also interlocking of particles. It is termed internal friction and is designated by either the angle, *phi* (ϕ), or the coefficient, mu (μ), of internal friction. The total frictional force along a potential failure surface, at limiting equilibrium, can be expressed as:

$$F = N\mu = N \tan \phi \qquad (3a)$$

in which F is the frictional force; N is the normal force that is pushing the particles together; and μ, or the tangent of the angle ϕ, represents the internal friction. Sometimes friction is expressed in terms of a stress—that is, the frictional force per unit area. In addition to frictional strength, particles may be cemented together or drawn together by physicochemical forces, providing another source of strength termed cohesion. This is the major source of strength in intact rocks, but it is lost when the rock mass is broken down into individual fragments. Cohesion is usually regarded as independent of the normal stress; total shear strength is, thus, conventionally expressed as cohesion plus the stress acting on the internal rock resistance, namely:

$$s = c + \sigma \tan \phi \qquad (3b)$$

in which s and c are, respectively, shear strength and cohesion expressed as stresses, *sigma* (σ) also is a stress, and ϕ is again internal friction.

Instability and critical height

Instability processes may be classified in many ways. The following approach is unconventional but very useful in understanding hillslope forms; it is based on the rate of undercutting of hillslopes relative to the rate at which the rock breaks down under the action of weathering (*q.v.*). At high rates (rapid undercutting or resistant rock), little of the cohesive bonding in the rock mass is destroyed and the rock may be considered intact. At medium rates, weathering destroys the cohesive strength of the rock mass near the surface of the valley slopes, but not at depth. At slower rates (slow undercutting or easily weathered rock), cohesion is lost throughout and the shear strength of the hillside material is entirely frictional.

Streams cutting down vertically into a solid, intact, rock mass will produce stable vertical slopes until a critical valley depth is reached. It can be shown that this depth is related to the physical properties of the rock and the stresses that are operative upon it, namely:

$$H = \frac{kc}{\gamma} \tan (45 + \phi/2) = \frac{kq_u}{2\gamma} \qquad (4)$$

in which H is the critical height of slope (valley depth); *gamma* (γ) is the unit weight (density) of the rock; k is a constant approximately equal to four, varying slightly with the shape of the potential failure surface; and c and ϕ are cohesive and frictional components of rock strength, respectively. The parameter q_u is the unconfined compressive strength of the rock (see ROCKS, PHYSICAL PROPERTIES OF). The critical height of a rock cliff with $q_u = 500$ kilograms per square centimetre (7,100 pounds per square inch) and $\gamma = 2{,}500$ kilograms per cubic metre (150 pounds per cubic foot), typical values for a medium-strength sandstone, would be 4,000 metres (13,000 feet). The critical height of a vertical rock slope is rarely attained in nature. The reason is that the process of valley cutting itself produces extensive weakening of the rock mass. If the rock mass is thought of as a set of vertical columns of rock packed together tightly, then this claim can be appreciated. Each column exerts a lateral thrust on adjacent columns; if a column is removed, there is a sudden drop in the lateral pressure on adjacent columns. This pressure release, or stress relaxation, results in a slight expansion of the rock mass toward the gap. The effect is a weakening of the rock and eventual destruction of cohesion in the vicinity of the gap. Analogous processes occur on the sidewalls of valleys as streams cut down creating canyons in rock masses.

The destruction of the cohesive strength in a rock mass by pressure release may take many forms, depending on the type of rock. Massive rock formations (rocks with few fractures) tend to produce vertical tension cracks at some distance from the canyon wall. This reduces the critical height of the wall by an amount equal to the depth of the tension crack. Instability thus takes the form of slabs of rock peeling off the valley wall. In highly fractured rocks, pressure release weakening takes a different form. As the lateral stress is relaxed, some cracks begin to open; the shear strength along these cracks is accordingly reduced and an extra part of the shear force must be taken up by other cracks. This increase in shear force opens up other cracks, and a chain reaction occurs until the slope material becomes completely cohesionless. Substitution of $c = 0$ in Equation 4 (above) yields zero, indicating that vertical slopes cannot exist in cohesionless material. If cohesion is lost throughout the rock mass, rock avalanches will lower the valley slope to an angle θ, at which the slope is just stable. It can be shown that θ should be identical to ϕ. Highly fractured material in a dense state of packing has ϕ values in the range 45° to 75°; clay usually has ϕ values in the range 5° to 25°, and as a result, slopes cut in clay tend to be gentler.

If a rock mass is sufficiently strong, or if undercutting is sufficiently rapid, cohesion may be destroyed only in the near-surface parts of the rock mass, and slope angles will be steeper than ϕ. The actual value will be restricted by the depth of the valley, as indicated by the expression relating slope angle, critical height, and rock properties; namely:

$$\frac{\sin \theta \cos \phi}{1 - \cos (\theta - \phi)} = \frac{H\gamma}{kc} \qquad (5)$$

in which θ is the angle at which the slope is just stable, ϕ represents internal friction, H is slope height, γ is rock density, k is a constant, and c represents rock cohesion. Examination of this equation shows that θ increases as the left-hand side decreases and, therefore, as H decreases and c increases. Note that Equation 4 is simply a special case of Equation 5, solved for H rather than θ, with $\theta = 90°$. As an example, consider a stream downcutting into a clay mass characterized by $\gamma = 2{,}500$ kilograms per cubic metre (150 pounds per cubic foot), $\phi = 20°$, and $c = 1$ kilogram per square centimetre (14 pounds per square inch), with valley-side slopes at 45°. The critical height would be 104 metres (341 feet).

Once the valley depth exceeded this value, instability would occur. On steep slopes, instability takes the form of a wedge sliding on a planar, failure surface. On gentler slopes, such as in clay masses, the failure surface

is usually curved, and instability takes the form of a rotational slip. This is illustrated in Figure 2; at limiting equilibrium, the shear force and strength along the failure arc are just equal.

Threshold slope angles

Under slower conditions of undercutting, widespread destruction of cohesion allows slope angles to become independent of slope height. The actual angle depends on the extent of weathering because this affects not only the cohesive strength but also the frictional strength of the rock material. With little weathering, so that in effect the rock mass is a densely packed, cohesionless aggregate, the angle $\phi = 45°–75°$, as noted previously. Further weathering, particularly loosening of the surface material by frost action, would produce a less densely packed aggregate. Such loose rock debris is often called talus or scree; the reduction of interlocking strength during weathering into scree results in a decrease in ϕ to about 32°–38°. The ϕ value of talus is often referred to as the angle of repose. Further weathering would break down rock fragments into earth. Deposits of earth and talus, termed taluvium, have ϕ values between 40° and 45°; fully weathered material may have values of ϕ between 4° and 35°, depending on the mineralogical composition and grain size of the soil.

The limiting angle of slope (threshold slope) for material containing fine earth debris, however, is no longer determined simply by ϕ. The angles are equal only for materials in which the voids, or pores, between particles are occupied by air. When water fills these voids, the pore pressures are greater than atmospheric pressure. This pore pressure tends to reduce the interparticle contact force imposed by the normal force and, therefore, it also reduces the amount of friction developed. In regoliths containing appreciable earthy material, the small voids allow saturation at times of prolonged rainstorms; and when the pore pressure is greater than zero, the threshold slope is less than the angle ϕ. Pore pressure increases with depth below the water table, but the exact relationship depends on the pattern of groundwater flow. Several studies have shown that the threshold slope of hillslopes mantled by weathered debris may often be given by the equation:

$$\tan \theta = \tfrac{1}{2} \tan \phi \qquad (6)$$

where θ is the threshold slope angle. Taluvium would, therefore, be expected to be stable only at angles near to 25° and colluvial mantles at angles from 2° to 20°, depending on the type of soil. This last expression may not hold for mantles that weather into clay, however; cohesion may develop in regoliths of fine earth and the angle of limiting stability in this case would depend on slope height.

Soil creep

Under very slow conditions of undercutting, instability processes are infrequent, but slower mass movements, commonly called soil creep, may, under suitable conditions, become very important. There are at least two mechanisms of soil creep. One is a rheological effect, long studied by engineers, and sometimes referred to as shear creep. It occurs at depth as well as in the near-surface zone, but only on slopes greater than a critical steepness, which varies with the soil type. The other, called seasonal soil creep, is confined to surface layers (usually within the upper 50 centimetres [or 20 inches]) but does not depend upon a minimum slope steepness. Seasonal soil creep is probably more universal than shear creep, and more is known about it.

This slow downslope movement of the surface layers is primarily the result of alternate expansion and contraction of the soil. It may stem from either wetting-drying or freeze-thaw of the surface soil. Expansion of the soil mass is along paths perpendicular to the ground slope, but during contraction the soil moves back in a direction between these expansion directions and the vertical. The effect of an expansion-contraction cycle, therefore, is to move material slightly downslope. The rate of soil creep on a hillside slope depends on slope angle, the amount of expansion-contraction per cycle, and the frequency of cycles. Creep is important in humid areas, where alternate wetting and drying of the soil mantle is fairly frequent; field measurements indicate rates of one to ten cubic centimetres per centimetre per year (the volume moving past a line one centimetre wide, along the contour, in a year) in these environments. Creep rates appear to be even higher in Arctic and Alpine areas, where the magnitude of the annual freeze-thaw cycle is at a maximum; in these areas creep probably merges imperceptibly into solifluction (earth flowage). Creep is probably at a minimum in tropical deserts, where neither of the two mechanisms of expansion-contraction can be very effective.

Evidence and theory suggest that, on straight hillside slopes, creep rate is independent of distance downslope from the divide. This has important implications for slope-profile development. If, during a unit of time, the volume of soil moving past a point downslope is equal to the volume moving down from an upslope point, clearly there is no net removal of soil, and, therefore, no lowering of the hillside along that stretch of slope. All of the net loss of material must occur at the hillcrest; as a result, crests are rounded convexly. Continued creep will tend to extend the convexity downslope.

Fluvial processes. Three main types of fluvial processes of erosion act on hillslopes: rain-splash, surface-soil wash, and solution. In terms of the effect on slope-profile geometry, rain-splash should be grouped with soil creep rather than with other fluvial processes. The term rain-splash is applied to the splashing of soil particles off the ground surface by the impact of falling raindrops. On hillside slopes this tends to produce a net movement of material downslope because particles are splashed further downslope than upslope from any point. Experiments have shown that the movement of soil downslope due to rain-splash increases with slope angle. Whether or not rain-splash attains any real importance, however, depends on the effectiveness of the vegetation cover in sheltering the ground surface from raindrop impact. Field data show that bare soil surfaces may erode at rates that are many times greater than rates on grass-covered plots. Much of the accelerated soil erosion on watersheds cleared for farming can be attributed to rain-splash. On natural slopes, with unaltered vegetation cover, rain-splash erosion is much greater on the bare surfaces of semi-arid areas than in humid areas; this is accentuated by the high intensity of rainstorms in the former areas. Data suggest that volumetric rates of rain-splash erosion in semi-arid areas may approach creep rates in humid areas, although generally they are probably lower. The layer of moving debris is much thinner and the velocity much greater in the case of rain-splash. As in the case of soil creep, the downslope movement of material that is due to rain-splash appears to be independent of distance downslope on straight slopes except at the hillcrest. Net loss, therefore, is concentrated near the summit of the slope, producing, and extending downslope, a convex slope element.

Rain-splash erosion

Soil wash and overland flow

Surface-soil wash, the transport of soil debris downslope by water flowing over the ground surface, may take many forms. At one extreme there are isolated rivulets or rills and, at the other extreme, a continuous sheet of water. Early workers maintained that rill erosion and unconcentrated wash had radically different effects on slope-profile geometry, but there is little evidence to support this view. The hydraulics of overland flow in all of its various forms on hillslopes seem to be similar to open channel flow modified by the thinness of the moving fluid. Although the detailed mechanics of debris transport by moving water, in both channels and on slopes, are poorly understood, certain relationships are well established. First, the tractive force acting to move debris along, or suspended above, the bed of a flowing fluid increases with the depth of flow and the slope of the water surface. Second, the threshold tractive force necessary to initiate movement along the bed increases as the particle size increases. For a given particle size there is, therefore, a critical combination of depth and slope that must be attained before movement will occur. If it is assumed that there is a definite limit to the depth of flow that will occur at any point, it is evident that, at that

point, there is a threshold gradient below which no movement will occur. Neither channel slope nor hillside slope can be flattened to angles below that corresponding to the threshold traction condition. This threshold gradient is not the same at all of the points along a hillside slope. Average particle size decreases downslope owing to various processes, including abrasive action during movement, and as a result, the threshold tractive force necessary for debris transport is smaller at increasing distances from the divide. Moreover, because the water discharge of overland flow, and therefore depth, increases with distance from divide, there is a still more pronounced decrease in threshold gradient downslope.

The ultimate slope profile developed by surface wash processes should, therefore, be concave upward. But this threshold slope profile is not necessarily static; at any point on the profile weathering may reduce particle size over time. This fact means that the equilibrium concave slope will depend on the extent of weathering of the surface.

The importance of surface-soil wash depends on two factors: the frequency of occurrence and depth of overland flow, and the resistance offered by soil particles to movement by the flow. In very dry areas, soil wash is small because overland flow is rare; in very wet areas, soil wash is also relatively minor because of the dense vegetal cover, which not only intercepts much of the rainfall but also protects the ground surface against the force of overland flow. Data on the sediment yield of drainage systems (*q.v.*) indicate that maximum erosion rates occur in semi-arid areas with an annual effective precipitation of about 40 centimetres (16 inches). Direct measurement of surface-soil wash in these areas shows rates of 200–500 cubic centimetres per centimetre per year (30–79 cubic inches per inch per year) and, sometimes, even higher (1,000 cubic centimetres per centimetre per year has been measured in the Sangre de Cristo Mountains, New Mexico), in contrast to rates less than 0.3 cubic centimetres per centimetre per year (0.05 cubic inches per inch per year) on grass-covered slopes in humid areas.

Subsurface solution

One of the main reasons why surface-soil wash is so ineffectual in humid areas is that the combination of lower rainfall intensities and denser vegetal cover on hillslopes increases infiltration of rainwater into the soil mantle and allows little overland flow. In these areas, water moves to stream channels primarily through subsurface routes; it is to be expected, therefore, that subsurface erosion is more important than surface-soil wash. Some of this erosion is direct mechanical movement of fine material through pores in the soil. A much more important process, however, is transport of elements in solution. This affects particularly calcium, sodium, magnesium, and potassium cations (positively charged ions). Direct evidence of the importance of solution in humid areas is afforded by studies of solute and sediment concentration in stream water; except for infrequent storms, the dissolved load is usually appreciably higher than the sediment load. Some of this dissolved load originates from the chemical weathering of bedrock at depth beneath the ground surface and has no direct effect on slope profiles. Solution is also responsible, however, for movement of material out of the soil mantle; this is indicated by the development of distinct soil profiles (vertical layering or zoning) on flatland. On hillside slopes, subsurface movement of water is downslope rather than vertical, particularly above impermeable layers, so that soil profiles are poorly developed. Studies show that much of this subsurface seepage occurs just above the interface between the soil mantle and the underlying bedrock, and it is probable that much of the solution process is concentrated here. In this way the interface is directly lowered, and, indirectly, the ground surface is also lowered, although by a smaller amount depending on the fraction of residual material undissolved. This residual fraction is very low in limestones, which typically produce very shallow soils. In other rocks, much of the solution process is taken up by thickening of the soil mantle rather than lowering of the ground surface. The thickness of weathered material on tropical hillsides testifies to the high rate of solution under hot, moist conditions. Solution is possibly the major erosion process operative on stable slopes in humid areas. Little is known about the effect solution has on slope profiles. Uniform lowering of the soil-rock interface would produce parallel retreat of hillside slopes. Precipitation of salts on the downslope parts of hillsides, the result of the evaporation of water as it seeps downslope, would, however, result in a gradual decrease in the net loss of material with increasing distance from the divide, and hence, to slope decline.

HILLSLOPE DEVELOPMENT

The equilibrium slope. Each process or combination of erosional processes acting on hillsides will produce an equilibrium slope form that is lowered at the same rate as undercutting of the hillside, assuming that undercutting continues at a constant rate for a sufficiently long period of time. Under conditions of rapid undercutting, instability processes control the equilibrium slope form; hillsides will be straight and inclined at angles of limiting stability. Under very slow conditions of undercutting, slower mass movements and fluvial processes determine slope geometry; equilibrium slopes are rounded, being dominantly convex where creep or rain-splash prevails, and dominantly concave where surface-soil wash is most important.

It is unlikely, however, that undercutting rates will remain constant over long periods of geologic time; as a result, it must be expected that slope forms also will change through time as one equilibrium form is replaced by another. According to W.M. Davis, as discussed at the outset of this article, undercutting of slopes by streams decreases through geologic time, and slope forms evolve accordingly. The Davisian system is not entirely satisfactory because, as emphasized by Walther Penck and others, it focusses attention on only one of an infinite number of possible courses of stream behaviour and landscape evolution. There is no single course of hillslope evolution even for a given rock type in a given climatic setting; the actual sequence depends upon changes in the rate at which slopes are undercut over time. Any general framework of hillslope evolution can present only the equilibrium slope forms for different rates of undercutting for different rock type-climatic environments. These forms may then be arranged in any permutation to accommodate different courses of stream erosional behaviour.

Classification of equilibrium slope forms

A convenient classification of equilibrium slope forms, on the basis of the rate of undercutting (in relation to the rate of weakening of the rock mass), follows directly from the previous discussion of erosional processes on hillsides. Four general classes may be distinguished in order of decreasing rates of undercutting:

1. Slopes produced by downcutting in fully cohesive rock masses will be vertical provided that valley depth is less than the critical value given by Equation 4; subject to this proviso, these vertical walls may be regarded as equilibrium–or time-independent forms. Such slopes do exist in nature, but only in very strong rock masses. Moreover, such massive formations are limited in thickness and probably always are thinner than the critical depth for the canyon walls that can be cut in them. These equilibrium forms, therefore, are short-lived in terms of geological time; as soon as streams cut through these massive formations into weaker rock underneath, nonvertical equilibrium slopes, discussed below, will be formed. Vertical cliffs, thus, tend to form parts of valley slope profiles in deep valleys, but they rarely form the entire rock wall. Many hills in the Colorado Plateau of the southwestern United States, for instance, show vertical slopes of sandstone that cap gentler slopes in weaker shales.

2. Slopes produced by downcutting at slower rates, which allow partial weathering of the rock mass and destruction of cohesive bonds in the near-surface zones, are nonvertical, as noted previously. Actual slope angles depend upon the extent of penetration of the cohesion-destroying forces and also on the valley depth, as indi-

cated by Equation 5. In areas where valleys are widely spaced and separated by broad, flat divides, the concept of equilibrium slope forms is inapplicable because valley depth increases with time, and slope steepness, therefore, is also a time-dependent phenomenon. Valley depths cannot increase indefinitely, however, because as they get deeper, the divides between them become narrower; eventually, divides become narrow ridges formed by the intersection of the valley slopes themselves. In this situation divides are lowered at the same rate as slopes are undercut; valley depth and slope now become time-invariant features. Because the initial valley spacing influences the depths to which valleys may develop before equilibrium is attained, the drainage density has a direct influence on slope form as indicated previously by Equation 2.

3. Slopes cut sufficiently slowly to allow complete loss of cohesion in the rock, in the vicinity of valleys, stand at an angle dictated by the internal friction angle ϕ, itself dependent on the stage of rock weakening and the characteristic pore-pressure pattern. Theoretically, straight equilibrium slopes may occur at any value between 90° and 0°, depending on these two parameters. Certain ϕ values are characteristic of different materials, and certain pore-pressure patterns are probably more common than others. It is, therefore, not surprising that some slope angles occur repeatedly in nature, whereas others are rare. Some common equilibrium slope angles reported by various authorities are as follows: clay, 10° to 15°; sandy soil, 20° to 21°; earthy rubble (taluvium), 25° to 27°; scree, 32° to 38°; and densely packed, fractured rock, 43° to 48°. The steepness of these straight threshold slopes depends not only on the type of rock mass but also on the rate of undercutting of the rock mass in relation to the rate of weakening of it.

4. Under conditions of downcutting, which are slower than those just necessary to produce fully weathered threshold slopes, instability processes become more infrequent and other processes—slow mass movements and fluvial processes—begin to mold the equilibrium slope forms. Soil creep and rainsplash produce convex hilltops, and equilibrium profiles may be either convex–straight or entirely convex, depending on the rate of the process relative to the rate of undercutting. Surface-soil wash produces concave forms beginning at the slope base and extending upslope. In areas where both soil wash and creep are significant and undercutting is slow, the typical slope form will be convex–concave. The effect of solution on the equilibrium slope form is still not fully understood and is probably dependent on the environment.

Influence of rock type and climate. The influence of rock type on the equilibrium slope form is immediately evident from this classification. The rate at which the rock mass weakens affects the class of slope form that is developed. Within any single class it may also affect the equilibrium profile. Highly fractured rock masses, for instance, may produce at least four different threshold slopes depending on the stage of rock disintegration, whereas a rock that weathers directly into a fine earth debris is restricted to much fewer possibilities. Because of its influence on the type of residual mantle that is produced by weathering, the rock type may indirectly influence the balance of processes that yield convex and concave forms. Impermeable clays encourage overland flow and surface-soil wash, whereas more permeable soil allows infiltration of rainfall and produces creep; this was neatly demonstrated in a study of slope forms in the Badlands National Monument, South Dakota.

Climate, runoff, and vegetation

Climate exerts a major influence on the form of equilibrium slopes in two main ways. First, it may allow fluvial processes and slow mass movements to assume a significant role well before weathering of the hillside material has been completed. In semi-arid areas, *in situ* debris mantles derived from the weathering of highly fractured rock frequently do not pass beyond the scree stage. Soil-size particles tend to be washed downslope by overland flow as soon as they are produced, and the hillside material remains as scree. More fully weathered threshold slopes, therefore, should not be as common as in humid temperate areas where surface-soil wash is a minor process. This may also be true for some parts of Arctic and Alpine areas that, in terms of vegetation, may be regarded as semi-arid. Second, climate exerts a strong control on the relative importance of slow mass movements and the various fluvial processes, and, thus, on the form of equilibrium slopes developed under conditions of very slow downcutting. In semi-arid areas the dominance of surface-soil wash is associated with very long, shallow concave profiles. Where this is developed on *in situ* rock material it is termed a pediment. Many pediment-like slope forms, however, are produced in semi-arid areas by extensive deposition of colluvium in valley bottoms, after it has been washed off the main hillsides. In humid temperate areas, surface wash and rainsplash are minor processes, but soil creep is relatively important. Slope profiles produced under conditions of slow undercutting are, therefore, largely, or sometimes wholly, convex. Concavities do occur at the slope base on many profiles, but whether they are caused by contemporary solution, surface wash in Pleistocene times (from 10,000 to 2,500,000 years ago), or some other process, is uncertain. Cold areas and hot, moist areas appear to be intermediate between semi-arid and humid temperate conditions in terms of the relative importance of convex- and concave-producing processes, but very little research has been undertaken in these areas so far.

From M.A. Carson, "Models of Hillslope Development under Mass Failure," *Geographical Analysis*, vol. 1, no. 1, p. 92; © 1969 by the Ohio State University Press, all rights reserved

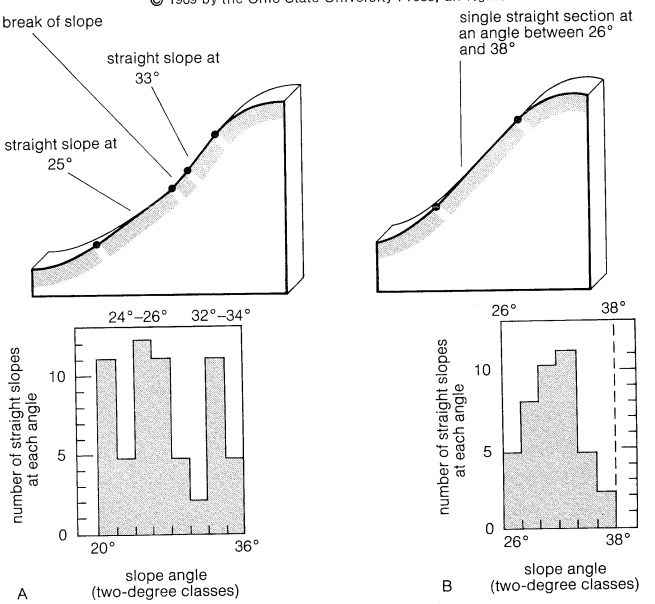


Figure 3: Differences between hillslopes attributable to (A) parallel slope retreat and (B) slope decline. Both hillslopes occur in jointed rocks in humid temperate climate, but the (top) actual hillslope profiles and (bottom) corresponding frequency histograms contrast (see text).

Slope retreat. In the Davisian system of landscape evolution, hillside slopes supposedly change through time, from equilibrium forms associated with high rates of undercutting to forms associated with slower rates. The main problem involved in the construction of such an evolutionary sequence, once the various equilibrium forms are known (at the time Davis proposed the cycle of erosion, very little was known about these forms), is to determine how, geometrically, the change from one equilibrium profile to another is achieved. There are two possible ways in which steeper hillslopes may be replaced by gentler slopes: hillsides may retreat from the initial position of the slope base and produce a new, gentler slope at its foot until continued retreat eventually results in complete replacement of the steeper slope by the gentler one; or steep slopes may gradually flatten to gentler slopes as though they were hinged at the slope base. Both views have their adherents, and the argument

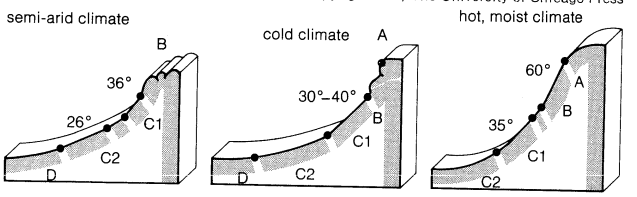
between the parallel-retreat school and the slope-decline school was the central issue in slope studies.

There is little doubt that both modes of change may occur—indeed, possibly even on the same hillside. This may be illustrated by some work on highly fractured rocks (shales, shaly gritstones, quartzitic gritstones, and limestones) in central and southwestern England, as shown in Figure 3. Most of the hillslopes in these two areas are mantled by scree or taluvium; a few are cloaked by a colluvial mantle. Equilibrium slope forms are dominated by straight sections at 32°–37° (scree slopes), where undercutting is rapid, and by straight sections at 25°–26° (taluvial slopes), where undercutting is slower. On slopes cut in shales and shale-grit, the change from one angle to another appears to be the result of the retreat of the scree slope and upslope extension of the taluvial replacement. This is indicated by two observations. Many slope profiles are multisegmented, with upper scree slopes at 32° and lower taluvial slopes at 25°—26°; upper slopes are very unstable and landslides appear to strip material off of these slopes resulting in retreat. This is supported by the histogram (Figure 3) of straight-slope angles, which shows peaks at 32° and 25° and a gap between these peaks. If slopes were to decline gradually from 32° to 25°, straight slopes would be expected at all angles between the two limits. On slopes cut in quartzitic gritstone and limestone, in contrast, the change from a scree-mantled slope to a taluvium-mantled slope appears to be caused by a gradual hinge decline; slopes occur at all angles between 37° and 26°, and no double-segment slope profiles occur.

Multi-segment slope profiles

Two lines of evidence do suggest, however, that retreat of slopes is a common mode of adjustment from one equilibrium form to another. The first is that multisegment slope profiles are very frequent even in relatively homogeneous rock material. The second is that similar angles of straight slopes occur repeatedly in different areas, whereas others are relatively infrequent. Figure 4 shows typical slope profiles on highly fractured rock

From (centre) based on photo from A. Jahn, *Zeitschrift fur Geomorphologie, Suppl. Bd. 1, Morphologie des Versants* (1960), Gebruder Borntraeger Verlagsbuchhandlung; and A. Rapp, *Geographiska Annaler* (1960), Generalstabens Litografiska Anstalt; (right) C.K. Wentworth, *Journal of Geology* (copyright 1928) The University of Chicago Press



A Convex, soil-mantled upper slope
B Fairly straight rock cliff
C1 Straight, dry, talus slope near to angle of repose of debris
C2 Moist taluvial slope; sometimes straight, sometimes concave merging with valley floor
D Very gentle concave slope mantled by very fine colluvial material

Figure 4: Slope profiles on jointed rock (see text).

in three climate settings: semi-arid, cold, and hot, wet conditions. All of them show distinct segments, as they steepen from slope base to hillcrest, hinting that the steeper segments are in the process of retreat, producing gentler slopes beneath them. In addition, all of the three profiles contain talus slopes at 30°–40° and taluvial-mantled slopes. Slope retreat does appear to be a common mode of adjustment from one equilibrium form to another. Moreover, judging by the abundance of multisegment slope profiles, this adjustment is rather a slow one. In an evolutionary system, then, it seems probable that many, perhaps most, of the slopes are not equilibrium forms but are in transition states from one equilibrium condition to another. In view of this, it is not surprising that slope studies have produced many conflicting ideas.

This framework of discussion of slope form and slope development is inevitably a much more simplified picture than actually exists in nature. Most of it relates to a uniform rock mass, whereas many natural slopes are carved in materials of mixed lithology. This may result in complex slope forms. Moreover, it has been assumed that constant climatic conditions prevail, and this, in the long term, is a purely hypothetical situation; climatic zones have fluctuated in position over the earth, especially during the last two to three million years of geologic time. Slope forms in humid temperate areas may have been affected by colder conditions; slope forms in semi-arid areas have probably been modified by former moist conditions. Indeed, some workers believe that it is impossible to explain present-day slope forms in terms of present-day processes, for this reason. Finally, any model is only as good as the "tools" with which it is built. The tools involved here are the concepts of fluid mechanics, soil mechanics, rock mechanics, hydrology, geochemistry, and related disciplines. Advances in these subjects will inevitably lead to refinements in current models of slope development.

BIBLIOGRAPHY. The three classic works by W.M. DAVIS, *Geographical Essays* (1909); W. PENCK, *Die morphologischen Analyse* (1924; Eng. trans., *Morphological Analysis of Landforms*, 1953); and L.C. KING, *The Morphology of the Earth* (1962), all discuss the development of hillslopes. The articles by A.N. STRAHLER, "Equilibrium Theory of Erosional Slopes Approached by Frequency Distribution Analysis, *Am. J. Sci.*, 248:673–696, 800–814 (1950); and S.A. SCHUMM, "The Role of Creep and Rainwash on the Retreat of Badland Slopes," *Am. J. Sci.*, 254:693–706 (1956), are two of the first attempts to measure slope form and process quantitatively. M.A. CARSON and M.J. KIRKBY, *Hillslope Form and Process* (1972); A. YOUNG, *Slopes* (1972); and S.A. SCHUMM and M.P. MOSLEY, *Slope Morphology* (1973) are devoted to this topic.

(M.A.C.)

Himachal Pradesh

Occupying a region of scenic splendour in the western Himalayas, Himachal Pradesh is a state of the Indian Union at the extreme north of the Indian subcontinent. It is bounded on the north by Jammu and Kashmir, on the east by China (Tibet), on the southeast by Uttar Pradesh, on the south by Haryana, and on the west by Punjab. Towering snow-clad mountains are divided by deep valleys with thick woods, green fields, and cascading streams, as well as lakes and waterfalls. Himachal means Snowy Mountain (*hima*, "snow"; *acal*, "mountain"), the state (*pradesh*) taking its name from the Himalayas.

The physical setting

The area of the state is 21,490 square miles (55,658 square kilometres). Its population in 1971 was more than 3,400,000. The capital is Simla, at an elevation of 7,262 feet above sea level, the largest and most popular hill resort in India and the summer headquarters of pre-independence British viceroys.

Formerly a union territory, Himachal Pradesh emerged as a full state of the Indian Union on January 26, 1971.

History. The history of the area dates back to the Vedic period, the earliest known tribal group being Dāsas, who had a reputation in the arts of warfare. Later the Aryans assimilated the tribes. The area was exposed to successive invasions through the centuries, ending with the British dominion over the native princes of the hill states.

Himachal Pradesh was constituted as an administrative unit comprising 30 hill states in 1948. This event was preceded by a movement for the end of feudalism and the introduction of responsible government. One of the princely states, Suket, virtually surrendered to the peaceful demonstrators, hastening the process of change.

Achievement of statehood

Between 1948 and the achievement of statehood in 1971, Himachal Pradesh went through various changes in size and administrative setup. First a substate and then a union territory directly administered by the central government, Himachal Pradesh was enlarged in 1954 by the merger of Bilāspur (a former Indian state and then a chief commissioner's province), and again in 1966 by the merger of Punjab hill areas including Simla, Kāngra, Kulu, and Lahaul and Spiti. Y.S. Parmar, who since the 1940s had led the hill people of Himachal Pradesh in the quest for responsible government, became the first chief minister of the new state.

Physical geography. The terrain varies with hills of

low and high altitudes, wooded valleys, flowing rivers, and rocky soils. The mountains rise to about 22,000 feet and include the Himalayan ranges of the Pīr Panjāl, Hathi Dhār, and Dhaula Dhār. The mountain valleys are known locally as *dūn*s (*e.g.*, the Kiarda Dūn in the Sirmūr district); the hill stations of Kulu, Manāli, and Dharmsāla are in the Kulu and Kāngra valleys. The major rivers are the Chenāb (Chandra Bhāga), Rāvi, and Beās in the west and the Sutlej and Yamuna (Jumna) in the east.

The climate ranges from mild to cold. Icy winds sweep the state, and the hills and valleys remain covered with snow for months. In the valleys flowers and fruits grow in abundance, and their rivers abound in fish. The forests harbour wildlife, including the ibex, goral, bear, and snow leopard.

Population. The population of the state in 1971 was approximately 3,400,000, and its population density was about 159 per square mile. The population presents a mixture of hill tribes, including Gaddis, Gūjars, Kinners, Lahaulis, and Pangwalas. The vast majority of the people are Hindus, who worship in temples dotted throughout the state, but there are also Buddhists, Sikhs, Muslims, and Christians. About 60 dialects are spoken, the main language being Western Pahari. Pahari is derived from Sanskrit and Prākrit. Caste distinctions and outdated social customs are slowly fading under the impact of influences from outside, the spread of education, and the growing consciousness of changes taking place throughout the country.

Himachal Pradesh is the least urbanized state in India, the urban population accounting for just 7 percent of the total. There are few towns, of which only the capital, Simla, has a population exceeding 50,000. The population of Sundarnagar is 21,000. The populations of Mandi, Nāhan, Chamba, Solon, and Dharmsāla are between 10,000 and 17,000; those of Bilāspur, Kāngra, Dalhousie, Una, and Kulu are between 5,000 and 9,000; Palampur and Kasauli range down to less than 4,000. Over 90 percent of the people depend for their livelihood on farming, horticulture, and livestock raising.

Administration. The head of the state is the governor, appointed by the president of India. The council of ministers, headed by a chief minister, is responsible to the 68-member legislative assembly that is elected directly on the basis of adult suffrage. The state has four representatives in the lower house (Lok Sabha) of India's Parliament and three in the upper house (Rajya Sabha). The state is divided into ten districts—Chamba, Mandi, Bilāspur, Mahāsu, Sirmūr, Kinnaur, Simla, Kāngra, Kulu, and Lahaul and Spiti.

Social conditions. In the second half of the 20th century, there has been expansion of education and public health facilities and an improvement of communications. The people, however, remain at the subsistence level, and the state's vast natural resources have yet to be tapped systematically. Development plans lay emphasis on road building, power generation, and exploitation of forest products for industrialization.

Growth of educational institutions

In the years since 1948, education has made rapid strides. The state now has 18 degree-granting colleges connected with Punjab University as compared with one in 1950. There is a medical college at Simla and an agricultural college at Solon, and research is carried on at the Indian Institute of Advanced Study (Simla) and the Central Research Institute (Kasauli). School enrollment has gone up in the wake of a phenomenal rise in the number of primary, middle, and higher schools. In addition to some 500 hospital-dispensaries and over 70 primary health centres, there are more than 70 family-planning centres engaged in the campaign to check the birth rate. More than 3,000 out of 13,000 inhabited villages had been electrified by 1971, and a substantial beginning had been made in providing drinking water to the people in the remote areas.

Economy. Himachal Pradesh, with its perennially snow-fed rivers, has a hydroelectric power potential of 8,500,000,000 watts. There is immense scope for the development of forests and setting up of forest-based industries, including newsprint. Hydroelectric development has only begun with installations in Mandi district on the Uhl River, a tributary of the Beās, and in Bilāspur district at the Bhākra Dam across the Sutlej.

Economy based on the land

The state's economy is still almost entirely based on the products of the land. The chief crops are wheat, maize (corn), barley, rice, and potatoes. Excellent varieties of plums, peaches, apricots, pomegranates, and other fruits are produced. The state's high-quality apples are exported to neighbouring states and abroad. Sheep and goat rearing are widespread, and wool and pashm (the fleece of the Tibetan goat) are important exports. The coniferous forests have also been a source of revenue.

There is very little industry. A foundry was established at Nāhan in 1872; taken over by the Indian government in 1952, it manufactures agricultural implements. There are also a turpentine factory at Nāhan and a brewery at Solon. Some woolen goods are made on handlooms and small machines, and several small plants process foods, notably fruits.

Concentrated efforts have been made to improve the state's economy through a series of development plans. The outlook for the economy is regarded as hopeful in view of the prospects of exploitation of the abundant power potential, mineral deposits, and forests and the promotion of tourism.

Transportation and communication. Except for the rail line from Kālka (in Haryana) to Simla, the capital, and the narrow-gauge track connecting Pathānkot (Punjab) and Jogindarnagar (H.P.) through the Kāngra valley, there are no railways or waterways in the state. Roads are the communications lifeline of Himachal Pradesh, and the state spends 30 percent of the development plan outlay to expand the road system. There are now over 4,500 miles of roads compared with less than 300 miles in 1950. The nationalized transport system operates more than 140 routes in the state.

Cultural life. The fairs and festivals of the hill people are occasions of joyful song and dance. Besides folk dances, there are folk songs with romantic themes and folktales woven around mythological figures. Both men and women partake in the dance and music and in the folk dramas.

Kāngra is a famous centre of paintings of delicate beauty with love as their central theme. In addition to this 18th-century school of painting, Himachal can boast of beautiful and useful crafts, such as the exquisitely designed shawls of Kinnaur and the embroidered handkerchiefs of Chamba.

The Simla hills, Kulu and Manāli valleys, and Dalhousie are tourist attractions. Skiing, golfing, fishing, trekking, and mountaineering are activities for which Himachal Pradesh is ideally suited.

Pilgrims journey to the Kulu and Kāngra valleys to worship at shrines of legendary antiquity. From Kāngra, one can proceed to Dharmsāla, a pleasant hill station that is the home of the Dalai Lama, who fled from Tibet in 1959 in the wake of Chinese occupation of Lhasa. The Kulu Valley is known as the Valley of the Gods; its pine and deodar forests, flower-spangled meadows, and fruit orchards provide the setting for the colourful Dussehra fair held every autumn. The temple gods are taken in caparisoned palanquins from the hilltops, and through singing and dancing the hill people give vent to their joy.

BIBLIOGRAPHY. Y.S. PARMAR, *Himachal Pradesh: Its Proper Shape and Status* (1965) and *Himachal Pradesh: Case for Statehood* (1968), present the historical perspective of the movement for a separate state of Himachal Pradesh, by the first chief minister of the state. R.K. KAUSHAL, *Himachal Pradesh* (1965), is a historical study that deals with developments leading to the formation of the state. H.C. SARASWAT, *Himachal Pradesh* (1970), provides a synoptical account of the state, its people, social customs, and economy. The *Census of India*, Paper 1 of 1971 (1971), gives the population and other vital statistics. The *Himachal Pradesh: Fourth Five Year Plan* (1969), embodies the programs formulated by the state government for development in agriculture, power and irrigation, transport and communications, social services, and other fields.

(C.Ra.)

Himalayan Mountain Ranges

The Himalayas, of Asia, include the highest mountains in the world, with more than 30 peaks rising to heights of 24,000 feet (7,300 metres) above sea level. One of these peaks is Mt. Everest, the world's highest, which reaches a height of 29,028 feet. The great heights of the mountains rise above the line of perpetual snow. The vast permanent snowfields attracted the attention of the pilgrim mountaineers of ancient India, who coined the Sanskrit name Himalaya—from *hima*, "snow," and *ālaya*, "abode"—for this great mountain system. In modern times, the Himalayas have constituted the greatest attraction and the greatest challenge to mountaineers throughout the world.

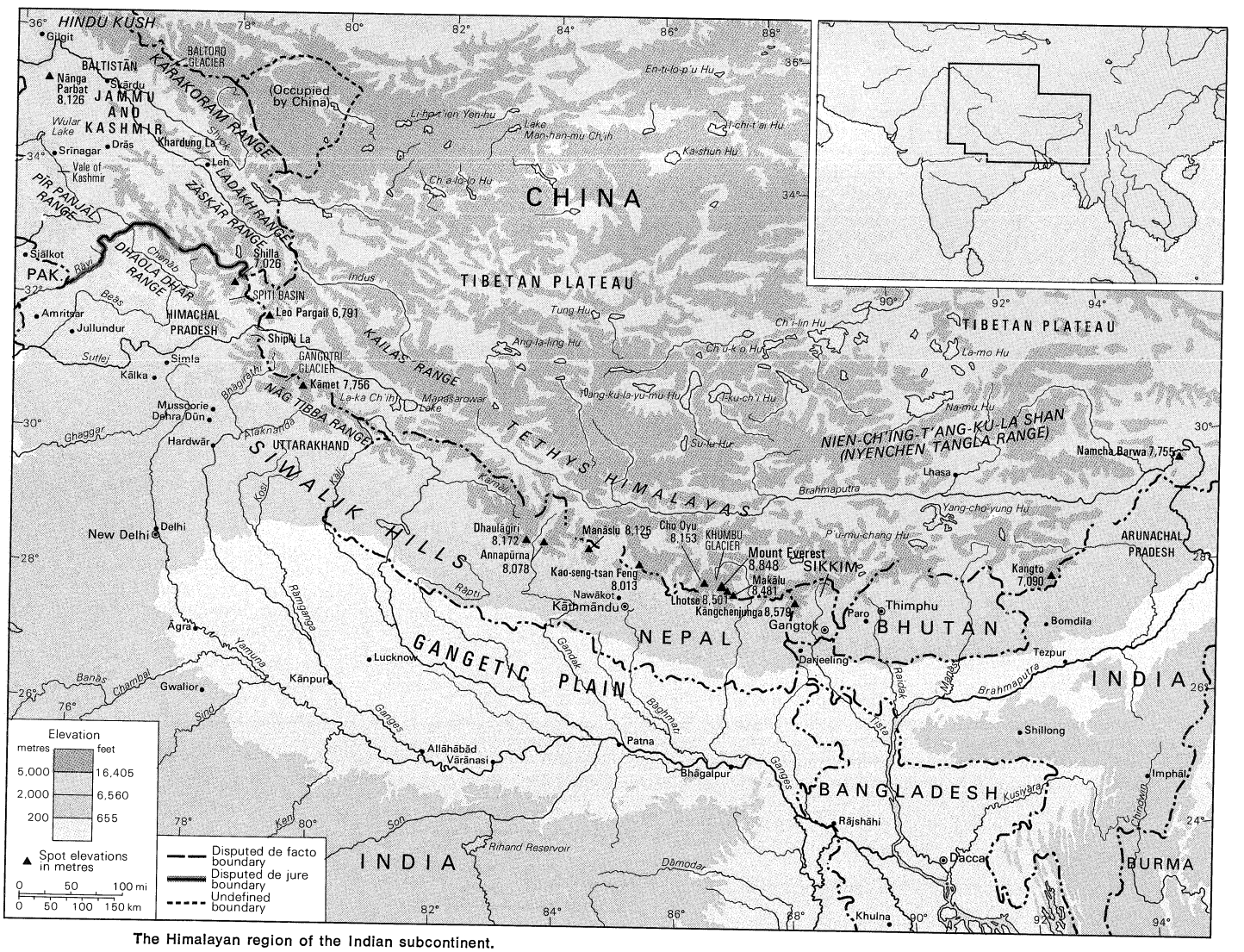
Forming the northern border of the Indian subcontinent and an almost impassable barrier between it and the lands to the north, the ranges form part of the great mountain belt stretching halfway around the world from northern Africa to the east coast of Asia. The Himalayas themselves stretch uninterruptedly for about 1,550 miles (2,500 kilometres) from west to east between Nānga Parbat (26,660 feet), in the disputed state of Jammu and Kashmir, and Namcha Barwa (25,445 feet), in Tibet. Between these eastern and western extremities lie the three Himalayan kingdoms of Nepal, Sikkim, and Bhutan. The Himalayas are bordered to the northwest by the mountain ranges of the Hindu Kush (*q.v.*) and Karakoram (*q.v.*) and to the north by the high Plateau of Tibet. The width of the Himalayas from south to north varies between 125 and 250 miles. Their total area amounts to about 229,500 square miles (594,400 square kilometres).

Though India has sovereignty over most of the Himalayas, Pakistan and China occupy parts of them. In the state of Jammu and Kashmir, Pakistan has administrative control of 32,362 square miles of the range lying north and west of a cease-fire line established between India and Pakistan in 1948. China's occupation of 14,000 square miles in the Ladākh district of Kashmir, as well as Chinese incursions to the south of the McMahon Line (a 1914 boundary line limiting Tibetan sovereignty in the Assam Himalayas of northeast India) in the North East Frontier Agency (now Arunachal Pradesh) in 1962, have accentuated further the boundary problems faced by India in the Himalayan region. (For an associated physical feature, see EVEREST, MOUNT.)

Physical geography. The Himalayas' most characteristic features are their soaring heights, snowcapped and steep-sided jagged peaks, valley glaciers often of stupendous size, topography deeply cut by erosion, seemingly unfathomable river gorges, complex geological structure, and a rich temperate and alpine vegetation. Viewed from the south, the Himalayas appear as a gigantic crescent, with its main axis rising above the snow line, where snowfields feed the valley glaciers and constitute the sources of most of the Himalayan rivers. The greater part of the Himalayas, however, lies below the snow line. The mountain-building process that created the range is still active and is accompanied by erosion by rivers and landslides of great dimension.

From south to north, the Himalayan ranges can be grouped into four parallel, longitudinal mountain belts of varying width, each having distinct physiographic features and its own geological history. They are designated

Principal divisions



The Himalayan region of the Indian subcontinent.

as the Outer, or Sub-Himalayas; the Lesser, or Lower Himalayas; the Great, or Higher, Himalayas; and the Tethys, or Tibetan Himalayas. Farther north lie the Trans-Himalayas in Tibet proper, eastward continuations of some of the most northerly Himalayan ranges. From west to east the Himalayas are divided broadly into three mountainous regions: Western, Central, and Eastern.

Physiography. The Outer Himalayas comprise flat-floored structural valleys and the Siwalik Hills, which border the Himalayan mountain system to the south. Except for small gaps in the east, the Siwalik run for the entire length of the Himalayas with a maximum width of 62 miles in the Indian state of Himachal Pradesh. In general, the 900-foot contour line marks their southern boundary; they rise to another 2,500 feet to the north. The main Siwalik range has steeper southern slopes facing the Indian plains and descends gently northward to flat-floored basins, called *dūns*. The best known of these is the Dehra Dūn, in Uttarakhand, which is in the mountainous parts of Uttar Pradesh.

Northward, the Siwalik Range abuts against a 50-mile-wide massive mountainous tract, the Lesser Himalayas, where mountains rising to 15,000 feet and valleys with altitudes of 3,000 feet run in different directions. There is a general conformity of altitude among neighbouring summits, which creates the appearance of a highly dissected plateau. The three principal ranges of the Lesser Himalayas, the Nāg Tibba, the Dhaola Dhār, and the Pīr Panjāl, have branched off from the Great Himalayan Range lying farther north. The Nāg Tibba, the most easterly of the three ranges, is 26,795 feet high near its eastern end, in Nepal, and forms the watershed between the Ganges (*q.v.*) and the Yamuna, in the Uttarakhand.

To the west, the picturesque Vale of Kashmir, a structural basin (*i.e.*, an elliptical basin in which the rock strata are inclined toward a central point), forms an important section of the Lesser Himalayas. It extends from southeast to northwest for 100 miles, with an average elevation of 5,100 feet, having a width of 50 miles; it is traversed by the meandering Jhelum River, which runs through the Wular Lake, the largest freshwater lake in India.

The backbone of the Himalayan system is formed by the Great Himalayas, a single high range rising above the line of perpetual snow.

Highest peaks of Himalayas

The Great Himalayan Range rises to its maximum height in Nepal, having in that section nine of the 14 highest peaks of the world. From west to east they are: Dhaulāgiri (26,810 feet), Annapūrna 1 (26,504 feet), Manāslu (26,760 feet), Kao-seng-tsan Feng (Gosainthan; 26,291 feet), Cho Oyu (26,750 feet), Mt. Everest (29,028 feet), Makālu (27,824 feet), Lhotse (27,923 feet), and Kāngchenjunga (28,208 feet).

Farther east the range changes from a southeasterly to an easterly direction as it enters Sikkim. After this, it runs eastward for another 260 miles through Bhutan and the eastern part of Arunachal Pradesh as far as the peak of Kangto (23,260 feet) and finally turns northeast, terminating in Namcha Barwa.

There is no sharp boundary between the Great Himalayas and the ranges, plateaus, and basins lying to the north of the Great Himalayan Range, generally grouped together under the name of the Tethys Himalayas and extending far northward into Tibet. In Kashmir the Tethys are at their widest, forming the Spiti Basin and the Zāskār Mountains, the highest peaks of which, to the southeast, are Leo Pargail (22,280 feet), rising north of the Sutlej River opposite Shipki La (pass), and Shilla (23,050 feet).

Geology. A study of the geological history of the Himalayas reveals that marine sediments of Paleozoic and Mesozoic eras (between about 65,000,000 and 570,000,000 years ago) deposited on the floor of the ancient Tethys Sea; the frontal part of the crystalline massif (mountain mass) of peninsular India; estuarine deposits along the flanks of the embryonic mountains; and finally products of surface erosion of the rising mountains—all contributed to the formation of the present-day range. The uplift of the Himalayas took place in at least three distinct and widely separated phases. The first phase of the major mountain-building movement took place at the close of the Eocene Epoch (about 38,000,000 years ago), although the beginning of the main Himalayan uplift started in Middle and Upper Cretaceous times (from about 100,000,000 to 65,000,000 years ago), with the advancement of the crystalline massif of peninsular India toward the Plateau of Tibet. This movement caused the rise of the Tethys Himalayas, along with the greater part of the Great Himalayas. In the second phase of upheaval, which occurred in the Miocene Epoch (7,000,000 to 26,000,000 years ago), the estuarine deposits and the Indian Massif formed the ranges of the Lesser Himalayas. The final mountain-building phase started at the end of the Tertiary Period (about 7,000,000 years ago), lifting the detrital deposits accumulating at the base of the Himalayas to form the Siwalik Range, the foothills of the Outer Himalayas. Since the middle Pleistocene Epoch (about 1,500,000 years ago), the Himalayas have risen at least 4,500 feet, an occurrence witnessed by early man. That the Himalayas still continue to rise is evidenced by the upheaval of the younger river terraces.

Formation of the ranges

Precambrian metamorphic rocks (rocks formed by heat and pressure from 570,000,000 to 4,600,000,000 years ago) form the bulk of the Himalayas. These rocks represent the frontal part of the Indian Shield, which, according to the theory of continental drift (see CONTINENTAL DRIFT), pushed northward, uplifting the Himalayas as it pressed against the Asian landmass. Only in the Spiti Basin and in a few other localities can large outcrops of marine sediments from Paleozoic and Mesozoic times be seen.

Plutonic rocks (formed deep down from a molten state), such as granites and granodiorites of pre-Miocene age (*i.e.*, more than 26,000,000 years old), outcrop in extensive areas in north Kashmir, forming the whole of the Ladākh Range and the greater part of the area of Baltistān to the west of the Drās River and to the south of the Karakoram Range (*q.v.*). The intrusion during post-Miocene times (*i.e.*, within the last 7,000,000 years) of tourmaline granite into an older series of gneisses (rocks formed by heat and pressure and made of bands that differ in colour and composition) and schists (crystalline rocks, the constituent minerals of which are usually arranged in a foliated or parallel pattern), aided by upthrusts in many areas, has given rise to many of the high peaks of the Himalayas, such as Makālu, Manāslu, and Nānga Parbat, which are typical examples. Mt. Everest and its two associated peaks, Lhotse and Cho Oyu, are, however, formed of limestones and pelitic (clay and mudstone) rocks, the latter dipping toward the north. It is possible that the entire formation of Mt. Everest is thrust up over a foundation of gneiss, as nappes or overturned folds.

The nappe structure of the Himalayas can also be seen elsewhere in the ranges. The Krol Nappes of the Simla region, in Himachal Pradesh, and the Garhwal Nappes, of Uttarakhand, are typical. These nappes are also evident in Nepal in the Nawākot and Kāthmāndu areas, which were formed along the line of the main central thrusts. This thrust zone borders the Great Himalayas to the south, rising abruptly in height and showing changes in the sequence of geological beds. The Pīr Panjāl Range, for example, owes its origin to thrust faulting; in the Simla region Krol limestones of Carboniferous Period (*i.e.*, from 280,000,000 to 345,000,000 years old) are overthrust onto much younger deposits of Pliocene Epoch (from 7,000,000 to 2,500,000 years old).

The geological structure is much simpler in the Outer Himalayas, where the foothills are mainly composed of Tertiary formations (from 2,500,000 to 65,000,000 years old), grouped under the Lower, Middle, and Upper Siwaliks. These consist mostly of freshwater deposits, such as sandstones, shales, and conglomerates. The Lower Siwaliks are from 1,800 to 6,000 feet thick, the Middle Siwaliks from 3,000 to 4,500 feet thick, and the Upper Siwaliks from 4,500 to 6,000 feet thick. At the same time that the Siwalik deposits were occurring, lacustrine (lake) deposits known as Karewas (flat-topped terraces) were

being formed in the Vale of Kashmir. Both the Karewas and Siwaliks show evidence of glaciation during Pleistocene times (from 10,000 to 2,500,000 years ago).

Major Tibetan rivers

Drainage system. The Himalayas are drained by 19 major rivers, of which the Indus (*q.v.*) and Brahmaputra (*q.v.*) are the largest, each having catchment basins about 100,000 square miles in extent in the mountains. Of the other 17 rivers, five belong to the Indus system—the Jhelum, Chenāb, Rāvi, Beās, and Sutlej, with a total catchment area of 50,958 square miles; nine belong to the Ganges system—the Ganges (*q.v.*), Yamuna, Rāmganga, Kāli (Sarda), Karnāli, Rāpti, Gandak, Bāghmati, and Kosi, draining another 84,098 square miles; and three belong to the Brahmaputra system—the Tista, Raidak, and Manās, draining another 70,769 square miles.

Most of the Himalayan rivers flow in troughs, the trends of which are generally determined by the branching ranges of the Great Himalayas. The rivers of the Indus system as a rule follow northwesterly courses, whereas most of the rivers of the Ganges–Brahmaputra systems take easterly courses while in the mountain region.

To the north of India, the Karakoram Range, with the Hindu Kush (Mountains) (*q.v.*) on the right and the Ladākh Range on the left, forms the great water divide, shutting off the Indus system from the rivers of Central Asia. The counterpart of this divide on the east is formed by the Kailas Range and its eastward continuation, the Nien-ch'ing-t'ang-ku-la Shan (Nyenchen Tangla Range), which prevent the Brahmaputra from flowing northward. South of this divide, the Brahmaputra flows eastward for about 900 miles before cutting across the Great Himalayan Range in a transverse gorge, although many of its Tibetan tributaries flow in an opposite direction, as the Brahmaputra may once have done.

The Great Himalayan Range, which normally would form the main water divide throughout its entire length, functions as such only in limited areas. This situation exists because the major Himalayan rivers, such as the Indus, Brahmaputra, Sutlej, and at least two headwaters of the Ganges—the Alaknanda and Bhāgirathi—are older than the mountains they traverse. It is believed that the Himalayas were uplifted so slowly that the old rivers had no difficulty in continuing to flow through their channels and, with the rise of the Himalayas, even acquired a greater momentum, which enabled them to deepen their valleys more rapidly. The elevation of the Himalayas and the deepening of the valleys thus proceeded simultaneously, with the result that the mountain ranges emerged with a completely developed river system cut into deep transverse gorges, ranging in depth from 5,000 feet to 16,000 feet and in width from six to 30 miles. The earlier origin of the drainage system explains the peculiarity that the major rivers drain not only the southern slopes of the Great Himalayan Range but to a large extent its northern slopes as well, the water divide being north of the crest line.

The role of the Great Himalayan Range as a watershed can, nevertheless, be seen between the Sutlej and Indus valleys for 360 miles; the drainage of the northern slopes is carried by the north-flowing Zāskār and Drās rivers, which drain into the Indus. Glaciers also play an important role in draining the higher altitudes and in feeding the Himalayan rivers. Several glaciers occur in Uttarakhand, of which the largest, Gangotri, is 20 miles long. The Mahalangur *himāl* ("snowfield"), with its Khumbu Glacier, drains the Everest region in Nepal. The rate of movement of the Himalayan region glaciers varies considerably; in the neighbouring Karakoram ranges, for example, the Baltoro Glacier moves about six feet a day, while others, such as the Khumbu, move only about a foot daily. Most of the Himalayan glaciers are in retreat.

Glaciers

Soils. Not much is known about the Himalayan soils. The north-facing slopes generally have a fairly thick soil cover, supporting dense forests at lower altitudes and grasses higher up. The forest soils are dark brown in colour and silt loam in texture and occur mainly in Uttarakhand; they are ideally suited for growing fruit trees. The mountain-meadow soils are well-developed but vary in thickness and in their chemical properties. Some of the wet, deep, upland soils of this type in the Eastern Himalayas—for example in the Darjeeling Hills and in the Assam Valley—have a high humus content that is good for growing tea. Podzolic soils (infertile, acidic forest soils) occur in a 400-mile-long belt along the valley of the Indus and of its tributary the Shyok, to the north of the Great Himalayan Range, and are also found in patches in Himachal Pradesh. Farther east, saline soils occur in the dry Ladākh Plateau. Of the soils that are not restricted to any particular area, alluvial soils (deposited by running water) are the most productive, though they occur in limited areas, such as the Vale of Kashmir, the Dehra Dūn and on the high terraces flanking the Himalayan valleys. Lithosols consisting of imperfectly weathered rock fragments deficient in humus content cover many large areas at high altitudes and are the least productive soils.

Climate. The Himalayas, as a great climatic divide affecting air- and water-circulation systems, exercise a dominating influence upon meteorological conditions in the Indian subcontinent, to the south, and in the Central Asian highland, to the north. By its situation and stupendous height, the Great Himalayan Range obstructs the passage of cold continental air from the north into India in winter and also forces the southwest monsoonal (rain-bearing) winds to give up most of their moisture before crossing the range northward, thus causing a heavy precipitation of rain and snow on the Indian side but arid conditions in Tibet. The average annual rainfall on the south varies between 60 inches at Simla and Mussoorie in the Western Himalayas and 120 inches at Darjeeling in the Eastern Himalayas. At places such as Skārdu, Gilgit, and Leh, in the Indus Valley, to the north of the Great Himalayan Range, only three to six inches of rainfall occur.

The precipitation pattern

Local relief and situation determine the meteorological variations experienced not only in different parts of the Himalayas but even on different slopes of the same range. Because of its favourable location on top of the Mussoorie Range facing the Dehra Dūn, the town of Mussoorie, at a height of about 6,100 feet, receives 92 inches of rainfall annually, as against 62 inches recorded in the town of Simla, which lies behind a series of ridges at a height of 6,600 feet. The Eastern Himalayas, being at a lower latitude than the Western Himalayas, are relatively warmer; the lowest minimum temperature so far recorded was at Simla, in the Western Himalaya, −13° F (−25° C). The average minimum temperature for the month of May, recorded in Darjeeling at 6,380 feet elevation, is 52° F (11° C). In the same month, at an altitude of 16,500 feet in the neighbourhood of Mt. Everest, the minimum temperature is about 17° F (−8° C); at 19,500 feet it falls to −8° F (−22° C), the lowest minimum being −21° F (−29° C). At this time during the day, in areas sheltered from strong winds that blow at more than 100 miles an hour, the sun is often pleasantly warm, even at that altitude.

There are two periods of wet weather—the winter rains and the rains brought by the southwest monsoon winds. Winter precipitation is due to the depressions advancing into India from the west, causing heavy falls of snow. Within the regions where western disturbances are felt, condensation takes place in upper air levels at a height of 10,000 feet from the surface; as a result, precipitation is much greater over the high mountains. It is at this season that snow accumulates around the Himalayan high peaks and that the Western Himalayas receive more rain than the Eastern Himalayas. In January, for example, Mussoorie in the west receives almost three inches, while Darjeeling to the east receives less than an inch. By the end of May, meteorological conditions are reversed. Southwest monsoon currents passing over the Eastern Himalayas reach heights of 18,000 feet; in June, therefore, Darjeeling receives about 24 inches, and Mussoorie less than eight inches. The rains cease in September, after which the finest weather in the Himalayas prevails until the beginning of winter in December.

Plant and animal life. *Plant life.* Himalayan vegetation can be broadly classified into four groups: tropical,

Types of vegetation

subtropical, temperate, and alpine. This division is based mainly on altitude and rainfall. Local variations in relief and climate, as well as exposure to sun and winds, cause considerable variation in the composition of the vegetation within each group. Tropical evergreen rain forest is confined to the humid foothills of the Eastern and Central Himalayas. The evergreen dipterocarps—a group of timber- and resin-producing trees—are common; their different species grow on different soils and on hill slopes of varying steepness. *Mesua ferrea* (rose chestnut) occurs on porous soils at altitudes between 600 and 2,400 feet; bamboos grow on very steep slopes; oaks and chestnuts grow on the lithosol, covering sandstones from Arunachal Pradesh westward to central Nepal at altitudes of from 3,600 to 5,700 feet. Alder trees grow along the watercourses on the steeper slopes. At higher elevations they are succeeded by mountain forests in which the typical evergreen is *Pandanus furcatus*, a type of screw pine. Besides these trees, some 4,000 species of flowering plants, of which 20 are palm, are estimated to occur in the Eastern Himalayas.

With the decrease of rainfall and the increase of altitude westward, the rain forests give place to tropical deciduous forests, where the timber tree, sal, is the dominant species, thriving best on high plateaus at 3,000 feet (wet sal), as well as higher up, at 4,500 feet (dry sal). Westward, steppe forest (*i.e.*, forest on an extensive plain), steppe, subtropical thorn steppe, and subtropical, semidesert vegetation occur successively. Temperate forests extend from about 4,500 to about 11,000 feet and contain conifers and broad-leaved temperate trees. Evergreen forests of oaks and conifers have their westernmost outpost on the hills above Murree, some 30 miles northwest of Rāwalpindi, in Pakistan; these are typical of the Lesser Himalayas, being conspicuous on the outer slopes of the Pīr Panjāl, in Kashmir, India. *Pinus roxburghi* (chir pine) is the dominant species at altitudes of from 2,700 to 5,400 feet. In the inner valleys this species may occur even at an altitude of 6,300 feet. Deodar cedar, a highly valued timber tree, is another species particular to the Himalayas, occurring mainly in the western part of the range. Stands of this species occur between 6,300 and 9,000 feet and also tend to grow at still higher altitudes in the upper valleys of the Sutlej and the Ganges. Of the other conifers, blue pine and spruce make their appearance between about 7,300 and 10,000 feet.

Alpine plants

The alpine zone begins above the tree line between 10,500 and 11,700 feet and extends as far as 13,700 feet in the Western Himalayas and 14,600 feet in the Eastern Himalayas. In this zone all the wet and moist alpine vegetation is to be found. Juniper is widely distributed, preferring sunny sites, steep and rocky slopes, and drier areas; on Nānga Parbat they are found even at an altitude of 12,750 feet. Rhododendron occurs everywhere but more abundantly in the wetter parts of the Eastern Himalayas, where it grows in all sizes from trees to low scrub. Mosses and lichens grow in shaded areas at lower levels where the humidity is high; flowering plants occur at high altitudes, especially on Nānga Parbat and Mt. Everest.

Animal life. The animal life of the Eastern Himalayas is derived mainly from that of the South Chinese and Indo-Chinese region. It is primarily the type of animal life found in the tropical forest and is only secondarily adapted to the subtropical, mountain, and temperate conditions prevailing at higher altitudes and in the drier western areas. The animal life of the Western Himalayas, however, has more affinities with that of the Mediterranean, Ethiopian, and Turkmenian regions. The past presence in the region of some African animals, such as the giraffe and the hippopotamus, can be inferred from fossil remains in the Siwalik deposits of the Outer Himalayas. The animal life at higher altitudes above the tree line consists almost exclusively of species, adapted to the cold, that have originated in the area, having evolved from the wild life of the steppes (extensive plains) after the Himalayan uplift. Elephants, bison, and rhinoceroses are restricted to certain areas of the forested *terai* (moist or marshy lands, now largely drained) at the base of the low hills in the Outer Himalayas. The Indian rhinoceros was once abundant all over the foothill zone of the Himalayas but is now becoming extinct; the musk deer and the Kashmir stag, or hangul, are also on the point of extinction. The Himalayan black bear, the clouded leopard, the langur monkey (a long-tailed Asian monkey), and the cat are some of the other denizens of the Himalayan forests. Himalayan goat antelopes are also found.

Exotic game animals

In higher altitudes above the tree line, the snow leopard, the brown bear, the red panda, and the Tibetan yak can occasionally be seen. The yak has been domesticated and is used as a beast of burden in Ladākh. The typical inhabitants of higher altitudes above the tree line are, however, diverse types of insects, spiders, and mites, which form the only animal life that can live as high up as 20,700 feet.

Catfish of the genus *Glyptothorax* live in most of the Himalayan streams, on the banks of which is found the Himalayan water shrew. Lizards of the genus *Japalura* are widely distributed. *Typhlops*, a genus of blind snake, is common in the Eastern Himalayas. The butterflies of the Himalayas are extremely varied and beautiful, especially the genus *Troides*.

The bird life is equally rich but is more in evidence in the east than in the west. Among some of the common Himalayan birds are different species of magpie—the black-rumped, the blue, and the racket-tailed; titmouse; chough (related to the jackdaw); whistling thrush; and redstart. A few strong fliers, such as the lammergeier (bearded vulture), the black-eared kite, and the Himalayan griffon (an Old World vulture), can also be seen in Sikkim. The snow partridge and the Cornish chough are found at elevations of 18,600 feet.

The people. Of the three principal ethnic types in the Indian subcontinent—the Indo-Aryan, the Mongolian, and the Dravidian—the first two are well represented in the Himalayas, although mixed in varying proportions in different areas. Waves of immigration into the mountains have occurred from all directions in the past and have caused intermingling of peoples. Generally speaking, the Great Himalayas and the Tethys Himalayas are inhabited by Tibetan and other Mongoloid people; the Lesser Himalayas are the home of the tall, fair Indo-Aryans. In the Outer Himalayan region of Jammu and Kashmir the Indo-Aryans are called Dogras and are divided into two main castes, Brahmins and Rājputs. In the Kashmir Valley the same type is represented by the Kashmīrī people. The Gaddis and Gūjars, who live in the hilly areas of the Lesser Himalayas, also belong to the Aryan type. The Gaddis are essentially a hill people; they possess large flocks of sheep and herds of goats and come down with them from their snowy abode in the Outer Himalayas only in winter, returning again to the highest pastures in June. The Gūjars are a migrating, pastoral people living from their herds of sheep, goats, and a few cattle, for which they seek pasture at various altitudes.

The Champa, Ladākhī, Balti, and Dard peoples live to the north of the Great Himalayan Range in the Kashmir Himalayas; the Dard are Aryan, the others Mongoloid. The Champa lead a nomadic pastoral life in the Upper Indus Valley. The Ladākhī have settled on terraces and alluvial fans flanking the Indus in Kashmir. The Balti have spread farther down the Indus Valley and have adopted Islām.

The Aryan racial type is represented by the Kanets in Himachal Pradesh and the Khāsas in Utturakhand. In Himachal Pradesh the majority of the inhabitants of Kinnaur and Lāhul-Spiti districts are Mongoloid, having immigrated from Tibet.

In Nepal, the Tibeto-Nepalese and Indo-Nepalese form the two main ethnic divisions, which are further subdivided into a large number of ethnic groups, including the Newārs, the Tamangs, the Gurungs, the Magars, the Sherpas, and the Kirāts. The Kirāts were the earliest inhabitants of the Nepal Valley. The Newārs are also one of the earliest Nepalese groups. The Tamangs inhabit the high valleys of Ganesh Himal (Nepal, southwest of Himachal Pradesh). The Gurungs live on the southern slopes of the Annapūrna Massif (mountain mass), pastur-

ing their cattle as high as 12,000 feet. The Magars inhabit western Nepal but migrate seasonally to other parts of the country. The Sherpas, who live to the south of Mt. Everest, are famed mountaineers.

The people of Sikkim belong to three distinct ethnic groups—the Lepchās, the Bhotiyas (Bhutias), and the Nepalese. Generally speaking, Nepalese and Lepchās live in western Bhutan and Bhotiyas of Tibetan origin in eastern Bhutan. Arunachal Pradesh is the homeland of several groups—the Abors or Adis, Akas, Apa Tanis, Daflas, Khamptis, Khowas, Mishmis, Mombas, Miris, and Singpho. Ethnically, all these groups are Indo-Mongoloid; linguistically, they are Tibeto-Burman. Each group lives in a distinct river valley, practicing shifting cultivation (*i.e.*, constantly changing the land on which they raise crops).

Early explorations

Mapping. The first Himalayan sketch map of some accuracy was drawn up by Father Antonio Monserrate, a Spanish missionary to Akbar's court in 1590. In 1733 a French geographer, Jean-Baptiste Bourguignon d'Arville, compiled the first map of Tibet and the Himalayan range based on systematic exploration. In the middle of the 19th century, the Survey of India organized a systematic program to measure correctly the heights of the Himalayan peaks. The Nepal and Uttarakhand peaks were observed and mapped between 1849 and 1855. Nānga Parbat, as well as the peaks of the Karakoram to the north, were surveyed between 1855 and 1859. The surveyors did not allot individual names to the innumerable peaks observed but designated them by figures and roman numerals. Thus, at first Mt. Everest was simply labelled as "h"; this was later changed to Peak XV in 1849–50. Not until 1856 were the computations sufficiently advanced for it to be realized that Peak XV was higher than any other peak in the world. By 1862 more than 40 peaks of 18,288 feet and above had been climbed for surveying purposes.

In the early 1970s the Survey of India was preparing large-scale maps of the Himalayas from aerial photographs. Parts of the Himalayas have also been mapped by German geographers and cartographers with the help of ground photogrammetry.

Transport. The difficulty of transport in the Himalayas has always constituted a barrier to economic growth. Only in recent years have new highways been built, making the Himalayan region accessible from both the north and the south. Of these, the 75-mile, all-weather Tribhuvan Rajpath road connecting Kāthmāndu, capital of Nepal, with India and the 65-mile road connecting Kāthmāndu with Kodari, on the Tibetan border, deserve special mention. The Hindustān–Tibet road—Indian National Highway No. 22—which passes through Himachal Pradesh has also been considerably improved by the government of India; this 300-mile highway, which runs through Simla, once the summer capital of India, connects the Punjab plains with the Indo-Tibetan border near the Shipki La (pass). There are only two main railroads, both of narrow gauge, penetrating into the Lesser Himalayas from the plains of India—one in the Western Himalayas, between Kālka and Simla, and the other in the Eastern Himalayas, between Silīguri and Darjeeling. There is another narrow-gauge line in Nepal, running 29 miles from Raxaul to Amlekhganj and connected with Kāthmāndu by an electrically operated aerial cableway, which transports cargo in baskets. Two other short railroads run to the Outer Himalayas—one, the railroad of the Kulu Valley, from Pathānkot to Jogindarnagar; the other from Hardwār to Dehra Dūn. A short railway, formerly running between Wazīrābād and Jammu through Siālkot, is now permanently closed.

There are two major airstrips in the Himalayas—one at Gaucher, Kāthmāndu, capital of Nepal, and the other in Srīnagar, capital of Kashmir—which are served by national and (except Srīnagar) international airways. Besides these, there are also many airstrips of local importance in the hills and in the *terai* district of Nepal. In Nepal there are 650 miles of dirt road, 200 miles of paved road, 50 miles of railroads, and 42 miles of aerial ropeway.

From the Punjab plains the only direct approach to the Vale of Kashmir is by the National Highway No. 1A from Jullundur to Uri through Jammu, Banihāl, Srīnagar, and Bāramūla. It crosses the Pīr Panjāl Mountains through a tunnel at Banihāl. The old road from Rāwalpindi to Srīnagar through Pakistan has lost much of its former importance. Within the Kashmir Himalayas, the 235-mile Srīnagar to Leh road is the longest. It was built in 1960 and connects Leh in Ladākh with the Nubra Valley, passing over the 17,730-foot-high Khardung La —the first of the high passes on the historic caravan trail to Central Asia from India. Many other new roads have been built in recent years.

Sikkim commands the historic Kālimpong to Lhasa caravan trade route, which passes through Gangtok. Before 1956, there was only one (30-mile) motorable highway running between Gangtok and Rongphu, on the Tista River near the Sikkim–West Bengal border, which then continued southward to Silīguri for another 70 miles. Since then, several roads passable by jeep have been built in the southern part of Sikkim, and some 95 miles of highway have been built in northern Sikkim, connecting Gangtok with Lachen (Lachung). A ropeway from Gangtok to Sherathang, south of the Nathu La Pass, is also in operation.

Bhutan has a 120-mile road, constructed in 1962, connecting Phuntsoling on the West Bengal–Bhutan border to Paro in western Bhutan; a 20-mile bifurcated extension runs northward to Thimphu, the capital. There are also aircraft-landing facilities at Paro and Thimphu.

Arunachal Pradesh is connected with the Brahmaputra Valley by roads running from Namsai to Chowkham, Sadiya to Roing, Pāsighāt to Dibrugarh, Along to Sonarighat, North Lakhimpur to Hapoli, and Tezpur to Bomdila.

Arable lands, minerals, and water resources

Economic resources. The Himalayas abound in economic resources. These include rich arable land, extensive grassland and forest, workable mineral deposits, and easily harnessable waterpower. The most productive arable lands in the Western Himalayas are in the Vale of Kashmir, the Kāngra Valley, the Sutlej Basin, and the terraces flanking the Ganges and Yamuna in Utterakhand; these produce 1,600,000 tons of cereals—rice, corn (maize), wheat, and millet—each year. In the Central Himalayas in Nepal most of the arable land is in the foothills and on the adjacent plains; this land produces four-fifths of the total rice production of the country, amounting to 1 percent of world production. The region also produces large crops of corn and wheat. In addition to cereals, most of the cash crops of the country—jute, sugarcane, and oilseeds—are grown in this region.

Most of the fruit orchards of the Himalayas lie in the Vale of Kashmir and in the Kulu Valley of Himachal Pradesh. Such fruit as apples, peaches, pears, and cherries, for which there is a great demand in the cities of India, are grown extensively. There are rich vineyards on the shores of Dal Lake in Kashmir, which produce grapes of good quality from which wine and brandy are made. On the hills surrounding the Vale of Kashmir grow walnut and almond trees, the nuts of which are exported to India where oil is extracted from them. Bhutan also has fruit orchards and exports oranges to India.

Of the plantation crops, tea is grown mainly on the hills and on the plain at the foot of the mountains in the Darjeeling district. Tea in limited quantity is also grown in the Kāngra Valley. Plantations of cardamom, a spice used in curry, are to be found in Sikkim, Bhutan, and the Darjeeling Hills. Medicinal herbs are grown in plantations in the Uttarkāshi and Pithorāgarh districts of Uttarakhand.

Transhumance (the seasonal migration of livestock) is widely practiced during the summer months in the Himalayas pastures, called *margs*, in Kashmir. Sheep, goats, and yak are raised on the rough grazing lands available.

Woodlands occupy at least one-third of the Himalayas, covering more than two-thirds of the total area of Bhutan and Sikkim. They constitute the greatest asset of the

mountains, although their fuller use is hampered because of inaccessibility. Logs of timber are floated down the Himalayan streams to sawmills located at the foot of the mountains. Forest-based industries, to manufacture matches, rayon, and paper pulp, are being established in Bhutan.

The Himalayas are rich in minerals, although their exploitation is restricted to the more accessible areas. Jammu and Kashmir state is the most mineralized region. Sapphires are found in the Zāskār Mountains, and alluvial gold in the nearby bed of the Indus. There are deposits of copper ore in Baltistān, and iron ores are found in the Vale of Kashmir. Ladākh contains borax and sulfur deposits.

The coal deposits of the Jammu Hills have an estimated reserve of 100,000,000 tons. Bauxite occurs in Jammu and Kashmir. Nepal, Bhutan, and Sikkim have extensive deposits of coal, mica, gypsum, and graphite and ores of iron, copper, lead, and zinc.

The Himalayan rivers have a tremendous hydroelectric potential, which has been harnessed more intensively since the five-year plans were introduced in the 1950s in India. The biggest multipurpose river-valley project is located at Bhākra-Nāngal on the Sutlej River in the Outer Himalayas; completed in 1963, it has a storage capacity of 348,218,000,000 cubic feet of water and a total installed capacity of 1,050 megawatts. Three other Himalayan rivers, the Kosi, Gandak, and Jaldhāka, have been harnessed by India, which then supplies the power to Nepal and Bhutan.

BIBLIOGRAPHY. S.G. BURRARD and H.H. HAYDEN, *A Sketch of the Geography and Geology of the Himalaya Mountains and Tibet*, 4 pt. (1907–08; rev. ed., 1933–34), the first official publication of the government of India on the subject, profusely illustrated with charts, diagrams, cross sections, sketches, and maps; S.P. CHATTERJEE, *Physiography of India* (1968), a systematic study dividing the country into physiographic regions; JYOTI BHUSAN DAS GUPTA, *Jammu and Kashmir* (1968), an account of the geography, international importance, and internal dynamics of the area, containing a detailed bibliography; FREDERIC DREW, *The Jummoo and Kashmir Territories* (1875), the most comprehensive and authoritative book on the land and people of the state of Jammu and Kashmir; G.O. DYHRENFURTH, *To the Third Pole* (1955), a general discussion including the highest peaks; A. GANSSER, *Geology of the Himalayas* (1964), the first comprehensive geology of the Himalayas, well supplemented by sketches, photographs, and profiles; PEARCE GERVIS, *This Is Kashmir* (1954), a travel book including accounts of the Vale of Kashmir, the cave of Amarnāth, Baltistān, Gilgit, Ladākh, Poonch, and Jammu; GEOFFREY GORER, *Himalayan Village*, 2nd ed. (1967), a description of the social life of the Lepchās; TONI HAGAN *et al.*, *Mount Everest* (1963), the most comprehensive book on the scientific study of the Everest region; SIR EDMUND HILLARY and DESMOND DOIG, *High in the Thin Cold Air* (1962), the story of the Himalayan Scientific and Mountaineering Expedition of 1960-61 led by Sir Edmund Hillary; A.G. JHINGRAN, *Himalayan Geology* (1971), a collection of papers on various aspects of Himalayan geology; P.P. KARAN and W.M. JENKINS, JR., *The Himalayan Kingdoms* (1963), a geographical account of Bhutan, Sikkim, and Nepal; TAKEHIDE KAZAMI, *The Himalayas* (1968; orig. pub. in Japanese, 1967), a travel book on Nepal including over 100 colour photographs; G.D. KHOSLA, *Himalayan Circuit* (1956), a travel account and social study of the Valleys of Lahūl and Spiti beyond the Great Himalaya; B.C. LAW, *Mountains and Rivers of India*, 21st International Geographical Congress, Calcutta (1968), a study of the origins and physical nature of the mountains and rivers of the Himalayas and India, and their geographical relationships; FOSCO MARAINI, *Una Spedizione alle montagne del Pakistan organizzata e finanziata nel 1958 dal Club Alpino Italiano . . .* (1960; Eng. trans., *Karakoram*, 1961), the story of the successful ascent of Gasherbrum IV peak of Karakoram Mountain by the Italian Alpine Club; KENNETH MASON, *Abode of Snow* (1955), a history of Himalayan exploration and mountaineering; S.D. PANT, *The Social Economy of the Himalayans* (1935), an economic and social study of the Himalayas between the Ganges and the Kali rivers; MANECK B. PITHAWALLA, *An Introduction to Kashmir: Its Geology and Geography* (1953), a brief review of the physical basis and socio-economic conditions of Kashmir; R.H. PHILLIMORE, *Historical Records of the Survey of India*, 4 vol. (1945), a systematic account of survey and mapping in India from earliest times; RAMA RAHUL, *The Himalaya Borderland* (1969), an overall survey of the southern half of the Himalayan borderland; U. SCHWEINFURTH, *Die Horizontale und Vertikale Verbreitung der Vegetation im Himalaya* (1957), a comprehensive study of the natural vegetation of the Himalaya, with summaries in English, French, Spanish, and Russian; H.W. TILMAN, *Nepal Himalaya* (1952), a vivid account of three journeys into Nepal.

(S.P.C.)

Hindenburg, Paul von

Paul von Hindenburg, a German field marshal in World War I, later became the second president of Germany's postwar Weimar Republic. Without special military or political qualifications, Hindenburg is usually said to have attained these posts largely because of his impressive bearing, suggesting great inner strength and calmness of mind. He thus met an apparently deeply felt need for a father figure during the crisis-ridden war and postwar years from 1914 to 1933.

Culver Pictures



Hindenburg

Paul Ludwig Hans Anton von Beneckendorff und von Hindenburg was born in the Prussian town of Posen (Poznán, Poland) on October 2, 1847. On his father's side he came from old Prussian Junker (aristocratic) stock; on his mother's, from a middle-class family—a fact he preferred to ignore but that may help account for his cautious, adaptable ways. A cadet at the age of 11, he served in the Austro-Prussian (Seven Weeks') War of 1866 and in the Franco-German War of 1870–71. In due time he was promoted to general, retiring in 1911 after an honourable but not especially distinguished career.

World War I commander

Hindenburg was called back into service in August 1914 to be the nominal superior of Maj. Gen. Erich Ludendorff. Acclaimed as one of the army's best strategists, Ludendorff was to drive a Russian invasion force from East Prussia. For this achievement, the rocklike Hindenburg, rather than Ludendorff, received the nation's applause. Soon Hindenburg's standing overshadowed that of Emperor William II; he was promoted to the rank of field marshal general, and in 1916 the Emperor was pressured into giving him command of all German land forces, with Ludendorff his coresponsible chief aide. Unable to win the war on land, they tried starving Britain into surrender by unrestricted submarine warfare, thus drawing the United States into the war and causing Germany's ultimate defeat. When they conceded defeat, Hindenburg let Ludendorff take the blame.

After the overthrow of William II, Hindenburg collaborated briefly with the new republican government. He directed the withdrawal of German forces from France and Belgium and had his staff organize the suppression of left-radical risings in Germany. With both tasks accomplished (and the old officer corps preserved in the process), he retired once more in June 1919. Living quietly in Hanover, he occasionally expressed antirepublican

views, but, on the whole, cultivated his image of nonpartisan national stature.

In April 1925, after the death of Friedrich Ebert, Hindenburg was elected the republic's second president, despite his professed monarchism. He adhered, if not to the spirit, then at least to the letter of the republican constitution. Yet his personal confidants, among them especially Maj. Gen. Kurt von Schleicher, longed for a new authoritarian regime and urged him to use his prestige and render the government more independent of parliamentary controls. Though tired of the frequent Cabinet crises, Hindenburg, fearful of any unconstitutional action and of added responsibilities, procrastinated.

When the depression set in and the government again broke up, he did appoint a Cabinet resting on his, rather than on the Reichstag's (parliament's) confidence. He authorized Chancellor Heinrich Brüning to dissolve the Reichstag should it prove uncooperative and promised to issue emergency decrees in lieu of Reichstag-enacted laws. The Reichstag was dissolved in July 1930; new elections produced an even less cooperative successor in which the antiparliamentarian National Socialists emerged as the second largest party. Brüning now governed almost exclusively by decree. Since the president's signature was required on each decree, however, Hindenburg could veto any governmental decision. Increasingly feeble, moody, and influenced by his military and landowning friends, the Marshal forced the government to spend huge amounts on the army and navy and hopelessly indebted estates at the expense of unemployment relief and other imperative needs. At the same time, Brüning's deflationist policies aggravated the economic difficulties. Unrest, sparked above all by the Nazis, kept mounting.

When Hindenburg's presidential term expired in April 1932, he ran again for the presidency as the only candidate who could defeat Hitler. He was re-elected, but mainly by the support of Brüning's Catholic Center Party and the Social Democrats, rather than the conservative nationalist circles, to whom he felt closest and who now supported Hitler. Those who did vote for him clung to him as a bulwark against Nazi lawlessness and brutality. Yet the President's confidants considered the Nazis a useful, if unpleasant, movement with whom they were sure they could come to terms. They saw in Brüning an obstacle to such an accommodation and persuaded the Marshal to dismiss the Chancellor, who had just helped to re-elect him.

Two governments, one headed by Franz von Papen, a former cavalry officer, the other by Schleicher, failed to win the support of the Nazis. Hitler insisted on becoming chancellor in any government in which his party participated, but, despite a deluge of petitions and letters, Hindenburg, who distrusted the Führer's noisy aggressiveness, would not concede him that post. In November 1932, however, when the Nazis lost 10 percent of their vote in new Reichstag elections, Papen and Hitler agreed on forming a government with Hitler as chancellor, Papen as vice chancellor, and non-Nazis in most other posts. Hindenburg was assured by Papen that Hitler could easily be controlled. When Schleicher failed in his efforts to obtain parliamentary support for his government, Hindenburg, frustrated and tired, asked for his resignation. On January 30, 1933, Hindenburg appointed Hitler chancellor of a new Cabinet in which only two other Nazis, Wilhelm Frick and Hermann Göring, held offices.

Hindenburg and Hitler

Papen's safeguards proved ineffective. Hitler quickly secured almost unlimited political power through terror, manipulations, and false promises. Hindenburg on his part accommodated himself to the new situation and, in effect, became a warm supporter of Hitler, although making an occasional innocuous gesture that seemed to set him apart from the Führer and the Nazi Party. He died on August 2, 1934, still a revered, though by then rather remote, national figure.

BIBLIOGRAPHY. GENERALFELDMARSCHALL VON HINDENBURG, *Aus meinem Leben* (1920; Eng. trans., *Out of My Life*, 1920), his evasive autobiography; EMIL LUDWIG, *Hindenburg und die Sage von der deutschen Revolution* (1935; Eng. trans., *Hindenburg and the Saga of the German Republic;* British title, *Hindenburg and the Saga of the German Revolution;* 1935), a somewhat uneven but helpful study; JOHN W. WHEELER-BENNETT, *Wooden Titan: Hindenburg in Twenty Years of German History, 1914–1934* (1936, reprinted 1963; British title, *Hindenburg: The Wooden Titan*, 1967), a readable and informative account; ANDREAS DORPALEN, *Hindenburg and the Weimar Republic* (1964), focussed on Hindenburg's presidency; TREVOR N. DUPUY, *The Military Lives of Hindenburg and Ludendorff* (1970).

(A.Do.)

Hinduism

Hinduism consists of the beliefs, practices, and socio-religious institutions of the South Asian peoples known as Hindus, principally the peoples of India and parts of Pakistan, Ceylon, Nepal, and Sikkim. Hinduism also has a following among overseas Hindu communities that are situated in parts of Southeast Asia, East and South Africa, Surinam, and in islands such as Fiji, Mauritius, and Trinidad.

This article is divided into the following topics:

I. General nature and characteristic features
  Common characteristics of Hindu belief
  Three *mārgas:* paths to salvation
II. Forms of Hinduism
  Vedism and Brahmanism
  Vaiṣṇavism
  Śaivism
  Tantrism and Śāktism
  Folk Hinduism
  Modern Hinduism
III. Rituals, social practices, and institutions
  Sacrifice and worship
  Sacred times and places
  Rituals and social status
IV. Cultural expressions of Hindu values and ideas
  Symbols and images
  Visual arts
  Theatre and dance
V. Place of Hinduism in Indian and world religions
  Hinduism and other religions of Indian origin
  Hinduism and Islām
  Hinduism and Christianity

I. General nature and characteristic features

The name Hinduism means the civilization of the Hindus (originally, the inhabitants of the land of the Indus River). Introduced *c.* 1830 by English writers, it properly denotes the Indian civilization of approximately the last 2,000 years, which gradually evolved from Vedism, the religion of the ancient Indo-European peoples who settled in India in the last centuries of the 2nd millennium BC. Because it integrates a large variety of heterogeneous elements, Hinduism constitutes a very complex but largely continuous whole, and since it covers the whole of life, it has religious, social, economic, literary, and artistic aspects. As a religion, Hinduism is an utterly diverse conglomerate of doctrines, cults, and ways of life.

The distinction between the level of popular belief and that of elaborate ritual technique and philosophical speculation is very marked and attended by many stages of transition and varieties of coexistence. Primitive magic and fetishism, animal worship, and belief in demons occur beside, and often combined with, the worship of more or less personal gods, as do mysticism, asceticism, and abstract and profound theological systems or esoteric doctrines. For example, worship of female local deities does not, in the same milieu, exclude the belief in pan-Indian higher gods or even in a single High God. Such deities are also frequently looked upon as manifestations of a High God.

Religious tolerance

In principle, Hinduism incorporates all forms of belief and worship without necessitating the selection or elimination of any. The Hindu is inclined to revere the divine in every manifestation, whatever it may be, and is doctrinally tolerant, leaving others—including both Hindus and non-Hindus—whatever creed and worship practices

suit them best. A Hindu may embrace a non-Hindu religion without ceasing to be a Hindu, and since the Hindu is disposed to think synthetically and to regard other forms of worship, strange gods, and divergent doctrines as inadequate rather than wrong or objectionable, he tends to believe that the highest divine powers complement each other for the well-being of the world and mankind. Few religious ideas are considered to be finally irreconcilable. The core of religion does not even depend on the existence or nonexistence of God or on whether there is one god or many. Since religious truth is said to transcend all verbal definition, it is not conceived in dogmatic terms. Hinduism is, then, both a civilization and a conglomerate of religions, with neither a beginning, a founder, nor a central authority, hierarchy, or organization. Every attempt at a specific definition of Hinduism has proved unsatisfactory in one way or another, the more so because the finest Indian scholars of Hinduism, including Hindus themselves, have emphasized different aspects of the whole.

COMMON CHARACTERISTICS OF HINDU BELIEF

There are, however, some—noncumulative and not even universally applicable—characteristics by which, it is said, a Hindu may be recognized.

Doctrine of Atman-Brahman. First, there is the belief in an uncreated, eternal, infinite, transcendent, and all-embracing principle, which, "comprising in itself being and non-being," is the sole reality, the ultimate cause and foundation, source, and goal of all existence. This ultimate reality is called Brahman. As the All, Brahman causes the universe and all beings to emanate from itself, transforms itself into the universe, or assumes its appearance. Brahman is in all things and is the Self (Ātman) of all living beings. Brahman is the creator, preserver, or transformer and reabsorber of everything. Though it is Being in itself, without attributes and qualities and hence impersonal, it may also be conceived of as a personal High God, usually as Viṣṇu (Vishnu) or Śiva (Shiva), who is characterized by sublime and adorable qualities. This fundamental belief in, and the essentially religious search for, ultimate reality, thus conceived—*i.e.*, for the One that is the All—have continued almost unaltered for over 30 centuries and have been the central focus of India's spiritual life.

Identity of individual self and the sacred

Concepts of iṣṭadevatā and Trimūrti. Although those Hindus who particularly worship either Viṣṇu or Śiva generally consider one or the other as their "favourite god" (*iṣṭadevatā*) and as the Lord (Īśāna) and Brahman in its personal aspect, Viṣṇu is often regarded as a special manifestation of the preservative aspect of the Supreme and Śiva as that of the "destructive" function. Another deity, Brahmā, the personification of the impersonal Brahman, remains in the background as a demiurge. These three great figures (Brahmā, Viṣṇu, and Śiva) constitute the so-called Hindu Trinity (Trimūrti, "the One or Whole with three forms"). This conception attempts to synthesize and harmonize the conviction that the Supreme Power is singular with the plurality of gods in daily religious worship. Although the concept of the Trimūrti assigns a position of increasing importance to some great gods, it never has become a living element in the religion of the masses. Moreover, Brahmā since ancient times has had no cult worth mentioning, and many Hindus worship neither Śiva nor Viṣṇu but one or another of the innumerable other Hindu gods.

Authority of the Veda and the Brahmin class. Another characteristic of Hindu belief is recognition of the Veda, the most ancient body of religious literature, as an absolute authority revealing fundamental and unassailable truth. At the same time, however, its content—*i.e.*, the Eternal Reality as Word—has long been practically unknown, so that even though it is venerated from a distance by every traditional Hindu and even though those Indians who reject its authority (such as Buddhists and Jains) are regarded as unfaithful to their tradition, it is, in fact, hardly drawn upon at all, not even for edification.

Also characteristic of Hinduism is the belief in the sacrosanctity of the Brahmins, a noble class possessing spiritual supremacy by birth. As special manifestations of Brahman and "bearers" and teachers of the Veda, Brahmins are considered to represent an ideal norm of ritual purity and social prestige. This characteristic of Hinduism is, however, losing its importance in modern times, especially in north India.

Ahiṃsā: respect for life. Further characteristics include vegetarianism and the application of the *ahiṃsā* idea to animal life. *Ahiṃsā*, consisting in respect of and consideration for life and fellow feeling with all living beings, is based on belief in the unity of all life. One of the cardinal virtues of Indian humanity, *ahiṃsā* is, according to Indian thinkers, the keystone of their ethics. Thus the protection and veneration of the cow are deemed especially important. On the other hand, about a quarter of the population do not hold eating of beef in abomination, and the double doctrine of *ahiṃsā* and vegetarianism has never found full acceptance.

Doctrines of transmigration and karman. Rarely disputed and thus generally accepted by Hindus are the doctrine of transmigration and rebirth and its complement, the belief in *karman*—that is, of previous acts, as the factor that determines the condition into which a being, after a stay in heaven or hell, is reborn. The whole process of rebirths is called *saṃsāra*, and the beings entrapped in it wander through a restless and unending stream of periodic returns to life in one form or another. Any earthly process is viewed as cyclic, and all worldly existence is subject to the cycle. *Saṃsāra* has no beginning and, in most cases, no end; it is not a cycle of progress or a process of purification but a matter of perpetual enslavement. The *karman*, acting like a clockwork that, while running down, always winds itself up, binds the *ātmans* (selves) of beings to the world, compelling them to go, in unremitting restlessness, through an endless series of births and deaths. This belief is indissolubly connected with the traditional Indian views of society and earthly life. It has given rise to an acquiescence that may verge upon fatalism—the belief that any misfortune is the effect of *karman*, of one's deeds, and so one's own doing—and to the conviction that the course of world history is conditioned by the collective *karman*.

Saṃsāra liberation

Such doctrines also encourage the view that mundane life is not true existence (the so-called Indian pessimism) and that human endeavour ought to be directed toward a permanent interruption of the mechanism of *karman* and transmigration—that is, toward final emancipation (*mokṣa*), toward escaping forever from the impermanence that, clouding all joy of living, is an inescapable feature of mundane existence. In this view there can be no other goal of this effort than the only permanent and eternal principle, the Holy, the One, God, Brahman, which is totally opposite to any phenomenal existence. Anyone who has not fully realized that his being is identical with Brahman is thus seen as afflicted, wretched. The only possible solution consists in the sudden realization that the kernel of human personality (*ātman*) really is Brahman and that it is man's very attachment to worldly objects that prevents him from reaching salvation and eternal peace. (Hindus sometimes use the largely Buddhist term Nirvāṇa to describe this state.)

THREE MARGAS: PATHS TO SALVATION

There is much difference of opinion among Hindus about the way (*mārga*) to final emancipation (*mokṣa*). Three paths to salvation (variously valued but nonexclusive) are presented in an extremely influential religious text, the *Bhagavadgītā* ("The Lord's Song"; *c.* 200 BC), according to which it is not acts themselves but the desire for their results that produces *karman* and thus enslavement. These three ways to salvation are (1) the *karma-mārga* ("the path of duties"), the disinterested discharge of ritual and social obligations; (2) the *jñāna-mārga* ("the path of knowledge"), the use of meditative concentration preceded by a long and systematic ethical and contemplative training—*yoga*—to gain a supra-intellectual insight into one's identity with Brahman; and (3)

the *bhakti-mārga* ("the path of devotion")—*i.e.*, devotion to a personal God. These ways are regarded as adapted to various types of men.

Although the search for *mokṣa* has never been the ideal of more than a small minority of Hindus, liberation has been for all a religious norm by which their lives were affected. For *mokṣa* has determined not only the hierarchical values of Indian social institutions and religious doctrines and practices but also the function of Indian philosophy, which is to discuss what man ought to do in order to find true happiness and what he has to realize, in direct experience, in order to escape from *saṃsāra* (bondage) and obtain spiritual freedom. While those who have not been reached by Brahmin thought have only vague ideas about the doctrines of *karman* and *mokṣa*, in semipopular milieus these doctrines gave rise to much speculation.

Dharma: comprehensive system of duties, morals, and religious law

For the ordinary Hindu, the main aim of worldly life lies in conforming to social and ritual duties, to the traditional rules of conduct for his caste, family, and profession. Such requirements comprise an individual's *dharma* (law and duties); *i.e.*, that stability, law, order, and fundamental equilibrium in the cosmos, nature, and society that should not be infringed for fear of undesirable results. *Sanātana* (traditional) *dharma*—a term used by Hindus to denote their own religion—is a close approximation to "religious practices" in the West. Religion for Hindus is thus mainly a tradition and a heritage, a way of life and a mode of thought. In practice, it is the right application of methods of securing welfare in this life and a good condition in the hereafter.

II. Forms of Hinduism

The several forms of Hinduism may be conveniently grouped into the following categories: (1) Vedism, the religion of the earliest Indo-European peoples who settled in India in the last quarter of the 2nd millennium BC and named after their sacred texts known collectively as the Veda; (2) Brahmanism, the religion into which Vedism gradually grew, deriving its name from both its ancient Brahmin priestly elite and the Supreme God Brahman, whom they worshipped; (3) Vaiṣṇavism (Vaiṣṇava), later sectarian cult forms that regard Viṣṇu or his earthly incarnations as the Supreme Being; (4) Śaivism (Śaiva), rival religious forms that regard Śiva as supreme; (5) Tantrism (Tantra) and Śāktism (Śākta), esoteric, magical, and devotional developments within various religious forms; and (6) folk Hinduism, an amalgam of numerous local folk beliefs and practices that have notable continuities with and differences from the other forms of Hinduism. Modern developments require separate treatment, not only because of new religious forms but also because of new amalgams of earlier forms.

VEDISM AND BRAHMANISM

Vedism, or the Vedic religion, is the earliest known form of religious practice of the ancient Indo-Aryan settlers of India. What is known of Vedism is derived from surviving texts of the Veda and certain ritual practices still performed by India's Brahmin elite. Vedic religion consisted principally of two categories of ritual practice: (1) domestic rites, known as *gṛhya*, and (2) great sacrifices and other communal rites, known as *śrauta*. Much of Vedism has died out, and whatever of it has survived has become so integral a part of Hinduism that it is no longer distinguishable as a separate element. Many features of Hinduism, however, have their roots in the Vedic past; and some characteristic ideas inherited from that past and developed in a few main currents—primarily theories of salvation—have up to the present time largely determined the character of the Indian view of life and the world.

The Veda is considered to be eternal, not even created or revealed by a god or gods, whose only task has been to promote the inspiration of gifted men (*ṛṣi*, or *rishi*, "seers") of archaic times who were able to "see" this eternal truth directly, without the intermediary of normal sense perception, and to express it in a human language, Sanskrit. Because of this origin and their character as bearers of supranormal power, the metrical portions (*mantra*) of the Vedic corpus are regarded, when solemnly recited and accompanied, as an infallible means of coming into contact with the unseen.

The sacred: nature, man, and God. The contact with the unseen or sacred also included man's contributions by ritual acts to the maintenance of the universe—of which Vedic man felt he was an indissoluble part—and to the periodic regeneration, through sacrificial practices, of both the powers for good and the cosmic processes that make earthly life and welfare possible. Vedic man was deeply convinced that the world is an actually existing organized cosmos (*sat*) governed by order and truth (*ṛta*) and that it is always in danger of being damaged or destroyed by the powers of chaos (*asat*). This conviction found mythological expression in the continual conflict between gods (*devas*) and demoniac antigods (*asuras*).

Ṛta: cosmic order

Gods were conceived as presiding over certain provinces of the universe or as responsible for important cosmic or social phenomena. Their deeds are timeless and exemplary presentations of mythic events replete with power and universal, eternal significance. To reproduce themselves in time and thus retain their vitality and efficacy, mythical events need repeating—that is, celebration and confirmation by means of the spoken word and ritual acts. Since the Vedic gods were not fully anthropomorphic—many minor deities are hardly distinguishable from impersonal powers—they were often representatives of natural functions, and not individuals, and the functions were liable to various applications and interpretations. Thus Indra, the greatest and most anthropomorphic god of the early Veda, was, in the view of the noble patrons of the Vedic poets, primarily a fighter and a warrior god invoked to bring booty and victory. Agriculturalists, however, emphasized Indra's fecundity, celebrating until recent times his very ancient festivals to produce fertility, welfare, and happiness. Indra, however, was essentially a representative of useful force in the cosmos and nature and the great champion of an ordered and habitable world. He won repeated victories over the snake-demon Vṛtra, the representative of obstruction and chaos, and these resulted in the separation of heaven and earth (the support of the former and the stabilization of the latter), the rise of the sun, and the release of the waters: in short, in the organization of the universe.

Vedic and Brahmanic theological principles and rites. Vedic religion—as it is now known—is primarily a liturgy differentiated in various types of ritual designed for almost any conceivable purpose. These rites are described in ritual handbooks (*sūtra*) in the minutest detail—no operation, no gesture, no formula is meaningless or left to an officiant's discretion. On the basis of a complicated speculative system, all are explained and shown to be effective in later prose works called *Brāhmaṇas*. The often very complicated ritual technique was devised mainly to safeguard man's life and survival, to enable him to face the many risks and dangers of existence, to thwart the designs of human and superhuman enemies that cannot be counteracted by ordinary means, to control the unseen powers, and to establish and maintain beneficial relations with the supramundane sacred order. Belief in the efficacy of the rites was the natural consequence of the belief that all things and events are connected with or participate in one another. Hence it was also believed that a close correspondence exists between a sacred place—for instance, the sacrificial place of many Vedic rites or a place of pilgrimage or consecrated area (*maṇḍala*, "circle")—and a province of the universe or even the universe itself. Such places represent, within the reach of the officiants, the universe or as much of it as was relevant. In such places, direct communication with other cosmic regions (heaven, underworld) is possible because they are said to be at the point of contact between this world and the "pillar of the universal," "the navel of the earth." The sacred place is (by virtue of a system of connections) identical with the universe in its various states of emanation from, reabsorption into, integration with, and disin-

Cosmic role of sacred places

tegration from the sacred. This idea has as its corollary the possibility of ritually enacting the cosmic drama and, thus, of influencing, through the same system of connections, those events in the cosmos that continuously affect man's weal and woe.

In the major *śrauta* rites, requiring three fires and up to 16 priests or more, "the man who knows"—he who has an insight into the correspondences (*bandhu*) between the mundane and cosmic phenomena and the eternal transcendent reality beyond them and who knows the meaning of the ritual words and acts—may set great cosmic processes in motion for the sake of human interests. In these rites, Brahmin officiants repeat the mythic drama for the benefit of their patron, the "sacrificer," who, temporarily, becomes its centre and realizes through ritual "symbolism" his identity with the universe. Whatever magical elements may be involved in this ritual technique, its aim in establishing an efficacious contact with a transcendental order that is the source of all life and power is based on an essentially religious conception.

Such officiants are firmly convinced of the efficacy of their rites: "the sun would not rise, were he [the officiant] not to make that offering; this is why he performs it" (*Satapatha Brāhmaṇa* 2. 3. 1. 5). The oblations might not be used to propitiate the gods, even less thank them for favours bestowed. The efficacy of the rites—some of which are sporadically performed up to the present day—did not depend on the will of the gods, who were regents presiding over definite phenomena and were, according to later exegesis, only indispensable, hypothetical entities required as addresses of the oblations to make the acts effective. The offerings do not require the postulation of any High God to whose beneficence the merit earned from performing the ritual should be regarded as due.

The origin of the world is of special interest to the authors of the Ṛgveda, the oldest document. Among its many incomplete and often unsystematic speculations, one came into prominence. In Ṛgveda 10, 90, the phenomenally evolved universe is described as having issued from a primeval Person, Puruṣa, whose origin is not explained. In Puruṣa, God and matter are one; he is the All. Everything existing in the world is only one-quarter of his being; three-quarters of him is immortal and in heaven. It was from this Person, the ultimate reality, who is the universe and transcends it, that the "goddess" Virāj (literally, the one "extending and ruling far and wide") was born. From Virāj (actually the hypostatization of the conception of the universe as an evolving, expanding, and creative whole) Puruṣa arose again in an evolved state. Puruṣa then became the victim in a primeval and exemplary sacrifice performed by the gods, and from his dissected body there arose the particular elements of the phenomenal creation. The poem in the Ṛgveda in which this is dealt with is the first expression of the fundamental Indian idea that creation is the self-limitation of the transcendent—in this instance conceived as a Person—manifesting itself in the realm of human experience.

Puruṣa: primeval cosmic man

In the course of the Vedic period Puruṣa fused with the figure Nārāyaṇa ("Scion of man") and with Prajāpati ("the Lord of Beings"), the patron of procreation in popular belief. In the speculative thought of the ritualists, Prajāpati came to the fore as the creator god and in many respects as the highest divinity, the immortal father even of the gods, whom he transcends and encompasses while molding them into one complex. As the One, the concentrated All, or Totality, Prajāpati was identified with the highest and most general categories. By a process of emanation and self-differentiation (by dividing up himself), he created all beings and the universe. After this "creation," Prajāpati became the disintegrated and differentiated All of the phenomenal world and was exhausted. He then reintegrated himself to prepare for a new phase of creativity by means of a rite. Since the sacred act aims at the restitution of the organic structural norm, which ensures the ordered functioning of the universe, Prajāpati was identified with the rite; by identifying himself with Prajāpati, a sacrificer may temporarily reintegrate within himself what has been disintegrated, and thus restore oneness and totality in himself and the universe.

In the course of the *Brāhmaṇa* period, the fear of the impermanence of religious merit and its loss in the hereafter, as well as the fear-provoking anticipation of the transience of any form of existence after death, culminating in the much-feared repeated death (*punarmṛtyu*), assumed the character of an obsession. The means of escaping and conquering death and of attaining integral life devised in the *Brāhmaṇas* were of a ritual nature, but in one of the oldest *Upaniṣads* (early religio-philosophic texts), the *Bṛhadāraṇyaka Upaniṣad* (6th century BC), more emphasis was laid upon the knowledge of the cosmic connection underlying ritual. When the doctrine of the identity of Ātman and Brahman was established in the *Upaniṣads*, the true knowledge of the "Self" and the realization of this identity was (by those who were inclined to speculative and meditative thought) substituted for the ritual method.

Development of Ātman-Brahman doctrine from ritual speculation

In the following centuries, the main theories connected with the divine essence underlying the world were harmonized and synthetized, and the tendency to extol one god as the supreme Lord and Originator (Īśvara), who at the same time is Puruṣa and Prajāpati and Brahman and the inner Self (Ātman) of all beings, gained ground. For those who worshipped him, he became the goal of identificatory meditation, which leads to complete cessation of phenomenal existence and becomes the refuge of those who seek eternal peace.

Ethical and social doctrines. In Vedic times, "sin" (*enas*) or evil (*pāpman*) was put on a par with illness, enmity, distress, or malediction to the extent that it was conceived of as a sort of pollution that might be neutralized by ritual or devices for averting evil. A man might incur "sin" by any incorrect or improper behaviour, in particular improper speech, and be thus guilty of *anṛta* (*i.e.*, any infidelity to fact or nonconformity with what is true, real, and constitutes the established order) without regard to whether he wittingly or unwittingly had committed a crime. Other transgressions included making mistakes in sacrificing or coming into contact with corpses, ritually impure persons, or persons belonging to the lower classes of society. Such acts were only rarely considered to be misdeeds against a god or violations of moral principles of divine origin, and consciousness of guilt was much more rare than fear of the evil consequences (disease, untimely death, and any other form of mishap). Sometimes, however, a god (Agni, the evil-devouring fire, or Varuṇa, the god of order supposed to punish and fetter the "sinner") was invoked to forgive the neglect or transgression or to release a man from their concrete results. More usually, however, these results were done away with by means of purifications, such as the ceremonial use of water, and a variety of expiatory rites.

To the pure who earned ritual merits, the prospect of a safe "world" (*loka*) or condition was held out. The meticulous effort to purify oneself from every kind of evil also involved the observation of various customs with regard to the avoidance or expiation of inauspicious occurrences (this endeavour is called *śānti*). Ritual purity was the principal concern of the compilers of the manuals of *dharma* (religious law) that, belonging to the sacred tradition (*smṛti; i.e.*, remembered by human teachers), have contributed much to the special character of Hinduism. According to the authorities on *dharma*, ritual purity is the first approach to *dharma*, the resting place of the Veda (Brahman), the abode of prosperity (*śrī*), the favourite of the gods, and the means of clearing (soothing) the mind and of seeing (realizing) the *ātman* in the body.

Importance of ritual purity

VAISNAVISM

Vaiṣṇavism is the worship of Viṣṇu (Vishnu) and his various incarnations (usually said to number ten). During a long and complex development from Vedic times, there arose many Vaiṣṇava groups with differing beliefs and aims. Some of the major Vaiṣṇava groups are the Śrī-

vaiṣṇavas and Dvaitins ("philosophical or religious dualists") of south India, the followers of the teachings of Vallabha in western India, and several Vaiṣṇava groups of Bengal in eastern India who follow teachings derived from those of the saint Caitanya (1485–1533). The majority of Vaiṣṇava believers, however, take what they like from the various traditions (described below) and blend it with various local practices (described under *Folk Hinduism* below).

Views of nature, man, and the sacred. In the Veda, Viṣṇu is the god of far-extending motion and pervasiveness who, for man in distress, penetrates and traverses the spaces to make his existence possible. All beings are said to dwell in his three strides or footsteps (*tri-vikrama*): his highest step, or abode, is beyond mortal ken in his dear and highest resort, the realm of heaven. So Viṣṇu is also the god of the pillar of the universe and identified with the sacrifice. He imparts his all-pervading power to the sacrificer who imitates his strides and so identifies himself with him, conquering the universe and attaining "the goal, the safe foundation, the highest light" (*Śatapatha Brāhmaṇa*).

Composite historical nature of Viṣṇu

In the centuries preceding the beginning of our era, Viṣṇu became the Īśvara (immanent deity) of his special worshippers, fusing with the Puruṣa-Prajāpati figure; with Nārāyaṇa, whose cult discloses a prominent influence of ascetics; with Kṛṣṇa (Krishna), who in the *Bhagavadgītā* revealed a religion of wide catholicity, open to everybody desiring to lead a socially normal life while having a prospect of final liberation; and with Vāsudeva, adored by a group known as the Pāñcarātrins.

Whatever justification the different Vaiṣṇava groups (*e.g.*, the Śrīvaiṣṇavas of south India or the worshippers of Viṣṇu Viṭhobā in Mahārashtra) offer for their philosophical position, all Vaiṣṇavas believe in God as a person with distinctively high qualities and worship him through his manifestations and representations. Vaiṣṇava faith is essentially monotheistic, whether the object of adoration be Viṣṇu Nārāyaṇa or one of his incarnations (*avatāras*), such as Rāma or Kṛṣṇa. Preference for any one of these manifestations is largely a question of tradition. Thus, the south Indian Śrīvaiṣṇavas mostly prefer Viṣṇu, Rāma, or Śrī (Viṣṇu's consort); the north Indian groups, Kṛṣṇa. The *avatāra* doctrine, in accommodating the cults of various divine or heroic figures within a monotheistic framework, proved a powerful integrating force. Whenever the *dharma* declines and evil and general disaster threaten, God, the protector and preserver of the world, emanates himself and assumes an earthly form to guard the good, to destroy the wicked, and to confirm the *dharma* (*Bhagavadgītā* 4, 6 ff.). The benevolence and beneficial activity of these figures (Rāma, Kṛṣṇa, etc.) is rarely in doubt. In many mythical tales, Viṣṇu is depicted as a versatile figure of great adaptability, able, for instance, to disguise himself as a fascinating young woman in order to trick the *asura*s (antigods) out of the possession of the newly-produced draft of immortality. His absorbing many-sided character was a source of inspiration for various stories in which his figure displays sentimentality and exhibits a view of humanity. Moreover, the scene of his great deeds is usually laid in this world, especially India, in places often mentioned by name. The narratives are full of the miraculous, but their central figures give the impression of human, sometimes all too human, characters whose actions and reactions are within the limits of ordinary understanding.

Avatāras: Viṣṇu's incarnations

The theologians had to assume the task of explaining the relation between God, as the unaffected and unchanging cause of all things, and the universe. According to Rāmānuja (*c.* 1050–1137), a great south Indian thinker of Śrīvaiṣṇava persuasion, Brahman (*i.e.*, God) is a Person with high attributes, the object of a higher knowledge that is the only entrance to salvation. Since an absolute creation is denied, God is viewed as the sole cause of his own modifications; viz., the emanation, existence, and absorption of the universe. Although unlimitedly expansive, God is conceived to be essentially different from everything material, the absolute opposite of any evil, free from any imperfection, omniscient, omnipotent, possessed of all positive qualities (such as knowledge, bliss, beauty, and truth), of incomparable majesty, the inner soul of all beings, and the ultimate goal of every religious effort. The universe is considered a real transformation of Brahman, whose "body" consists of the conscious souls and everything unconscious in their subtle states, if Brahman is viewed as the cause, and of the same in their gross states, if he is viewed as the product. The *karman* (see above) doctrine is modified as follows: the Lord, having determined good and bad deeds, provides all individual souls with a body in which they perform deeds, reveals to them the scriptures from which they may learn the *dharma* (duties of life), and enters into them as their internal regulator. The individual acts at his own discretion but needs the Lord's assent. If the devotee wishes to please him, God induces him, with infallible justice and loving regard, to intentions and effort to perform good deeds by which the devotee will attain to him; if not, God keeps him from that goal.

Vaiṣṇava theology of Rāmānuja

The *Bhāgavata-Purāṇa* (written *c.* 10th century in Tamil-speaking south India) teaches a quite representative Vaiṣṇava theology: God is transcendent and beyond human understanding; he is the universal causality, creator and substratum, time and the bearer of all possibilities that are susceptible of actualization; through his incomprehensible creative ability (*māyā*) or specific power (*ātmaśakti*) he expands himself into the universe, which he pervades and which is his outward appearance (his immanence); so he is the All and everything and the inner Self of all beings. When God is conceived of as Brahman, he is immutable and as such the Puruṣa (cosmic Person) who is not the universe; if, however, his creation is thought to be in him, he is the world.

Accepting the *Bhāgavata-Purāṇa* as a high scriptural authority, Bengal Vaiṣṇavism considers God the ground and subsistence of whatever exists, from whom all objects have come, by whom they continue to be, toward whom they move, and into whom they enter at the final dissolution at the end of this world, unless they come to him in the state of emancipation (*mokṣa*). Between God and the world there is a relation of inconceivable difference in identity and identity in difference (*acintyabhedābheda;* literally, "unthinkable difference and nondifference"). The Lord creates the world merely because he wills to do so. Creation, or rather the process of differentiation and integration, is his sport (*līlā*). The world is real, but reality has two aspects; viz., the transcendent and eternally real and that which is progressively realized and, in the process, bound up with the eternal aspect.

Vaiṣṇava theological principles and rites. The day of the faithful Śrīvaiṣṇava Brahmin is usually devoted to five pursuits: purificatory rites, collecting the requisites for worship, acts of worship, study and contemplation of the meaning of the sacred books, and meditative concentration on the Lord's image. The performance of sacrifices and other rites, restraint of the senses, fasting and soberness, worship, recitation of the scriptures, and visits to sacred places are a lifelong obligation. In addition, to those who aspire to liberation, Rāmānuja recommends concentration on God, a virtuous way of living, and insensibility to luck and misfortune. According to Madhva (1199?–1278?), a faithful observance of all regulations of daily conduct—including bathing, breath control, etc. —will contribute to eventual success in the quest for liberation.

Devout Vaiṣṇavas are inclined to emphasize God's omnipotence and the far-reaching effects of his grace and attach much value to the repeated muttering of his name or sacred formulas (*japa*), to his praise and the commemoration of his deeds as potent means of self-realization and of unification with his essence. Special stress is laid on *ahiṃsā* as a virtue of those who are born to the divine estate.

Bhakti

A very pronounced feature of Vaiṣṇavism is the strong tendency to "devotion" (*bhakti*), which is generally considered to be "the heart of worship," the sole true religious attitude toward a personal God, and the very foun-

dation of the realization of man's relationship with him. Characterized by a continual consciousness of participating in God's essence, *bhakti* is the disinterested performance of all deeds for him, a passionate love and adoration of and a complete surrender to him. The widespread *bhakti* movement is a corollary of the Vaiṣṇava ideal of a loving personal God and aversion to a conception of salvation that puts an end to all consciousness or individuality. Attesting to the superiority of a mystic and emotional attitude to the meditative or preponderantly ritualistic means to the highest goal, the practical and theoretical development of the *bhakti* idea constitutes one of the main points of difference among the several Vaiṣṇava schools. The belief already expressed in the *Bhagavadgītā* (18, 62) that those who seek refuge in God with all their being will by his benevolence and grace (*prasāda*) win peace supreme, the eternal abode, was generally accepted: *bhakti* will result in divine intercession with regard to the consequences of one's deeds.

Among many followers of Rāmānuja, however, complete self-surrender (*prapatti*) came to be distinguished from *bhakti* as a superior means of spiritual realization. One of the chief purposes of the *Bhāgavata-Purāṇa* is the glorification of an intensely personal and passionate *bhakti* that gradually develops into a decidedly erotic mysticism, independent of all alternative means of salvation. According to this text, there are nine characteristics of *bhakti:* listening to the sin-destroying sacred histories; praising God's name; recollecting and meditating on his nature and salutary endeavour (resulting in a spiritual fusion of devotee and God); serving his image; adoring him; respectful salutation; servitude; friendship; and self-surrender. Meritorious works are now also an element of *bhakti.*

In later Bengal Vaiṣṇavism, the emphasis shifts from service and surrender to mutual attachment and attraction between God (*i.e.*, Kṛṣṇa) and man: God is said to yearn for man's identification with himself, which is his gift to the wholly purified devotee. All the mystical and devotional possibilities of the Kṛṣṇa legend are made subservient to religious practice; the divine sport and wonderful feats of this youthful hero are interpreted symbolically and allegorically. The highest fruition of *bhakti* is admission to the eternal sport of Kṛṣṇa and his beloved Rādhā, whose sacred love story is explained as the mutual love between God and the human soul. Various gradations of *bhakti* are distinguished, such as awe, subservience, parental affection. These are to be correlated with the persons of the Kṛṣṇa legend; the highest and most intimate emotion is said to be the love of Rādhā and her girl friends for Kṛṣṇa.

The work of Caitanya

These ideas were developed into the practice of devotional ecstasy by Caitanya (1485–1533), who had a profound and continuing effect on the religious sentiments of his Bengali countrymen and propagated the community celebration (*saṃkīrtana*) of Kṛṣṇa as the most powerful means of bringing about the proper *bhakti* attitude. Caitanya also introduced the worship of God, the director of man's senses, through the very activity of man's senses, which must be free from all egoism and completely filled with the intense desire (*preman*) for the satisfaction of the beloved (*i.e.*, Kṛṣṇa).

The influence of the *bhakti* movement had earlier led Rāmānuja to admit a twofold possibility of emancipation: besides the meditative method of the highest insight (*jñāna*) into the oneness of soul and God, which destroys the residues of *karman* and propitiates God to win his grace, there is the way of *bhakti.* Those who prefer the former way will reach a state of isolation, the others an infinitely blissful eternal life in, through, and for God, with whom they are one in nature but not identical. They do not lose their individuality and may even meet Viṣṇu in his Vaikuṇṭha heaven and enjoy delight beyond description.

Vaiṣṇava ethical and social doctrines. The *Bhagavadgītā,* by demanding that God's worshippers fulfill their duties—"better one's own duty ill-done than another's well-performed" (3.35)—and observe the rules of moral conduct, disinterestedly bridged the chasm between ascetic morality and the search for emancipation, on the one hand, and the exigencies of daily life, on the other. For those who must lead a normal life in this world, the *Bhagavadgītā* gave a moral code and a prospect of final liberation. In so doing, the work thus founded, on the basis of the Vaiṣṇava tradition, what may be called a social ethic. Since God is in all beings as their physical and psychical substratum—and collectively in human society—the wise should not see any difference between his fellow creatures and should love God in them equally. Like God himself, the devotee should be impartial—the same to friend as to foe. The serious endeavour to realize God's presence in man requires humility and a complete unconsciousness of oneself as a corollary of the consciousness of the Presence. It demands selfless dedication of all one's actions, duties, and ceremonies to the Lord and obliges one to promote both individual and social uplift and welfare.

Bhagavāt and Bhāgavata

According to the *Bhāgavata-Purāṇa,* the highest Bhāgavata—worshipper of the Bhagavāt (God: literally, "the adorable possessed of all excellences")—sees himself in all beings and all beings in the Bhagavāt; free from hatred and partiality and knowing God to be present in all beings, he loves him also in them. Those who cannot reach this level will in any case have friendly relations with coreligionists, irrespective of their birth or social status, and take compassion upon the infatuated. The true Vaiṣṇava should worship Viṣṇu or one of his *avatāras* (incarnations), construct temples, bathe in holy rivers, study religious texts, serve superiors, and honour cows. In social intercourse with the adherents of other religions he tends to be passively intolerant, avoiding direct contact, without injuring them or prejudicing their rights. He should not neglect other gods but must avoid behaving ritually like their followers. Misuse of the advantages of birth is severely condemned, and those who apply themselves mainly to the acquisition and enjoyment of wealth are hardly supposed to be qualified for *bhakti.* The concept of class divisions is accepted, but the idea that possession of their characteristics is the inevitable result of birth is decidedly rejected. Since sin is antithetical to *bhakti,* a Brahmin who is not free from jealousy, falsehood, hypocrisy, envy, injury, and pride cannot be the highest of men, and many persons of low social status may have some advantage over him in moral attitude and behaviour. The most desirable behaviour is compatible with *bhakti* but independent of class.

In establishing *bhakti* religion against any form of opposition and defending the devout irrespective of birth, the Bhāgavata religion as such did not propagate social reform; but the attempts to make religion an efficient vehicle of new spiritual and social ideas, especially Caitanya's movement, contributed, to a certain extent, to the emancipation of lowborn followers of Viṣṇu.

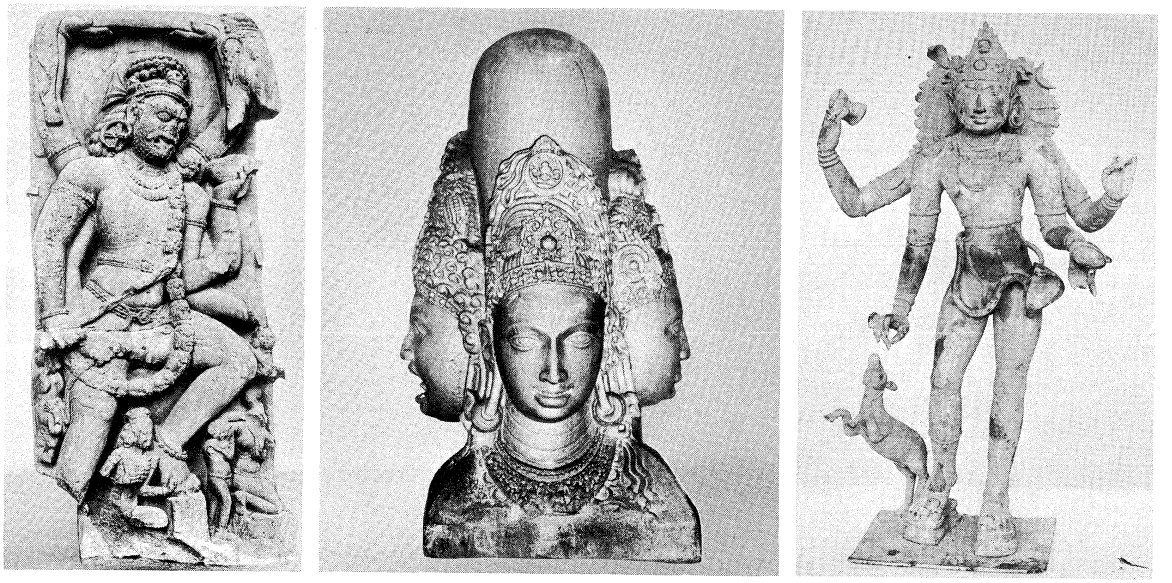
SAIVISM

Śaivism is the worship of Śiva (Shiva) in his various forms and manifestations. During a complex development from ancient, possibly in part from pre-Vedic, times, many different Śaiva groups arose. Despite the fact that major groups such as the Vīraśaivas of southern India and the Kashmir Śaivas contributed the theological principles of Śaivism, most Śaiva worship is not systematic but a complex amalgam of pan-Indian Śaiva philosophy and local or folk worship.

The Vedic god Rudra—called Śiva, "the Mild or Auspicious One," when this aspect of his ambivalent nature is emphasized—had a character and a position of his own, which remain clearly perceptible in some important features of the great god who together with Viṣṇu came to dominate Hinduism.

Composite historical nature of Śiva

In the minds of the ancient Indians Śiva must have been primarily the divine representative of the uncultivated, dangerous, unreliable, and much-to-be-feared aspects of nature. Śiva's character lent itself to being split up into partial manifestations—each said to represent only an aspect of him—as well as to assimilating divine or demoni-



Paradoxical nature of Śiva as indicated by differing representations.
(Left) Gajāntaka, destroyer of the elephant demon, sculpture from Orissa, 8th century. In the Indian Museum, Calcutta. (Centre) Caturmukha Linga, *liṅga*, or phallic symbol, with four faces, sculpture, 5th century. In worship at Nāchna-Kuthara, Madhya Pradesh, India. (Right) Bhikshatana, the naked ascetic with his begging bowl, early Cōla bronze, from Tiruvengadu District, 1048. In the Thanjāvūr Art Gallery, India.
Pramod Chandra

ac powers of a similar nature from other deities. Already in the Ṛgveda, deprecations and a frequent appeal to him for help in case of disaster—of which he might be the originator—were combined with the confirmation of his great power. In the course of the Vedic period, Śiva—originally a ritual and conceptual outsider, yet a mighty god whose benevolent aspects were readily emphasized—gradually gained access to the circle of honourable gods who preside over various spheres of human interest. Many characteristics of the Vedic Prajāpati, or Brahmā, the personification of Brahman, and of the great Vedic god of fire, Agni, have been accreted to the figure of Śiva.

Śaiva views of nature, man, and the sacred. In those circles that left the world the *Śvetāśvatara Upaniṣad* (*c.* 400 BC), Śiva rose to the highest rank. In grandiose terms its author desires both to show a way of escape from *saṃsāra* and to proclaim Śiva the sole eternal Lord but to establish, primarily, Śiva's existence. In the description of Śiva's nature, some of the most salient features of the later Śiva, the Īśvara (immanent deity), are clearly discernible: he is the ultimate foundation of all existence and the source and ruler of all life, who, while emanating and withdrawing the universe, is the goal of that identificatory meditation that leads to complete cessation from phenomenal existence. While Viṣṇu became a friend nearer to man, Rudra-Śiva developed into an ambivalent and many-sided lord and master. His "doubles" or partial manifestations remain to be distinguished: as Paśupati ("Lord of Cattle"), he in a way took over the fetters of the Vedic Varuṇa; as Aghora, he shows the uncanny traits of his nature (evil, death, punishment) but also their opposites.

It is not always clear in particular cases whether Śiva is invoked as a great *deva* (god) of frightful aspect, capable of conquering demoniac power, or as the boon-giving Lord and protector. The Īśvara idea of a Highest Being demonstrably beyond contingency is rather abstract; hence its propagators could not do without the imagery, popular belief, and mythical thought of the masses. Śiva might be the sole Principle above change and variation, yet he did not sever his connections with innumerable local deities and much-feared powers worshipped by the masses, who still continue to invoke him in magical rites. Whereas Viṣṇu champions the cause of the gods, Śiva sometimes sides with the demons.

Śiva is a typical example of polarity within the Highest Being since he reconciles in his person semantically opposite though complementary aspects: he is both terrible and mild, creator and agent of reabsorption, eternal rest and ceaseless activity. These seeming contradictions make him a paradoxical figure, transcending humanity and assuming a mysterious sublimity of his own. His character is so complicated and his interests are so widely divergent as to lead him in mythical narratives into conflicting situations. Yet, although Brahmin philosophers like to emphasize his ascetic aspects and the ritualists of the Tantric tradition (see below) his sexuality, the seemingly opposite strands of his nature are generally accepted as two sides of one character.

Power of chastity and sexuality

Śiva interrupts his austerity and asceticism (*tapas*), which is sometimes described as continuous, to marry Pārvatī—he is even said to perform ascetic acts in order to win her love—and he combines the roles of lover and ascetic to such a degree that his wife must be an ascetic (*yogīnī*) when he devotes himself to austerities and a lustful mistress in the other case. This dual character finds its explanation in the ancient double conviction that unrestrained sexual intercourse is conducive to the fertility of nature and that the chastity and continence of the ascetic produce marvellous events and have an uncommon influence upon the unseen. By his very chastity, an ascetic accumulates (sexual) power that can be discharged suddenly and completely so as to produce marvellous results such as the fecundation of the soil. From various mythical tales it is seen that both chastity and the loss of chastity are necessary for fertility and the intermittent process of regeneration in nature. Ascetics engaging in erotic and creative experiences are a familiar feature in Hinduism. The element of teeming sexuality in mythological thought counterbalances the Hindu bent for asceticism. Such sexuality, while rather idyllic in Kṛṣṇa, assumes a mystical aspect in Śiva, which is why the devotee can see in him the realization of the possibilities of both asceticism and the householder state. His marriage with Pārvatī is, then, a model of conjugal love, the divine prototype of human marriage, sanctifying the forces that carry on the race of men.

Comparison of Śiva and Viṣṇu

Unlike Viṣṇu and his incarnations, there is indeed little about Śiva that is human. His myths—fewer in number than those produced by the Vaiṣṇavas—tend to depict him as the absolutely mighty unique One, who is not responsible to anybody or for anything. Much less active than Viṣṇu, he is a god of poses that express aspects of his nature: as a dancer, he is the originator of the eternal

rhythm of the universe; he also catches the waters of the heavenly Ganges River, which destroy all sin; and he wears in his headdress the crescent moon, which drips the nectar of everlasting life.

Authors of Śaivite *Purāṇas* (sacred encyclopaedic texts) established two ingenious and complementary doctrines to explain the nature and omnipotence of God (the force that rules, absorbs, and reproduces the world and that in performing any one of these acts necessarily performs the other two as well), the existence of the world and the identity of God and the world. A theory of five "faces," or manifestations—each of which is given mythological names and related *mantras*—is of great ritual significance. It associates Śiva's so-called creative function, by which he provokes the evolution of the material cause of the universe, with his first face or aspect; its maintenance and reabsorption with his second and third faces or aspects; his power of obscuration, by which he conceals the souls in the phenomena of *saṃsāra*, with his fourth face or aspect; and his ability to bestow his grace, which leads to final emancipation, with the fifth face or aspect. The five functions are an emanation of the unmanifested Śiva who is the transcendent Brahman.

The faces became central elements of a comprehensive classification system. They were identified with parts of God's body, regions of the universe, various ontological principles, organs of sense and action, and the elements. The system was used to explain how Śiva's being is the All and how the universe is exclusively composed of aspects or manifestations of Śiva. In his fivefold nature, Śiva was shown to be identical with the 25 (five times five) elements or principles assumed by the very prominent Sāṃkhya school of Indian philosophy. The special significance of the number five in Śaivism can be understood as a philosophical elaboration of the time-honoured fourfold organization of the universe. (The four quarters of the sky also play a prominent part in religious practice.) According to this conception, a fifth aspect, when added to the four, is considered the most important aspect of the group because it represents each of the four and collectively unites all their functions in itself. The system finds its complement in the doctrine of the five Sadākhyas (five items that bear the name *sat*, "is" or "being") representing the five aspects of that state, which may be spoken of as the experience of "there is" (*sat*) and which have evolved from God's fivefold creative energy (*śakti*). In these God "dwells" in his aspect called Sadāśiva ("the Eternal Śiva"), which is regarded sometimes as a manifestation of and sometimes as identical with the Supreme Being.

Another Śaiva doctrine posits eight "embodiments" of Śiva as the elements of nature (ether, wind, fire, water, and earth), sun, moon, and the sacrificer or consecrated worshipper (also called *Ātman*). To each of these eight elements corresponds one of Śiva's traditional names or aspects, to the last one, usually Paśupati. The world is a product of these eight forms, consists of them, and can only exist and fulfill its task because the eight embodiments cooperate. Since each individual is also composed of the same eight realities (*e.g.*, the light of man's eyes corresponds to that of the sun), Śiva makes up the corporeal frame and the psychical organism of every living being. The eighth constituent is the indispensable performer of the rites that sustain the gods who preside over the cosmic processes and are really Śiva's faculties.

Regional variations in Śaivism

Although Śaivism is a much more coherent whole than Vaiṣṇavism, there evolved, in different parts of India, some branches with peculiarities of their own. According to the pronouncedly idealist monism of Kashmir Śaivism, an important religio-philosophic school, Śiva manifests himself through a special power as the first cause of creation, and he also manifests himself through a second power as the innumerable individual souls who because of a veil of impurity forget that they are the embodiment of the Highest. This veil can be torn off by intense faith and constant meditation on God, by which the soul transmutes itself into a universal soul and eventually attains liberation through a lightning-like, intuitive insight into its own nature. Those Hindus who adhere to this group consider their doctrine a manifestation of the highest Reality, Knowing Consciousness, neither personal nor impersonal; as Śiva in the form of the transcendent Word, which is his unspoken Thought, the content of which is the universe.

The Śaiva-siddhānta, a prominent religio-philosophic school of Tamil-speaking South India, assumes three eternal principles: God (who is independent existence, unqualified intelligence, and absolute bliss), the universe, and the souls. The world, since it is created by God (efficient cause) through his conscious power (instrumental cause) and *māyā* (material cause), is no illusion. The main purpose of its creation is the liberation of the beginningless souls, which are conceived as "cattle" (*paśu*) bound by the noose (*pāśa*) of impurity (*mala*) or spiritual ignorance, which forces them to produce *karman*. The *karman* process, however, proves to be a benefit; for as soon as the soul has sufficiently ripened and reached a considerable state of purity enabling it to strive after the highest insight, God graciously intervenes, appearing in the shape of a fully qualified and liberated spiritual guide (guru), through whose words God permits himself to be realized by the individual soul.

Śaiva theological principles and rites. Ascetic tendencies were much in evidence among the Pāśupatas, the oldest Śaivite tradition in north India, the last adherents of which now live in Nepal. Pāśupatas often gave offense because of their customs and ritual practices. Their *yoga*, consisting of a constant meditative contact with God in solitude, made them frequent burial places for cremated bodies. More extreme groups carried human skulls (hence the name Kāpālikas, from *kapāla*, "skull") and a bowl of liquor in which they projected and worshipped Śiva as Kapālika, "the Skull Bearer," or Bhairava, "the Frightful One," and by which they had to forget the world in intoxication. Their belief was that an ostentatious indifference to anything worldly was the best method of severing the ties of *saṃsāra*.

Vīraśaivas (Liṅgāyats)

The view and way of life peculiar to the Vīraśaivas, or Liṅgāyats (*liṅga*-bearers), in southwestern India is mainly characterized by a deviation from some common Hindu traditions and institutions, such as sacrificial rites, temple worship, pilgrimages, child marriages, and inequality of the sexes. Initiation (*dīkṣā*) is, on the other hand, an obligation laid on every member of the community. It is bestowed upon the newborn and converts, who receive the spiritual power of the guru, the eightfold shield—which includes the guru, *liṅga* (phallic representation of Śiva), *mantra*, sanctified food—to protect the devotee from nescience (ignorance of the supremacy of God) and to guide him to final beatitude, and the *liṅga* phallic symbol. The miniature *liṅga*, the centre and basis of all their religious practices and observances, which they always bear on their body, is God himself concretely represented. Worship is due it twice or three times a day. When a Liṅgāyat "is absorbed into the *liṅga*" (*i.e.*, dies), his body is not, as is customary in Hinduism, cremated but, like those of the ascetics of other groups, interred. Those Liṅgāyats who have reached a certain level of holiness are supposed to die in the state of emancipation.

Śaivism, though in doctrinal matters it is inclined to adoptive inclusivism, inculcates some fundamental lines of conduct: one should worship one's spiritual preceptor (guru) as God himself, follow his path, consider him to be present in oneself, and dissociate oneself from all opinions and practices that are incompatible with the Śaivite creed. Yet some of Śiva's devotees also worship other gods, and the "Śivaization" of various ancient traditions is sometimes rather superficial. Like many other Indian religions, the Śaiva-siddhānta has developed an elaborate system of ethical philosophy, primarily with a view to preparing the way for those who aspire to liberation. Since *dharma* leads to happiness, there is no divorce between sacred and secular duties. Any deed is to be performed as a service to God and in the conviction that all life is sacred and God-centred. A devout way of living and a non-emotional mysticism are thus much recom-

mended. Kashmir Śaivism developed the practice of a simple but most unusual method of salvation: by the recognition (*pratyabhijñā*)—direct, spontaneous, techniqueless, but full of *bhakti*—of one's identity with God.

Kashmir Śaivism

TANTRISM AND SAKTISM

Indian authors have often considered Vedism and Tantrism (Tantra)—a comprehensive term for religious doctrines and practices characterized by a "systematic succession of ceremonies"—complementary concepts. But it is more correct to say that the cult of the Vedic elite was gradually superseded by religious traditions dating from both Aryan and non-Aryan antiquity. Vedic religion, through a long and complicated process of amalgamation, transformation, assimilation, and interpenetration with other traditions, some doubtless pre-Aryan, developed into the extremely variegated complex called Hinduism. Nevertheless, Tantrism, which appears both in Buddhism and in Hinduism, is an important component of religion that, though primarily meant for esoteric circles, also influenced and penetrated, from the 5th century AD, many religious trends and movements.

Nature of Tantric tradition. Opinions of what Tantrism is are quite diverse. Generally speaking, Tantrism claims to show in times of supposed religious decadence a new way to the highest goal and bases itself upon mystic speculations concerning divine creative energy (*śakti*). Tantra is essentially a method of conquering transcendent powers and of realizing oneness with the highest principle by yogic and ritual means—in part magical and orgiastic in character—which are also supposed to achieve other supranormal goals.

Tantrics take for granted that all factors in both the macrocosm and the microcosm are closely connected. The adept (*sādhaka*) has to perform the relevant rites on his own body, transforming its normal, chaotic state into a "cosmos." The macrocosm is conceived as a complex system of powers that by means of ritual-psychological techniques can be activated and organized within the individual body of the adept. Contrary to the ascetic emancipation methods of other groups, the Tantrics emphasize the activation and sublimation of the possibilities of their own body, without which salvation is believed to be beyond reach.

Vāmācāra: left-handed Tantra

The Tantrics of the Vāmācāra ("the left-hand practice") sought to intensify their own sense impressions and so made enjoyment or sensuality (*bhoga*) their principal concern: the adept was to pursue his spiritual objective through his natural functions and inclinations, which were to be sublimated and then gratified in rituals in order to disintegrate his normal personality. This implies that cultic life was also largely interiorized and that the whole world, since it became completely ritualized, was given a new and esoteric meaning.

Tantric worship (*pūjā*) is complicated and in many respects different from the conventional ceremonies that, however, it has often influenced. Tantric devotees distinguish between an "external" and an esoteric meaning of their texts and interpret their texts by means of an ambiguous shadow language. Tantrics describe states of consciousness with erotic terminology and describe physiological processes with cosmological terminology. They proceed from "external" to "internal" worship and adore the goddess mentally, offering their hearts as her throne and their self-renunciation as "flowers."

According to Tantra, concentration is intended to evoke an internal image of the deity and to resuscitate the powers inherent in it so that the symbol changes into mental experience. This "symbolic ambiguity" is also much in evidence in the esoteric interpretation of ritual acts performed in connection with images, flowers, and other cult objects and is intended to bring about a transfiguration in the mind of the adept.

The *mantra*s (one-word spells, such as *hūṃ*, *hrīṃ*, and *klaṃ*) are an indispensable means of entering into contact with the power they bear and of transcending normal mundane existence. Most potent are the monosyllabic, fundamental, so-called *bīja* ("seed") *mantra*s, which constitute the main element of longer formulas and embody the essence of divine power as the eternal, indestructible prototypes from which anything phenomenal derives its existence. Even the cosmos itself owes its very structure and harmony to them. Also important is the introduction, through imposition of a finger (accompanied by a *mantra*), of spiritual qualities or divine power into the body (*nyāsa*).

Those Tantrics who follow the "right-hand path" attach much value to the *yoga* that developed under their influence and to *bhakti* and aspire to union with the Supreme by emotional-dynamic means, their *yoga* being a self-abnegation in order to reach a state of ecstatic blissfulness in which the passive soul is lifted up by the divine grace.

There is also a Tantric Mantra Yoga (discipline through spells), which operates with formulas, and a Haṭha Yoga, "[a method of] reintegration through force." Besides normal yogic practices—abstinences, observances, bodily postures, breath control that requires intensive training, withdrawal of the mind from external objects, concentration, contemplation, and identification that are technically helped by *mudrās* (*i.e.*, ritual intertwining of fingers or gestures expressing the metaphysical aspect of the ceremonies or the transformation effected by the *mantra*s) and muscular contractions—Haṭha Yoga consists of internal purifications (*e.g.*, washing out stomach and bowels), shaking the abdomen, and some forms of self-torture. The whole process is intended to "control the 'gross body' in order to free the 'subtle body.' "

Mudrās

Some Tantrics also employ Laya Yoga ("reintegration by mergence"), in which the female nature-energy (representing the Śakti), which is said to remain dormant and coiled in the form of a serpent (*kuṇḍalinī*) representing the uncreated, is awakened and made to rise through the six centres (*cakra*s) of the body, which are located along the central artery of the subtle body, from the root centre to the lotus of a thousand petals at the top of the head, where it merges into the Puruṣa, the male Supreme Being. As soon as the union of Śakti and Puruṣa has become permanent, according to this doctrine, wonderful visions and powers come to the adept, who then is emancipated. Some of the Tantric texts also pursue worldly objectives verging on magic or pseudo-medicine.

Tantric and Śākta views of nature, man, and the sacred. The Tantric movement is not rarely inextricably interwoven with Śāktism (Shāktism, Śākta). Śāktism consists of doctrines and practices that assume one or more *śakti*s, "creative energies," that are inherent in and proceed from God and are also capable of being imagined as female deities. Śakti is the deciding factor in the salvation of the individuals and in the processes of the universe because God acts only through his energy, which, personified as a goddess, is his spouse. Her role is very different in the various systems: she may be the central figure in a philosophically established doctrine, the dynamic aspect of Brahman, producing the universe through her *māyā* (mysterious power); a capricious demoniac ruler of nature in its destructive aspects; a benign Mother Goddess; or the queen of a celestial court. There is a comprehensive Śāktism that identifies the goddess (usually Durgā) with Brahman and worships her as the ruler of the universe by virtue of whom even Śiva exists. As Mahāyoginī (Great Mistress of Yoga), she produces, maintains, and reabsorbs the world. As the eternal Mother, she is exalted in the *Devīmāhātmya* ("glorification of the goddess") section of the *Mārkaṇḍeya-Purāṇa* (an important medieval Śākta encyclopaedic text). In the Bengal cult of the ogress Kālī, she demands bloody sacrifices lest her creative potency fail her; this cult propounds the belief that birth and death are inseparable, that joy and grief spring from the same source, that the frightening manifestations of the divine should be faced calmly.

Śakti

Whatever non-Aryan roots of Śāktism there may have been, the idea of a bisexual Primeval Being was also well known in the Veda. The goddess Vāc (her name means "Word") was then already the energetic and productive partner of Prajāpati; attention has already been drawn to



Various roles of sakti, the female aspect of the divine.
(Left) Pārvatī, the beneficent mother, bronze statue, south Indian, *c.* 900. Formerly in the collection of Srinivasa Gopalachari. (Centre) Durgā, the destroyer, detail from a Basohli school painting, *c.* 1700. In the Cleveland Museum of Art, Ohio. (Right) Ardhanārīsvara, united with lord Śiva as half-male, half-female, sandstone sculpture from Jhālawār, 6th century. In the Jhālawār Archaeology Museum, Rājasthān, India.

(Left and right) Pramod Chandra, (centre) by courtesy of the Cleveland Museum of Art, Ohio, Mr. and Mrs. William H. Marlatt Fund

the part played by the goddess Virāj in Puruṣa's becoming the world through a sacrifice. As Ardhanārīśvara (the "Lord Who Is Half Female"), Śiva presides over procreation. The Śāktas—often markedly associated with Śaivism—drew these conclusions: creation is the result of the eternal lust of the divine couple; the man who is blissfully embraced by a beloved woman who is Pārvatī's counterpart assumes Śiva's wonderful personality and, liberated, will continue the joy of amorous sport.

Vaiṣṇava Śāktism

In all of his incarnations Viṣṇu is united with his consort, Lakṣmī (Lakshmī). The sacred tales of his various relations with her manifestations led his worshippers to view human devotion as divine and hence as universal, eternal, and sanctified. In Vaiṣṇava Tantrism, Lakṣmi plays an important part as God's *śakti*—that is, as a central metaphysical principle. In his supreme state, Viṣṇu and his *śakti* are indissolubly associated with one another so as to constitute the personal supreme Brahman, also called Lakṣmī-Nārāyaṇa. In mythical imagery, Lakṣmī never leaves Viṣṇu's bosom. In the first stage of creation, she awakes in her dual aspect of action-and-becoming, in which she is the instrumental and material cause of the universe; Viṣṇu himself is the efficient cause. In the second stage, her "becoming" aspect is manifested in the grosser forms of the souls and the *māyā* power, which is the immaterial source of the universe. In displaying her power she takes into consideration the accumulated *karman* of the beings, judging mundane existence as merit and demerit. Presented in myth as God's wife and the queen of the universe, she is always intent on liberating, by her favour and compassion, the incarnated souls of the devout; that is, she allows them to re-enter into herself because they are really "parts," or rather "contractions," of her own essence. After entering her, the liberated soul takes part in the perfect embrace of the divine couple. Pāñcarātra Vaiṣṇavism emphasizes that Lakṣmī—who in the mythological sphere intercedes with her husband for the preservation of the world—spontaneously and by virtue of her own power differentiates herself from Viṣṇu because she has in view the liberation of the souls. This current of thought complicated its explanation of the relation between God and the universe—which was also at the same time an attempt at assigning to God's manifestations a place in a harmonious theological and cosmological system—by an evolutionist theory of successive creations. God is assumed to manifest himself also in three other figures, mythologically his brothers, who, each with his own responsibility, have not only a creative but also an ethical function, by which they assist those who seek to attain to final emancipation.

Śākta ritual and magical practices. The cult of the Śāktas, though often deteriorated, is based on the principle of the ritual sublimation of natural impulses to maintain and reproduce life. Śākta adepts are trained to direct all their energies toward the conquest of the Eternal. The ritual satisfaction of lust, consumption of consecrated meat or liquor, and sexual intercourse are all esoterically significant means of realizing the unity of flesh and spirit, of the human and the divine. They are not sinful acts but, on the contrary, effective means of salvation. Ritual copulation—which may also be accomplished symbolically—is, for both partners, a form of sacralization, the act being a participation in cosmic and divine processes. The experience of transcending space and time, of surpassing phenomenal duality, of recovering the primeval unity, the realization of the identity of God and his Śakti, and of the manifested and unmanifested aspects of the All, constitute the very mystery of Śāktism. The interpretation—metaphorical or literal—of the doctrines is, however, largely a matter of opinion and practice. Ritual practice is indeed as varied as the doctrines. Extreme Śākta communities perform the secret nocturnal rites of the *Śrīcakra* ("wheel of fortune"), described in the *Kulārṇava Tantra*, a leading text, and avail themselves of the natural and esoteric symbolic properties of colours, sounds, and perfumes to intensify their sexual experiences.

After elaborate purifications, the worshippers—who must be initiated, full of devotion toward the guru and God, have control over themselves, be well prepared and pure of heart, know the mysteries of the scriptures, and look forward to the adoration with eagerness—make the prescribed offerings, worshipping the mighty puissance of the Divine Mother, reciting the relevant *mantras*, and identifying themselves with their "soul." Once they have become aware of their own state of divinity, they are qualified to commune with the goddess. *Yoga* and worship (*pūjā*), individually and collectively, daily, fortnightly, and monthly, "for the delectation of the deity," are of special importance. If a woman or a girl is, in certain rituals, made the object of worship, the goddess is first invoked into her; the worshipper is not to cohabit with her until his mind is free from impurity and he has risen to divine status. A relationship with a low-caste woman helps to transcend all opposites; that with a woman who

belongs to another man is often preferred because it is harder to obtain, nothing is certain in it, and the longing stemming from the separation of lover and beloved is more intense—it is pure *preman* (agape, or divine love); adoration of a girl of 16 aims at securing the completeness and perfection of which this number is said to be the expression. The texts however reiterate how dangerous these rites are for those who are not initiated; those who fail and perform such ritual acts without merging their minds in the Supreme are likely to go to one of the hells.

Bengali Vaiṣṇava-Sahajīya cult

The esoteric Vaiṣṇava-Sahajīya cult, which arose in Bengal in the 16th century, was another emotional attempt at reconciling the spirit and the flesh. Displaying contempt for social opinion, its adherents, using the natural (*sahaja*) qualities of the senses and stressing the sexual symbolism of Bengal Vaiṣṇavism, reinterpreted the Rādhā–Kṛṣṇa legend and sought for the perpetual experience of divine joy: since Kṛṣṇa's nature is love and the giving of love and man is identical with Kṛṣṇa, the realization of love is, after an arduous training, to be experienced in man's nature. Women, being a ritual necessity as well as the embodiment of a theological principle, could even become spiritual guides, like Rādhā, conducting the worshipper in his search for realization. After reaching this state, he remains in eternal bliss, can dispense with guru and ritual, and be completely indifferent to the world, "steadfast amidst the dance of *māyā*."

Tantric and Śākta ethical and social doctrines. The ethical and social principles, though fundamentally the same as those promulgated in the classical *dharma* works, breathe a spirit of liberality: much value is set upon family life and respect for women (the image of the goddess); no ban is placed on travelling (conventionally regarded as bringing about ritual pollution) or on the remarriage of widows. Although Tantric and Śākta tradition did not oblige its followers to deviate in a socially visible way from the established order, it provided a ritual and a way of life for those who, because of sex or caste, could not participate satisfyingly in the conventional rites.

FOLK HINDUISM

In spite of the impact of the West, the propaganda of modern reform movements, and the spread of education and secularist modernization, the Hinduism of the masses is changing only slowly. For the ordinary Hindu, religion consists primarily in the manual and verbal performance of rites to promote his private interests. To him, the innumerable ceremonies, observances, fasts, feasts, pilgrimages, and visits to nearby temples constitute the essence of religion.

General characteristics of folk traditions. For millions, the main motive of religious practices is still the fear of ambivalent powerful beings, which infuse their minds with vague and primitive "theological" ideas. The lower castes often confine themselves to propitiating the meat-eating, sometimes benevolent but largely malevolent deities concerned with man's daily events, their ancestors or the founder of their community, and those various spirits that, causing evil and misfortune, have no permanent residence. These castes are content to escape the powers of the evil eye; to manipulate those spirits dwelling in wells, trees, stones, water, and ground; to counteract curses, witchcraft, plague, and cholera; and to worship village godlings who may give rain or a bountiful harvest. They believe in astrology, horoscopy, divination, and the reading of omens and auspicious moments. A large variety of purifications and ritual prohibitions, charms, and amulets to ward off any kind of misfortune (including bad luck in lawsuits and examinations) are, in the eyes of the majority, of greater importance than the Ātman-Brahman doctrine. Even the hope of heaven or the fear of hell has little vogue in various regions except among the higher castes.

It is difficult to draw a sharp line of distinction between popular Hinduism, the beliefs and practices of more or less Hinduized "external" groups, and Indian tribal religion. Many elements of tribal culture that in a definite region have not been adopted by those who are recognized as belonging to the Hindu fold are similar to what has been adopted. Tribal people and outcaste groups are, on the other hand, always willing to worship a few more gods or to imitate the rituals of lower caste Hindus if they can expect to gain some social or material advantages from so doing. Age-long processes of interpenetration and fusion have led to an adoption of many local and popular cults into general Hinduism—or, because it expressed itself mainly in Sanskrit, into the Great, or Sanskritic, tradition—and to the identification of regional gods with the great figures of the Hindu pantheon. Popular belief is integrated and refined rather than discounted or discarded. This process is facilitated by a tendency toward assimilation of local beliefs by pan-Indian Hinduism and by an unwillingness to deny the gods and cults of the lower groups (the worship of a local river deity, for instance, may be identified with that of the Ganges). The inheritors of the Little, or regional, traditions—which are essentially popular, pragmatic, and without a philosophical foundation and transcendent ideals—accepted, as a result of continual and complicated Hinduizing influences, vegetarianism, regular fasts, and food restrictions and began to object to the remarriage of widows, to observe Hindu festivals, to sing Hindu religious songs, to perform funeral and other ceremonious worship, and last, but most importantly, to imbibe the ideas embodied in religious and mythological narratives. Thus, various tribal or outcaste groups have a religion "anticipating" Hinduism, with some affinities to a simple Śaivism without sacred books or regular liturgy.

Relation of folk Hinduism to Indian tribal religions

While the higher caste Hindu pursues his approach to the divine individually, corporate worship in families, villages, and "sects" is far more common among the masses. These groups exhibit the utmost variation in beliefs and practices. As a rule, each community selects only a small segment from the whole spectrum of religious behaviour as the expression of its own religious life. As to the relation between religion and social structure, there are in many communities differently structured systems, each with its own religious behaviour, in which their members may be involved. As members of a joint family, they have to take part in the domestic cult and ritually to express family solidarity at such critical points as mourning or marriage; as members of a village community, they are expected to take part in its particular cult, which is a collective action of that community. Different castes, however, establish, also in respect of ritual, their own unity differentiating themselves from others. There is, on the other hand, ritual cooperation between different villages of the same region.

For many communities, spiritual reality is complex: while many women may address local spirits, family ancestors, and disease goddesses, some of the men may adhere to monotheistic ideas. The village's guardian spirit and the saint of a Muslim shrine may also be worshipped, Rāma's name is invoked in prayers, and the major deities are honoured chiefly during their periodic festivals. Marriage and other ceremonies combine ancient "Sanskritic" rites with popular and local features, and even members of the higher classes may accept a range of belief from the lowest to the highest. In many regions, each caste has both general Hindu and "parochial" rituals and beliefs, but the proportions in which the two are found together vary from caste to caste and from locality to locality. The upper castes everywhere, however, have a certain amount of "Sanskritic" ritual in common; but even those who are more or less exclusively devoted to, for instance, Śiva under one of his names do not necessarily constitute a Śaiva community. The *bhakti* movements have long since influenced the religious feelings of their followers. Religious problems are discussed by people of any profession or intellectual level. Divine assistance is implored on every imaginable occasion; ancient Vedic rites were recently used as a defense against atomic danger.

Regional varieties of folk religion. In the hilly and mountainous regions of north India, Śaivism, aligned with Śāktism, is prevalent. The awe and mystery of jungle and mountains are, there and elsewhere, personified as

forest "Mothers" or mountain deities, represented by piles of stones or branches of trees to which every passer-by contributes an offering. Mother Earth is a great goddess whose marriage (with the earth god or the sun) is festively celebrated and whose annual period of impurity is observed by a cessation of all agricultural activities. During the harvest season she is propitiated with wild orgies. In secluded parts of central India she is identified with Devī, the goddess mostly worshipped in North India. Very often, however, she shows herself in her malevolent form—*e.g.*, as Mother Death (Mārī) or as Kālī. There are also many lower caste groups that have adopted a Vaiṣṇava way of life either in order to raise their social status or to have a prospect of salvation.

Folk goddesses

In the east and northeast, where, broadly speaking, Śāktism is dominant, though Vaiṣṇavism is far from absent, popular belief has modified the transmigration doctrine by the assumption that a soul of the deceased reappears in a child born in the same family within a year after the person's death. Among the female deities, there are tutelary goddesses of young children and women in childbed: Ṣaṣṭhī, "the Sixth," worshipped on the sixth day after birth and represented by a compost pile of cow dung or earth that is placed in the birth room; and Caṇḍī, a form of the goddess Durgā, who lives in trees and is propitiated by lumps of earth. The snake goddess Manasā is personified in a plant of the same name or a stone rudely carved into the shape of a female seated on a snake; a day in the rainy season, when reptiles are most dangerous, is especially devoted to her priestless and unpretentious worship. In literary works, she is eulogized as the Great Mother who is expected to give a prosperous journey through life and, to a certain extent, "Sanskritized" by being identified with epic snake demons. Another example of fusion of more general and local Hindu institutions is the conviction that ghosts and demons are warded off by performing a ceremony in honour of the deceased at Gayā in modern Bihār state.

In many regions—*e.g.*, in western India, where Vaiṣṇavism is dominant—people admit that virtue will improve one's lot in a subsequent existence but, except for the Brahmins and the educated, seem to have only vague ideas of final union with the Supreme. Here also there is a workaday religion practiced by the masses to meet the requirements of everyday existence as well as a higher religion understood only by the Brahmins, who are called on to officiate on important occasions. In order to discover the divine will, exorcists and mediums, claiming to be possessed by mother goddesses and submitting themselves to self-torture, are called upon to prophesy about future events. In these regions also the worship of snakes—to which even temples are dedicated—is much in prominence. Practices based on the belief in scapegoats, ritual nudity, and black magic are not yet things of the past; but here also modernization has penetrated, especially into places near railways or industrial areas.

The whole of peninsular India is mainly devoted to Śaivism and devotional forms of Vaiṣṇavism. A very striking feature in the religion of south India is the propitiation of usually local female village deities of varied and ambivalent character, to whom in almost every settlement a simple shrine or other sacred place is dedicated. These deities have no relation to the universe and are only competent to deal with the facts of village life, such as diseases of the inhabitants and their cattle. Although there are special cholera and smallpox goddesses, their functions are not clearly marked in comparatively primitive milieus; but wherever higher civilization has penetrated, they are more differentiated and the subject of elaborate stories. In a few cases—*e.g.*, that of Māriyamma, the smallpox goddess of South India—such a goddess is known to a large region. These mothers, from whom all good and bad luck emanates, are almost universally worshipped with animal sacrifices; the priestly ministrants (*pūjārī*) officiating in their cult belong to the non-Brahmin groups. The goddesses may be represented by various symbols (stone pillars, sticks, clay figures) that need not even be permanent. Most of their shrines are mean little brick buildings or rough stone platforms under a tree. Offerings of rice, fruit, and flowers may be made every day or on fixed days; there often is a fixed annual festival but no uniformity and no calendar of festivals. An exception to this is the male deity Aiyaṉār, who in the countryside of Tamil Nadu state is worshipped as the watchman and patron of the villages but who is also implored to grant children and other blessings. He is a vegetarian and therefore ranks as "socially" superior to the female village goddess with whom he has entered into a complementary relation. Aiyaṉār is worshipped either as a village deity or in a temple dedicated to Śiva, where he is given the rank of a son of that god. In these Śiva temples he is legitimated by higher Hinduism and fulfills the function of a double of Śiva representing "All-India" or general Hinduism in the village, which does not regard him as an outsider. Śiva himself is also worshipped and given a consort, who, though considered a manifestation of Durgā, has various names according to the tradition of temple or village.

Aiyaṉār

The cult of the sacred snakes, especially cobras, is also practiced—partly to avert danger from these reptiles, partly to propitiate them with the aim of obtaining rain, fertility, or children; to that end women worship snake stones (*nāgalkals*) or erect stone figures of cobras. Every joint family of the Coorgs in Mysore state and most other peoples have a snake deity of their own that is said to embody their welfare. Here and there, Brahmins officiate in this cult, which usually takes place in small sanctuaries in private gardens. Though also known in other parts of India, the methods of exorcising evil spirits known as devil-dancing are most fully developed in south India. The notorious hook-swinging festival, Caḍak-pūjā, held for propitiatory purposes in cases of famine or other calamities—a man was suspended by hooks at the end of a long pole and swung around—though strictly prohibited, survived in this century. Greater festivals are, generally speaking, either celebrated at the chief agricultural seasons or connected with the expulsion of malign powers.

MODERN HINDUISM

The beliefs and customs of the masses are much the same in the present century as they were in the past. Contact with the dynamic West, Christianity, modern life, and technology since the early 19th century have, however, resulted in the emergence of a considerable number of movements and spiritual groups as diverse in their principles, ideals, and reactions to foreign influences as the many remarkable men who founded them. Most such movements distinguish themselves from traditional movements of a devotional variety by more pronouncedly ethical, social, and national preoccupations.

Brahmo Samaj and Arya Samaj. The first product of this cultural encounter, the Brahmo Samaj, founded by the Brahmin Rammohan Ray (1772–1833), aimed at a restoration of a monotheistic and Upaniṣadic Hinduism, purged of abuses and idolatry, which, expressing eternal truth in a comparatively undiluted form, might become the foundation of a universal religion. Though its appeal was to an elite, it stimulated many Hindus into a reflection on the postulates of their view of life. Some schisms, resulting mainly from the activities of Keshab Chunder Sen (1838–84)—in whom *bhakti*, mysticism, and asceticism were stronger than rational theism—led to the coexistence of some small groups, mainly in Bengal, whose ideals range from radicalism (abolition of castes and the remodelling of rituals) to the introduction of Christian elements and institutions (welfare work) or to philosophical reformation of points of ancient Hindu belief.

The work of Rammohan Ray and Keshab Chunder Sen

After the Indian Mutiny (1857), anti-Muslim and anti-Occidental ideas began to spread, and a religious nationalism crystallized, mainly in western and northwestern India, either in movements of reformation and modernization or in the propagation of what was considered the core of traditional Hinduism. The Arya Samaj, founded in 1875 by the militant Dayānanda Sarasvati (1824–83), is a typical example of the former tendency: op-

posing foreign religions as well as Hindu sectarianism and claiming to return to the authentic Vedic tradition, it actually propagates a refined nationalist and democratic Hinduism and a worship of God by means of praise, prayer, and meditation, but without symbols and local cults; emphasis is laid upon the importance of the five-fold daily worship and includes the respect for worthy persons and social activity.

Ramakrishna and Vivekananda: the Ramakrishna Mission. The main object of Ramakrishna (1836–86) was the propagation of Śaṅkara's Vedānta as a superior and comprehensive view of life and a philosophy of meditative experience, synthesizing on a higher level of spiritual consciousness, reached by inner realization, the plurality of human faiths. He proposed not only to prepare the individual devotee for eternity but also to translate his religious ideals into useful social activity. The work of Vivekananda (1862–1902), Ramakrishna's chief disciple, sought to include other religions and philosophies and deliberately turned the trend of Vedānta philosophy toward new values. From these beginnings grew the Ramakrishna Mission and its international "missionary" offshoots, known in Europe and North America as Vedānta societies. Internally, this mission has become an important force for regeneration and unification. Strongly inclined to tolerance and reconciliation, the Ramakrishna Mission has prevented many Indians from being converted to foreign religions. It is the first Indian society to promote its ideas abroad.

Roles of Ramana Maharishi and Rabindranath Tagore. Ramana Maharishi (1879–1950), a southerner, is a typical representative of those holy men who, living in retreat, attempt to achieve the realization of the Vedāntic truth by asceticism. He wrote very little, did not found a system, and communicated his ideas to his disciples, inducing them to search for the Self in themselves.

Although the cosmopolitanism of Rabindranath Tagore (1861–1941) was the utopia of a minority, his foundation, the *āśrama* ("abode of retired ascetics") Santiniketan—which developed into the international Visva-Bharati University—has become a centre of study and inspiration for those who accepted his partly traditional, partly innovatory *bhakti* and mystic view of life and belief in the creative power of Brahman and in the unity of all things and all lives in God.

Political teachers. The intimate relations between religion and social life induced the leaders of the nationalist movements to borrow political lines of action from the religious traditions and to transform ancient religious ideals—*e.g.*, Rāma's perfect rule—into expectations of the future. Ancient doctrines such as the *Bhagavadgītā* were reinterpreted in a political sense, the *avatāra* doctrine becoming a political messianism. Combining with the cult of Kālī and Gaṇeśa—who was made the patron of nationalism—these tendencies even developed into popular movements in Bengal and the Deccan.

The politician B.G. Tilak (1856–1920) professed self-realization through higher meditative knowledge, based on the *Bhagavadgītā* and determining social and political activity. In giving a new content to traditional beliefs and customs. Mahatma Gandhi (1869–1948), who characterized himself as a man of religion in the garb of a politician, far surpassed him. The ancient ideals of *ahiṃsā*, chastity, observances, and *satya* (Truth, which he identified with God) were the main principles of his undogmatic doctrine and social and political practice.

Observances (*vrata*) of restrictions help a man find the energy necessary for the realization of his ideals. By practicing *ahiṃsā* one is believed to acquire an incomparable power that causes the good qualities of one's opponents to triumph over their evil inclinations; asserting truth (*satyāgraha*, the name of his practical politics) implies patience, self-denial, and complete confidence in righteousness and morality. Gandhi was also among those who adapted the ancient institution of *āśramas* so as to suit the new circumstances and to become communities of people bound together by common ideals and activities. Whereas Vinoba Bhave (1895–) propagates, in Gandhi's spirit, the donation of fields and money as a means of self-realization by work (*karma-mārga*), the Hindu-Mahasabha advocates, like other groups of the right, a pure and orthodox anti-Occidental and anti-secularist Hinduism.

Aurobindo's idea of Mother India

One of the most striking personalities was Sri Aurobindo (1872–1950), a native of Calcutta. After studying classical philology at Cambridge University and a career as teacher, poet, publicist, and radical politician, he devoted himself, in his *āśrama* in the then French Pondicherry (now union territory), to spiritual realization, introducing mystical *bhakti* into politics, urging the necessity of a spiritual emancipation of Mother India, whom he identified with the divine Mother, and teaching a *yoga* by which to transform ordinary human beings into divine beings possessing love, wisdom, and power for good and so to achieve a transfiguration of material existence.

Patterns of change. Thus, notwithstanding the presence of both secular ideas and an accelerating process of secularization—which is also unintentionally brought about, for instance, by orthodox Hindu vaccinators making the belief in the goddess of smallpox superfluous—in the general change from predominantly rural to predominantly urban patterns of life and thought, Hinduism has remained an utterly varied complex of religious beliefs and practices, assimilative in adopting Western institutions in a familiar form yet clinging to the essence of its old—mainly Vedāntic—traditions. Vaiṣṇavas, Śaivas, and other religious groups apply modern methods of propaganda, publishing texts and periodicals, founding associations, and convening conferences. Through its best exponents, neo-Hinduism is steadily adapting itself to the exigencies of modern life.

III. Rituals, social practices, and institutions

SACRIFICE AND WORSHIP

Image worship and mental adoration have in later Hinduism taken the place of the great Vedic sacrificial rituals (*yajña*). Nevertheless, much of the comparatively uncomplicated Vedic private rites (*gṛhya*) has survived. The Vedic householder was expected to maintain a domestic fire into which he made his offerings. Normally he did this himself, but in many cases he might also employ a Brahmin officiant. In the course of time, the family priest claimed or was given a large part in these ceremonies, so that most Hindus have been dependent on Brahmins for the administration of the "sacraments" (*saṃskāra*), which, accompanying them from their conception to their cremation, are the main constituents of the domestic rites.

Domestic rites. These sacraments as transitional rites are intended to make a person fit for a certain purpose, for the next stage in his life, by removing taints (sins) or by generating fresh qualities. If the blemishes incurred in this or a previous life are not removed, man is impure and will acquire no reward for his ritual acts. The sacraments, while sanctifying critical moments, are therefore deemed necessary for unfolding man's latent capacities for development. Śūdras are mostly allowed to perform some *saṃskāras* without using Vedic *mantras*.

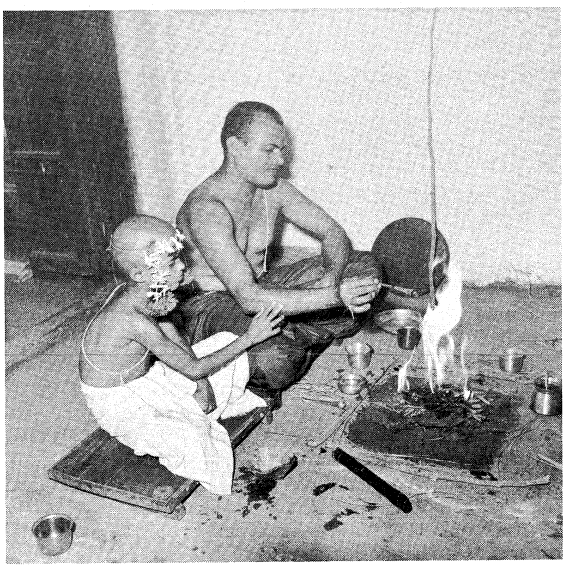
Saṃskāras: passage rites. In antiquity there was a great divergence of opinion about the number of passage rites, but in later times 16 were regarded as the most important. The impregnation rite consecrating the supposed time of conception consists of a ritual meal of pounded rice (mixed "with various other things according as the married man desires a fair, brown or dark son, a learned son or a learned daughter"), an offering of rice boiled in milk, sprinkling of the woman, intercourse; all acts are accompanied by *mantras*. In the third month of pregnancy, the rite called *puṃsavana* (begetting of a son) is to follow. The birth is itself the subject of elaborate ceremonies, the main features of which are an oblation of ghee (clarified butter) cast into the fire; the introduction of a pellet of honey and ghee into the newborn child's mouth, which according to many authorities is an act intended to produce mental and bodily strength; the muttering of *mantras* for the sake of a long life; rites to coun-

Rites associated with the newborn

teract inauspicious influences. There is much divergence of opinion as to the time of the name-giving ceremony; in addition to the personal name, there is often another one that should be kept secret for fear of sinister designs against the child.

In modern times most *saṃskāra*s (except impregnation, initiation, and marriage) have in many areas fallen into disuse or are performed in an abridged or simplified form without Vedic *mantra*s and a priest. This tendency was encouraged by the accommodating spirit of the Brahmins who allowed their clients easy atonements for the nonobservance of rites. The important initiation *upanayana* is held when a boy is between eight and 12 and marks his entry into the community of the three higher classes of society. In this rite he becomes a "twice-born one," or *dvija*. Traditionally, this was also the beginning of a long period of Veda study and education in the house under the guidance of a teacher (guru). Nowadays, the hair-cutting ceremony—formerly performed in a boy's third year—and the initiation are usually performed on the same day, the homecoming ceremony at the end of the period of study being little more than a formality.

McKim Marriott



A young boy performs his first *pūjā* after initiation into the "twice-born."

Marriage customs

Wedding ceremonies, the most important of all, have not only remained elaborate—and often very expensive—but have also incorporated various elements—among others, propitiations and expiations—which are not indicated in the oldest sources. Already in ancient times there existed great divergences in accordance with local customs or family or caste traditions. Yet a small number of practices is usually considered essential. The date is fixed only after careful astrological calculation; the bridegroom is conducted to the home of his future parents-in-law, who receive him as an honoured guest; there are offerings of roasted grain into the fire; the bridegroom has to take hold of the bride's hand; he conducts her around the sacrificial fire; seven steps taken by bride and bridegroom are to solemnize the irrevocability of the unity; both are in procession conducted to their new home, which the bride enters without touching the threshold.

Of eight forms of marriage recognized by the ancient authorities, two have remained in vogue; viz., the simple gift of a girl and the legalization of the alliance by means of a marriage gift paid to the bride's family. In the Vedic period, girls do not seem to have married before they had reached maturity. Child marriage and the condemnation of the remarriage of widows, especially among the higher classes, became customary later and have gradually, since the mid-19th century, lost their stringency.

The traditional funeral method is cremation (a family affair), burial being reserved for those who have not been sufficiently purified by *saṃskāra*s (*i.e.*, children) and those who no longer need the ritual fire to be conveyed to the hereafter, such as ascetics who have renounced all earthly concerns. An important and meritorious complement of the funeral offices is the *śrāddha* ceremony: offerings of food to Brahmins for the benefit of the deceased. Many people are still solicitous to perform this rite at least once a year even when they no longer engage in any of the five obligatory daily offerings.

Death rites and customs

Daily offerings. There are five obligatory offerings: (1) offerings to the gods (food taken from the meal); (2) a cursory offering (*bali*) made to "all beings"; (3) a libation of water mixed with sesame offered to the spirits of the deceased; (4) hospitality; and (5) recitation of the Veda. The conventional Hindu daily worship of the five protecting deities (Viṣṇu, Śiva, Pārvatī, Gaṇeśa, Sūrya), represented by figurines, continues in the present century.

Other private rites. The morning and evening adorations (*saṃdhyā*), being a very important duty of the traditional householder, though mainly Vedic in character, have, by the addition of Purāṇic and Tantric elements, become lengthy rituals. If not shortened, the morning ceremonies consist of self-purification, bathing, prayers, recitation of *mantra*s, especially the "Gāyatrī" *mantra* (Ṛgveda 3.62.10), a prayer for spiritual stimulation addressed to the sun. The accompanying ritual includes (1) the application of marks on the forehead characterizing the adherents of a particular religious community, (2) presentation of offerings (water, flowers) to the sun, and (3) meditative concentration. There are Śaiva and Vaiṣṇava variants, and some elements are optional. The observance of the daily obligations, including the care of bodily purity and professional duties, leads to mundane reward and helps to preserve the state of sanctity required to enter into contact with the divine.

"Gāyatrī" *mantra*

Temple worship. According to Vaiṣṇava authorities, the regular temple worship of God grants the same results as the ancient fire cult; it even has an advantage over it because temple worship may be continued after the death of the founder.

Temples. The erection of a temple, which belongs to whoever paid for it or to the community that occupies it, is a meritorious deed recommended to anyone desirous of heavenly reward. The choice of site, which should be serene and lovely, is determined by astrology and divination as well as by its situation with respect to the dwellings of men: a sanctuary of a benevolent deity should face the village. The construction of temples is, because of its "symbolic" value, described in great detail. There is much diversity in size and artistic value, ranging from little village shrines with crude statuettes to great temple-cities whose boundary walls, pierced by monumental gates (*gopura*), enclose various buildings, courtyards, pools for ceremonial bathing, and sometimes even schools, hospitals, and monasteries. From the point of view of construction there is no striking difference between Śaivite and Vaiṣṇavite sanctuaries, which are easily recognizable by the image or symbols in the centre, the images on the walls, the symbol fixed on the finial (crowning ornament) of the top, and Śiva's bull, Nandi, or Viṣṇu's bird, Garuḍa—theriomorphic duplicate manifestations of each god's nature—before the entrance. Services, which may be held by any qualified member of the community, are neither collective nor carried out at fixed times. Those present experience, as spectators, the fortifying and beneficial influence radiating from the sacred acts. Sometimes the faithful assemble to meditate, to take part in chanting, or to listen to an exposition of doctrine. The *pūjā* (worship) performed in public "for the well-being of the world" is, though sometimes more elaborate, largely identical with that executed for personal interest. There are, on the other hand, many regional differences, even within the same community.

Pūjā. Hindu worship (*pūjā*) consists essentially of an invocation, reception, and entertainment of God as a royal guest. It normally consists of 16 "attendances" (*upacāra*): invocation by which the omnipresent God is

invited to direct his attention to the particular worship; offering of a seat, of water (for washing the feet, for washing the hands, and for rinsing the mouth), of a bath, a garment, a sacred thread, perfumes, flowers, incense, a lamp, food, and homage; and a circumambulation of the image and dismissal by God.

Photo Atlas-Doumic, Paris



A temple *pūjā* to Krsna.

The Pāñcarātra Vaiṣṇavas in south India introduced the songs of the Dravidian poets into their temple cult and regard these poets and their great teachers as incarnations of God and worship their images also. The Śaivites also have songs of their own but were, generally speaking, more open to Tantric elements and to the admission in their cult of dances executed by dancing girls. In both religious groups, there are communities that cling to the traditional Sanskrit *mantra*s and others that also use other languages.

The significance of the temple door

The first phase of worship is the reverential opening of the temple door and the adoration of the powers presiding over it: according to the Vaikhānasa Vaiṣṇavas, the opening of heaven; and to the Śaivites, an act to secure the building's protection. The divine powers carved in the doorjambs promote the process of transmutation without which man cannot even enter into the presence of God, whose image is established in the cella (*garbhagrha*). This image is honoured with gifts, notably flowers, fruit, and perfumes. Small portions of the consecrated food (*prasāda*) are given to visiting worshippers. The offering into the fire (*homa*) of Vedic origin has been retained in nearly all extended *pūjā* ceremonies. The main purpose of the rites is the meditative identification of the worshipper with the divine Presence; the enactment, in a gradual process of development, of the realization of the worshipper's soul and God. The Vaikhānasas distinguish between the transcendent and unanalyzable Brahman and its immanent and analyzable aspect and invoke God to descend out of compassion from the immovable image—the permanent "seat" of the former—into a movable cult image in which he converses with the world, represented by the worshipper. Those denominations (both Śrīvaiṣṇavas and Śaivites) that adopted Tantric practices believe that God comes, during these ceremonies, also out of the worshipper's heart or that the latter's soul leaves his body to reach God's feet in heaven, to descend from there in a new body that is meditatively created.

A remarkable rite of yogic-Tantric origin is, also in other ritual contexts, the transmutation of water into the elixir of life and "immortality" (*amṛta*), the essential element of which is drawn from the spot between the worshipper's eyebrows, which is supposed to be the seat of Śiva's highest aspect.

Śaivas have to transform themselves into Śiva by means of complicated preparatory rites, because, they say, "Śiva alone can worship Śiva." Some authorities also enjoin a mental worship and sacrifice, without which "exterior" rites are rendered senseless. The merit of the performances is often said to be entrusted to God's keeping for the sake of the worshipper. Many Vaiṣṇavas emphasize that *pūjā* is to propitiate God disinterestedly.

SACRED TIMES AND PLACES

Festivals. Hindu festivals are complex combinations of religious ceremonies, semi-ritual spectacles, worship, prayer, lustrations, processions (supposed to set something sacred in motion and to extend their powerfulness over a certain region), music, dances (which by the rhythm of their movement have a compelling force), magical acts—the participants throw fertilizing water or, during the Holī festival, red powder (an aphrodisiac) at each other—eating, drinking, lovemaking, licentiousness, feeding the poor, and other activities of a religious or traditional as well as of a recreational character. Their original functions appear clearly from both ancient literature and anthropological research: they are intended to purify, to avert malicious influences, to renew society, to bridge over critical moments, to stimulate or resuscitate the vital powers of nature (hence the term *utsava*, meaning both a generation of power and a festival). Since such festivals relate to the cyclical life of nature, they are supposed to prevent it from stagnating. These cyclic festivals —which may extend for many days—continue to be celebrated.

Such festivals refresh the mood of the participants, further the consciousness of their own power, and thus help to compensate for their sensations of fear and inferiority concerning the unknown forces of nature. Such mixtures of worship and pleasure require the participation of the whole community and create harmony among its members, even if only some of the total number of participants may now be aware of their original character. There are also innumerable festivities in honour of individual gods, relating to particular temples, villages, and religious communities.

Holī: festival of spring

A very important festival, formerly celebrating Kāma, the god of sexual desire, survives in the Holī, a saturnalia connected with the spring equinox and in western India with the wheat harvest and in its boisterous and licentious form observed by the lower classes. There are local variants: among the Marāṭhās, heroes who died on the battlefield are "danced" by their descendants, sword in hand, until they believe themselves possessed by the spirits of the heroes. In Bengal, swings are made for Kṛṣṇa; in other regions, a bonfire is also essential. The mythical tradition that accounts for the festival describes how young Prahlāda, who, in spite of his father's opposition, persisted in worshipping Viṣṇu, was carried into the fire by the female demon Holikā, the embodiment of evil, who herself was believed to be immune to the ravages of fire. Through Viṣṇu's intervention, Prahlāda emerged unharmed, while Holikā was burned to ashes. The bonfires are intended to commemorate this event or rather to reiterate the triumph of virtue and religion over evil and sacrilege. This explains why objects representing the sickness and impurities of the past year—the new year begins immediately after Holī—are thrown into the bonfire, and it is considered inauspicious not to look at it. Moreover, people pay or forgive debts, reconcile quarrels, and try to rid themselves of the evils, conflicts, and impurities that may have accumulated during the preceding months, translating the central conception of the festival into a justification for dealing with continuing situations in their lives. The New Year festival, according to another Indian calendar, Dīwālī, though celebrated by all classes of society, is traditionally believed to have been given by Viṣṇu to the Vaiśyas (traders, etc.); it takes place in September–October, with worship and ceremonial lights in honour of

New Year festivals

Lakṣmī, the goddess of wealth and good fortune; fireworks to chase away the spirits of the deceased; and gambling, an old ritual custom intended to secure luck for the coming year. The nine-day Durgā festival, or Navarātrī, is, especially in Bengal, splendid homage to Śakti, and in south India, a celebration of Rāma's victory over Rāvaṇa.

Pilgrimages and fairs. Like processions, pilgrimages (*tīrthayātrā*) to holy rivers (*tīrtha*) and other places were already known in Vedic and epic times and are, up to the present day, one of the most remarkable aspects of Indian religious life. Many sections of the *Purāṇas* eulogize temples and the sacredness of places situated in beautiful scenery or wild solitude (especially the Himalayas). Although the whole of India, and especially Kurukṣetra (presumed to be the scene of the great war portrayed in the Mahābhārata) in the northwest, is holy ground offering an opportunity to reach emancipation and although the number of places of pilgrimage of regional significance amounts to many hundreds, some of them (Ayodhyā, Mathurā, Hardwār, Kāśī or Vārānasi [Benares], Kāñcī or Kānchipuran, Avantikā or Ujjain, and Dvāravatī or Dwārkā) have for many centuries possessed exceptional holiness. The reason for such sanctity derives from their location on the bank of a holy river, especially of the Ganges, or from their connection with legendary figures of antiquity who are said to have lived there, or

Paul Popper Ltd.



Pilgrims bathing in the Ganges River at Hardwār, India.

from the alleged local manifestation of a god. Many places are sacred to a specific god; the district of Mathurā, for example, is studded with places of pilgrimage connected with the Kṛṣṇa legends. Visits to holy places may bestow special benefits upon pilgrims; temples or ponds dedicated to Sūrya ("the Sun") are visited in order to recover from leprosy, other places to escape from astrological threats. Pilgrimages to Gayā (Bihār state)—where visitors are escorted around the sacred centres by Brahmin temple priests who maintain certain ritual connections with their clients—are undertaken for the sake of the welfare of deceased ancestors. In most cases, however, the devotee hopes for deliverance from sin or pollution, preservation of his religious merit, rebirth in a heaven, or even emancipation. The last prospect is held out to those who, when death is near, travel to Vārānasi to die near the Ganges.

On special occasions, be they auspicious or, like a solar eclipse, inauspicious, the devout crowds increase enormously. Most important shrines also organize gatherings (*melā*), half fairs, half religious demonstrations. These journeys, which are also undertaken, individually or in groups, in order to discharge a vow or to please the devotee's favourite god, confirm the devotees in their faith, providing them with an opportunity for spiritual retreat or to bring their inner life nearer to a state of perfection. They have contributed much to the spread of religious ideas and the cultural unification of India.

RITUALS AND SOCIAL STATUS

The traditional Hindu expresses his religious beliefs also in his social customs and institutions as well as in his relations with his fellow men. Four social classes (*varṇa*) were distinguished in the oldest period. Notwithstanding a certain openness and fluidity, these classes appear to have been a socio-religious reality. Such is evident from the Puruṣa hymn (Ṛgveda 10. 90), in which the statement that the Brahmin (Brāhmaṇa) was the Puruṣa's mouth, the nobleman (Kṣatriya) his arms, the Vaiśya his thighs, the Śūdra being born from his feet, gives us an idea of their functions and mutual relations.

The four social classes

The Brahmins, whatever their worldly avocations, are by virtue of their birth a perpetual incarnation of the *dharma*, guardians and dispensers of divine power, entitled to teach the Veda, sacrificing for others and accepting gifts and subsistence; the term alms is misleading, and the *dakṣiṇā* offered at the end of a rite to a Brahmin officiant is not a fee but an oblation through which the rite is made complete. Brahmins are, on account of their pre-eminence, the superiority of their origin, their sanctification owing to the *saṃskāra*s (rites of passage), and their observance of restrictive rules, held to be the highest of all human beings. The main duty of the nobility (the Kṣatriyas) is to protect the people, that of the third estate (the Vaiśyas) to tend cattle, to trade, and to cultivate land. Even if a king (theoretically of Kṣatriya descent) was not of noble descent, such an upholder of *dharma* was clothed with divine authority in the eyes of the masses. He was consecrated by means of a complex and highly significant ritual; he was Indra and other gods (*deva*) incarnate; the emblems or paraphernalia of his office represent sovereign authority; the white sunshade of state is, for instance, the residence of Śrī-Lakṣmī, the goddess of fortune. All three higher classes, claiming Aryan descent, had to sacrifice and to study the Veda, although the responsibilities of the Vaiśyas in sacred matters were less onerous.

While this tripartition seems, in the main, to have been inherited from prehistoric times, the fourth class (the Śūdras), whose sole duty it was "to serve meekly" (*Mānava Dharmaśāstra* 1, 91) the other classes, are partly descended from the subjugated non-Aryans, a fact which accounts for their many disabilities and exclusion from religious status. According to Hindu tradition, the Veda should not be studied in their presence, but they may listen to the recitation of epics and *Purāṇas*. They might perform the five main acts of worship (without Vedic *mantra*s) and undertake observances, but even now they have various ceremonies of their own, carried out without Brahminic assistance. Yet a distinction is often made between Śūdras of a more and those of a less pure and correct behaviour and way of living, the former tending to assimilate with higher castes, the latter to rank with the lowest in the social scale, who, often called *cāṇḍālas*, were already at an early date sweepers, bearers of corpses, or charged with other impure occupations. Ritual purity was indeed an important criterion; "impure" conduct and neglect of Veda study and the rules regarding forbidden food might suffice to stigmatize a "twice-born man" as a Śūdra. On the other hand, in later times the trend of many communities has been toward integrating all Śūdras into the Brahminic system. The Brahmins, who have far into modern times remained, on the whole, a respected, traditional, and not rarely intellectual upper class, were generally until the 1930s much in demand because of their knowledge of rites and traditions. Although Kṣatriya rank is claimed by many whose title is one of function or creation rather than of inheritance, this class is in many regions nowadays numerically insignificant. Moreover, none of the four *varṇa*s represented for a considerable time anything but hierarchically arranged groups of castes.

Castes. The origin of the castes (*jāti*, literally "birth"), the most conspicuous characteristic of Hindu society, cannot be explained from a simple splitting up of the *varṇas;* nor, in accordance with the Indian tradition, from the mixture of these through intermarriage, considered an infraction of the *dharma;* nor, in accordance with some one-sided modern theories, from exclusive commensality deriving from family worship, racial, or ritual differences, occupational differentiation, taboo, and totemism. This is not to deny that some of these factors contributed to the growth and extension of the highly complex division of the Hindu society into nearly 3,000 castes and subcastes—only small parts of which are actually coexistent in the same region—a division that, with the sanction of religion and tradition, survives, especially in South India, though tending to mitigation, up to modern times.

A caste is, in general, an endogamous hereditary group of families, bearing a common name; often claiming a common descent; as a rule professing to follow the same hereditary calling; clinging to the same customs, especially regarding purity, meals, and marriages; and very often further divided into smaller endogamous circles. Moreover, tribes, "guilds," or religious communities characterized by particular customs—for instance, the Vīraśaivas—could easily be regarded as castes. The status of castes is variable in different localities, and social mobility is possible, but their mutual relation is hierarchically determined; local Brahmin groups occupy the highest place, and differences in ritual purity are the main criterion of position in the hierarchy. Most impure are the outcastes, or, to use modern names, the exterior or scheduled castes, which, however, have among themselves numerous divisions, each of which regards itself as superior to others.

Traditional Hindus are inclined to emphasize that the ritual impurity and "untouchability" inherent in these groups does not essentially differ from that temporarily proper to mourners or menstruating women. This, and the fact that some exterior group or other might rise in estimation and become an interior one, or that individual outcastes might be well-to-do, does not alter the fact that the spirit of exclusiveness was in the course of time carried to extremes. The scheduled castes were subjected to various socio-religious disabilities before mitigating tendencies made progress; social discrimination was, after independence, prohibited, and the practice of untouchability made a punishable offense (which does not, however, mean abolition). Scheduled castes were barred from use of temples and other religious institutions and from public utilities such as tanks and schools. These groups also had many disabilities in relations with private persons. From the traditional Hindu point of view, this social system is the necessary complement of the principles of *dharma*, *karman*, and *saṃsāra*. Like hells and heavenly regions in the hereafter, the castes are the mundane, social frame within which *karman* is settled. A low social status is the inevitable result of sins in a former life but can, by virtue and merit, be followed by a better position in the next existence (see CASTE SYSTEMS).

Religious orders and holy men. Those members of the various denominations who are inclined to abandon all worldly attachment assume the character of an "inner circle" or "order" that, seeking a life of devotion, adopts or develops particular vows and observances, a common cult, and some form of initiation.

Initiation. Though Hindus, generally speaking, are free to join an order or inner circle, once they have joined it they must submit to its rites and way of living. The initiation (*dīkṣā*), a sort of purification or consecration involving a transformation of the aspirant's personality, is regarded as a complement to, or even a substitute for, a previous initiation ceremony that it strikingly resembles. Such religious groups integrate ancient, widespread ideas and customs of initiation into the framework of either the Vaiṣṇava or Śaiva patterns of Hinduism. Vaiṣṇavism emphasizes their character as an introduction to a life of devotion, as an entrance into closer contact with God, although happiness, knowledge, a long life, and a prospect of freedom from *karman* are also among the ideals aspired to by them. Śaivism is convinced of the absolute necessity of initiation for anyone desirous of final liberation and requires an initiation in accordance with their rituals. All communities are agreed that the authority to initiate belongs only to a qualified spiritual guide (guru), usually a Brahmin, who has previously received the special *gurū-dīkṣā* (initiation as a teacher) and is often regarded as representing God himself. The postulant is sometimes committed to a probationary period, to training in *yoga* mysticism, or to instruction in the esoteric meaning of the scriptures. The initiate receives a devotional name and is given the distinctive *mantra*s of the community, which, since they are sacred, must never be misused.

Dīkṣā

There are numerous, complicated forms of initiation: the Vaiṣṇavas differentiate between the members of the four classes; the Śaivas and Tantrics take account of the natural aptitude and competency of the recipients and distinguish between first-grade initiates, who obtain access to God, and higher grade initiates, who remain in a state of holiness.

Yoga. The initiate guided by his guru may apply himself to *yoga* (a "methodic exertion" of body and mind) in order to attain, through mortification, concentration, and meditation, a higher state of consciousness in which he may find the supreme knowledge, achieve spiritual autonomy, and realize his oneness with the Highest (or however the ultimate goal is conceived). *Yoga* may be atheistic or combined with various philosophical or religious currents. Every denomination attempted to implement yogic practices on a theoretical basis derived from its own teachings. There are many different forms of *yoga*, the practices vary according to the stage of advancement of the adepts. All serious *yogins*, however, agree in disapproving use of yogic methods for worldly purposes.

Āśramas: the four stages of life. In the West, the so-called life-negating aspects of Hinduism have often been overemphasized. The polarity of asceticism and sensuality, which assumed the form of a conflict between the aspiration to liberation and the heartfelt desire to have descendants and continue earthly life, manifested itself in Hindu social life as the tension between the different goals and stages of life. The relative value of an active life and the performance of meritorious works (*pravṛtti*) as over against renunciation of all worldly interests and activity (*nivṛtti*) was a much-debated issue. While one-sidedly religious and philosophical works such as the *Upaniṣads* place emphasis on renunciation, the *dharma* texts argued that the householder who keeps up his sacral fire, procreates children, and performs his ritual duties well earns religious merit. They also came, nearly 2,000 years ago, to elaborate the social doctrine of the four *āśrama*s ("stages of life"). This concept is a typically Brahmanic and legalizing attempt at harmonizing these conflicting tendencies in one system. A member of the three higher classes should first become a chaste student (*brahmacārin*); then a married householder (*gṛhastha*), discharging his debts to his ancestors by begetting sons and to the gods by sacrificing; then (as a *vānaprastha*) retire, with or without his wife, to the forest to devote himself to spiritual contemplation; and finally, but not mandatorily, become a homeless wandering ascetic (*sannyāsin*). The situation of the married ascetic was always a delicate compromise remaining problematic on the mythological level and often omitted or rejected in practical life.

Although the status of a householder was often extolled, and some authorities, regarding studentship as a mere preparation, went so far as to brand the other stages as inferior, there were always men who became wandering ascetics immediately after studentship. Theorists were inclined to reconcile the divergent views and practices by allowing the latter course of life to those who are, owing to the effects of restrained conduct in former lives, entirely free from worldly desire, even if they had not gone through the traditional prior stages.

Sectarian symbols. The typically Hindu ascetic (*sādhu*) usually wears some distinctive mark (*puṇḍra*) on his forehead and often carries some "symbol" of his religion.

Symbols of Viṣṇu and Śiva

If he is a Vaiṣṇava he may possess a discus (*cakra*) and a conch shell (*saṅkha*), replicas of Viṣṇu's flaming weapon and his instrument of beneficent power and omnipresent protection; a *śālagrāma* stone or a *tulasī* plant, which represent both Viṣṇu's own essence and that of his spouse Lakṣmī. If he is a Śaiva, he will impersonate Śiva himself and carry a trident (*triśūla*), denoting empire and the irresistible force of transcendental reality; wear a small *liṅga;* carry a human skull, showing that he is beyond the terror inspired by the transitoriness of the world; and smear his body with apotropaic (supposed to avert evil) and consecratory ashes. These emblems are sacred because the divine Presence, when invoked by *mantras*, is felt to be in them, which makes them a means of worship.

The attitude toward asceticism has always been ambivalent. On the one hand, there is a genuine regard for hermits and wandering ascetics and a desire to store up spiritual merit through feeding religious mendicants. On the other hand, the fact that fringe members of society may find a sort of respectable status among Śaiva ascetics often led to a decline in the moral reputation of the latter.

Role of Śaṅkara. In ancient times, religious communities were somewhat isolated and scattered throughout India. However, after Śaṅkara (8th century), as tradition has it, had founded and organized a Śaivite order composed of ten brotherhoods and hence called Daśanāmīs ("Those with Ten Names"), orders became an established institution with wider geographic affiliations. Some of these admitted Brahmins only; others were open to all four classes or even to women; some made a practice of nudity. Śaṅkara is also credited with having founded four great monasteries (*maṭha*), governed by the head of Śṛṅgeri monastery in Mysore, who enjoys great spiritual authority. These Śaiva communities are more inclined to individual asceticism and are less closely organized than the Vaiṣṇava Vairāgins ("the Dispassionate") or Gosvāmins ("the Masters").

IV. Cultural expressions of Hindu values and ideas

The structure of Indian temples, the outward form of images, and indeed the very character of Indian art are largely determined by religion and the traditional view of the world, which penetrated the other provinces of culture and welded them into a homogeneous whole.

SYMBOLS AND IMAGES

Indian art is highly symbolic, a "symbol" being understood as the representation or manifestation of an invisible idea and the relation between a sign and an idea being understood as one of community or unity. The much-developed ritual-religious symbolism presupposes the existence of a spiritual reality that, while being in constant touch with phenomenal reality, may make its presence and influence felt and can also be approached through the symbols that belong to both spheres.

Mental and external fashioning of images

The production of objects of symbolic value is therefore more than a technique. The artisan has to model a cult image after the ideal prototype that, with its canonical forms, appears in his mind only when he has brought himself to a state of supranormal consciousness. Undergoing a process of spiritual transformation himself, he also transforms the material of which the image is to be made into a receptacle of divine power. Like the artisan, the worshipper (*sādhaka*, "the one who wishes to attain his goal") must grasp the esoteric meaning of a statue, picture, or flowerpot and identify himself with the power residing in it. The usual offering, a handful of flowers, is a vehicle to convey his "life-breath" into the external image, which has already been transformed into an adequate internal vision of the same divine power.

Types of symbols. If he knows how to handle the symbols, the worshipper—who must achieve his object himself and cannot come into contact with God unless he insistently invokes him—has the disposition of an instrument for utilizing the possibilities lying in the depths of his own subconsciousness as well as a key to the mysteries of the forces dominating our world.

Yantra and maṇḍala. The general term for such an "instrument [for controlling]" is *yantra*, which, while denoting in a wider sense cult images, pictures, and other such aids to worship, is often especially applied to ritual diagrams. Any *yantra* represents some aspect of the divine and enables the devotee to worship it immediately within his heart while identifying himself with it. Except in its greater linear complication, a *maṇḍala* does not differ from a *yantra*, drawn with a highly complex ritual upon the ground on a purified and ritually consecrated place. The meaning and the use of both are similar, and both may be permanent or provisional. A *maṇḍala*, delineating a consecrated place and protecting it against disintegrating forces represented in demoniac cycles, is the geometric projection of the universe, spatially and temporally reduced to its essential plan. It represents in a schematic form the whole drama of disintegration and reintegration, and the adept standing in its centre identifies himself with the forces governing these. Just as in temple ritual, a vase is employed to bring down the divine power so that it may be projected into the drawing and into the person of the adept. Thus the *maṇḍala* becomes a support for meditation, an instrument to provoke visions of the unseen. A good example of a *maṇḍala* is the *śrī-cakra*, "the Wheel of Śrī" (*i.e.*, of God's Śakti) made up of four isosceles triangles with the apices upward symbolizing Śiva and five isosceles triangles with the apices downward symbolizing Śakti; the nine triangles are of various sizes and intersect with one another. In the middle is the power point (*bindu*), visualizing the highest, the invisible, elusive centre from which the entire figure and the cosmos expand. The triangles are enclosed by two rows of (eight and 16) petals, representing the lotus of creation and reproductive vital force. The broken lines of the outer frame make the figure a sanctuary, with four openings to the regions of the universe. A *yantra*, called *maṇḍala*, of the Puruṣa (spirit) of the site, is, as a "spiritual" foundation, also drawn on the site on which a temple is built. This rite is motivated by a replica of the myth of Puruṣa—an immortal primeval being who with his bulk obstructed both worlds until subdued by the gods; the parts of his body becoming site spirits.

Śrīcakra-maṇḍala

Liṅga and yoni. One of the most common objects of worship, whether in temples or in the household cult, is the *liṅga* (phallus). Often much stylized and an austere rather than obscene symbol, erect and representing the cosmic pillar, it emanates its all-producing energy to the four quarters. As the symbol of male creative energy it is frequently combined with its female counterpart (*yoni*), the latter forming the base from which the *liṅga* rises. Although the *liṅga* originally may have had no relation to Śiva, it has from ancient times been regarded as symbolizing Śiva's creative energy and is widely worshipped as his "fundamental form."

Visual theology in icons. The beauty of the cult objects contributes to their force as sacred instruments; their ornamentation facilitates the process of inviting the divine power into them. Statues of gods, however, are not intended to imitate ideal human forms but to express the supernatural. A divine figure is a "likeness" (*pratimā*), a temporary benevolent or terrifying expression of some particular aspect of a god's nature. Iconographic handbooks attach great importance to the ideology behind images and reveal, for instance, that Viṣṇu's eight arms stand for the four cardinal and intermediate points of the compass and that his four faces, illustrating the concept of God's fourfoldness, typify his strength, knowledge, lordship, and potency. The emblems express the qualities of their bearers—*e.g.*, a deadly weapon pointing to destructive force, many-headedness to omniscience. Much use is made of gestures (*mudrā*), conventional devices for denoting activities that express an idea; thus, the raised right hand, in the "fear-not" gesture (*abhaya-mudrā*), bestows protection. Every iconographic detail has its own symbolic value, helping the devotee to direct his energy

Purposes of Hindu iconography

to a deeper understanding of the various aspects of the divine and to proceed from external to internal worship. For many Indians, an installed and consecrated image becomes a container of concentrated divine energy; according to Hindu theists, an instrument for ennobling the worshipper who realizes God's Presence in it.

VISUAL ARTS

Religious principles in sculpture and painting. The dance executed by Śiva as king of dancers (Naṭarāja), the visible symbol of the rhythm of the universe, represents God's five activities: he unfolds the universe out of the drum held in one of his right hands; he preserves it by uplifting his other right hand in *abhaya-mudrā* (a gesture indicating "do not be afraid"); he reabsorbs it by the upper left hand, which bears a tongue of flame; his transcendental essence is hidden behind the garb of apparitions and grace is bestowed and release made visible by the foot that is held aloft and to which the hands are made to point. The other foot, planted on the ground, gives an abode to the tired souls struggling in *saṃsāra*. Another dancing pose adopted by Śiva is the famous *tāṇḍava*, executed in his destructive Bhairava manifestation, usually with ten arms and accompanied by Devī and demons. The related myth is that Śiva conquered a mighty elephant demon whom he forced to dance until he fell down dead; then, wrapped in the blood-dripping skin of his victim, the god executed a horrendous dance of victory.

Images sustain the presence of the god: when Devī is shown advancing against the buffalo demon, seated on her lion, she represents the affirmative forces of the universe and the triumph of divine power over wickedness. Male and female figures in uninterrupted embrace, as in Śaiva iconography, signify the union of opposites and the eternal process of generation. Lovers sculptured on temples are "auspicious symbols" put on a par with foliage, water jars, and other representatives of fertility.

Religious organization of sacred architecture. Temples must be erected on a site that is *śubha* (*i.e.*, suitable, beautiful, and auspicious), in the neighbourhood of water, because the gods will not come to other places. Temples are not, however, designed to be congenial to their surroundings because a manifestation of the sacred is an irruption, a break in phenomenal continuity. Since temples are said to constitute an opening in the upward direction and thus ensure communication with the gods, they are visible representations of a cosmic pillar and their site is said to be a navel of the world. Their outward appearance must raise the expectation of meeting with God. Their erection is considered a reconstruction and reintegration of Puruṣa-Prajāpati, enabling him to continue his creative activity, and the finished monuments are "symbols" of the universe that is the unfolded One. The owner (*i.e.*, an individual or community that paid for its construction, and the descendants) of the temple—also called the "sacrificer"—participates in the process of reintegration and experiences his spiritual rebirth in the small cella, aptly called the "womb room" (*garbhagṛha*), by means of meditative contact with God's presence, symbolized or actualized in his consecrated image. The cella is in the centre of the temple above the "navel"—*i.e.*, the "foundation stone"—a jar filled with the creative power (*śakti*) that is identified with the goddess Earth (who bears and protects the monument) three lotus flowers, and three tortoises (of stone, silver, gold) that represent earth, atmosphere, and heaven. The tortoise is regarded as a manifestation of Viṣṇu bearing the cosmic pillar, and the lotus as a symbol of the expansion of generative possibilities. The vertical axis or tube (coinciding with the cosmic pillar), which connects all parts of the building and is continued in the finial on the top, corresponds with the mystical vertical "vein" in the body of the worshipper through which his soul rises to unite itself with the Highest.

Requirements of temple architecture and planning

THEATRE AND DANCE

Theatrical performances of various types and artistic levels are also events to secure blessings and happiness; the element of recreation is indissolubly blended with edification and spiritual elevation. The structure and character of the classical Indian drama reveal its origin and function; it developed from the last part of a magico-religious ceremony that survives as a ritual introduction and begins and closes with benedictions. Drama is produced on festive occasions with a view to spiritual and religious success (*siddhi*), which must also be prompted by appropriate behaviour from the spectators; there must be a happy ending; the themes are borrowed from epic and legendary history; development and unravelling of the plot are retarded and so the envy of malign influences averted by the almost obligatory buffoon (*vidūṣaka*, "the spoiler"). The less cultured masses still have, in addition to the films, which often use the same religious and mythic themes, their uncomplicated *yātrās*, a combination of stage play and various festivities, which have contributed much to the spread of the "puranic" view of life.

Yātrās

Dancing is not only an aesthetic pursuit but a divine service. Hence there are halls for sacred dances annexed to some temples. The rhythmic movement has a compelling force, generating and concentrating power or releasing superfluous energy. It induces the experience of the divine and transforms the dancer into whatever he impersonates. Thus, many tribal dances consist of symbolic enactments of events (harvest, battles) that are desired to be accomplished successfully. Musicians and dancing girls accompany processions to expel the demons of cholera or cattle plague. Even today, religious themes and the various relations between man and God are "danced" and made visual by the codified symbolic meanings of gestures and movements (see SOUTH ASIAN PEOPLES, ARTS OF).

V. Place of Hinduism in Indian and world religions

HINDUISM AND OTHER RELIGIONS OF INDIAN ORIGIN

Buddhism and Jainism are, essentially, new interpretations rather than repudiations of the common Indian tradition.

Buddhism. Since Buddhism did not interfere with Hindu customs and usages, allowing its adherents to approach Hindu or local supernatural powers for immediate favours, Hindu criticism came mainly from Brahmin philosophers who opposed its adherents because they rejected the authority of the Veda and the Brahmins and the doctrine of the *ātman* (soul) and because they admitted persons of any age and caste to monastic life. The spread of Buddhism was often regarded as an indication of the degeneration of humanity. In the course of time, Buddha was recognized as an incarnation of Viṣṇu, but this recognition was often qualified by the addition that Viṣṇu assumed this form in order to mislead and destroy the enemies of the Veda. Yet this *avatāra* is rarely worshipped. Buddhist emblems also were often ascribed to Viṣṇu or Śiva. Some ancient Buddhist shrines have remained partly under the supervision of Hindu ascetics and are visited by pilgrims notwithstanding their much neglected condition.

After the rise of Buddhological studies in the West and the archaeological discoveries and restorations since the end of the 19th century had made Indians more aware of the Indian origin of Buddhism, the Republic of India adopted the Buddhist emperor Aśoka's lion capital, marking the place of Buddha's first teaching, as its national emblem. The Buddha jubilee in 1956 was an occasion for enthusiastic celebrations. In recent years the number of Indian Buddhists is again increasing, due mainly to the conversion of persons of low social rank for whom, after conversion, the disabilities of their birth no longer remain.

Jainism. With Jainism, which always remained an Indian religion, Hinduism has, especially in social institutions and ritual life, so much in common that nowadays Hindus generally tend to consider it a Hindu sect and many Jainas are much inclined to fraternization. The points of difference—*e.g.*, a stricter *ahiṃsā* practice and the absence of sacrifices for the deceased in Jainism—do not give any offense (see BUDDHISM; JAINISM).

HINDUISM AND ISLAM

Islām, so different from Hinduism in creed and institutions, was, in India, neither absorbed nor powerful enough to make India a Muslim country. The religious situation created by the presence of its numerous adherents always had explosive potentialities: the Muslims do not respect bovine life and regard Hindu cult practices as objectionable idolatry. Although the Indian Muslims, with few exceptions, are of native descent, they are theoretically outcastes with whom dealings must remain restricted by formal rules; but, like Christians, they are less polluting than the Hindu lower castes. The Islāmic way of life meets with opposition; Muslims and Hindus do not ordinarily intermarry or dine together. Up to the present day, this situation has raised acute and even devastating issues, but it does vary somewhat from region to region, from village to village, and from class to class. Very often, however, mutual differences are accepted. Though repudiating caste, Muslims often observe it in practice or have even after their conversion to Islām retained their original caste organization.

Hindus, inclined to worship the holy whatever its manifestations, may revere Muslim saints or take part in Muslim festivities, often to such an extent that the character of these celebrations has been altered. The Muslim faith of ordinary people was constantly exposed to the subtle infection of Hinduism. This led, especially among the mystically minded who speculated whether the sole reality should be called Allāh or Brahman, to various syntheses, such as Sikhism, which are, for the Hindu, Hindu sects, and to communities whose religious practice is partly Muslim, partly Hindu. Those who, like Gandhi, could not understand the intolerance of orthodox Islām sympathized with the moderation and eclecticism of these groups. Most of the educated class, however, always remained aware of the cleavage. To the Muslims—who, as part of an ecumenical community stretching over large parts of Asia and Africa, are concerned about the political and religious crisis of Islām since the late 19th century—the collapse of the Mughals after the Indian Mutiny (1857) was a severe blow that did not contribute to closer and better relations with Hindus. This is particularly true because anti-Muslim tendencies had, since the renascent Hinduism of the Marāṭhā movement and in later times in the Arya Samaj (see above), won ground, and Muslims became self-assertive and determined to maintain their distinctive position. After partition of the subcontinent into India and Pakistan—partly based on religious differences—and independence (1947), the political controversies between India and Pakistan constituted a complication.

HINDUISM AND CHRISTIANITY

For the eclectic and undogmatic educated Hindu, who believes that religion is a matter of personal realization, every religion is true and a path to truth. If the adherents of Christianity sincerely follow it, the Hindu's attitude toward it, notwithstanding what he believes to be the militant and essentially intolerant disposition of the followers of Christianity—which is regretted by Hindus—continues to be one of respect and understanding, of tolerance and even sympathy. The Hindu is ready to accept the ethical teachings of the Gospels, particularly the Sermon on the Mount (whose influence on Gandhi is well known) but rejects the theological superstructure. Many adherents of *bhakti* movements—the Christian influence on which has been grossly exaggerated—feel that the Christian conceptions, which are regarded as a kind of *bhakti*, do not realize in God the multiplicity of human relations of love and service. Educated Hindus, though assimilating some Christian ideas, often regard missionary propaganda as an attack on their national genius and time-honoured institutions and take offense at what they regard as the disrespectful utterances of Christian missionary literature. They are averse to the organization, the reliance on authorities, and the exclusiveness of Islām and Christianity, considering these obstacles to harmonious cooperation. They subscribe to Gandhi's opinion that missionaries should confine their activities to humanitarian service. Since independence, conversion has indeed been viewed with disfavour by many influential Indians, who often also find in Hinduism what might be attractive in Christianity. Movements that, like the Arya Samaj, advocate a Hindu theism designed to rival Islām and Christianity, make serious efforts to reconvert Christians to the Hindu community. The ordinary people tolerate the proximity of Christian converts, even if these transgress Hindu taboos, provided they form a more or less separate community. Thus Christians often form castes or endogamous bodies analogous to castes. They are sometimes admitted to temples to which untouchable Hindus have no entrance. In Malabar, Christians came, on account of their economic position, to be practically equalized with Brahmins. Nationalism has challenged the more serious-minded Indian Christians to express the genius of their faith in Indian modes and patterns. This has led, since 1921, to the emergence of Christian *āśramas* in the south. The dialogue between Hinduism and Christianity is more or less institutionalized at Bangalore in Mysore state where the Christian Institute for the Study of Religion and Society is located. Its bulletin offers an opportunity for discussion between, for example, Christians and supporters of the Ramakrishna Mission.

The Arya Samaj movement

BIBLIOGRAPHY

General Surveys: C.N.E. ELIOT, *Hinduism and Buddhism*, 3 vol. (1921, reprinted 1962), a useful historical survey, though mainly devoted to Buddhism; L. RENOU, *Religions of Ancient India* (1953), an excellent introduction to some of the main aspects of ancient and medieval Hinduism, and *L'Hindouisme*, 2nd ed. (1958; Eng. trans., *The Nature of Hinduism* (1962), a succinct account of the main facts; R.C. ZAEHNER, *Hinduism* (1962), a discussion of the essential facets of this subject; J. GONDA, *Die Religionen Indiens*, 2 vol. (1960–63), a comprehensive study with full bibliographies; K.W. MORGAN (ed.), *The Religion of the Hindus* (1953), a collection of articles by Hindu scholars.

On Vedism and Brahmanism: A.B. KEITH, *The Religion and Philosophy of the Veda and Upanishads*, 2 vol. (1925), theoretically somewhat dated; J. GONDA, *Viṣṇuism and Śivaism* (1970), a comparative treatment of their development, theology, rituals, and mutual relations.

On Vaiṣṇavism: J. GONDA, *Aspects of Early Viṣṇuism*, 2nd ed. (1969), an attempt at understanding the significance of Viṣṇu as it presented itself to the ancient Indians; R.C. ZAEHNER, *The Bhagavad-Gītā* (1969), a reliable translation and the most recent commentary; M.B. SINGER (ed.), *Krishna: Myths, Rites and Attitudes* (1968).

On Śaivism: W.D. O'FLAHERTY, "Asceticism and Sexuality in the Mythology of Śiva," *History of Religions*, 8:300–337 (1969), and 9:1–41 (1970).

On Tantrism and Śāktism: J.G. WOODROFFE, *Introduction to Tantra Śāstra*, 5th ed. (1969), and *Shakti and Shakta*, 6th ed. (1965). The same scholar published and translated, often under the pseudonym Arthur Avalon, many Tantric texts. K.W. BOLLE, *The Persistence of Religion* (1965), an essay on Tantrism; E.C. DIMOCK, JR., *The Place of the Hidden Moon: On Erotic Mysticism in the Bengal Vaiṣṇava-Sahajiyā Cult* (1966).

On folk Hinduism: L.S.S. O'MALLEY, *Popular Hinduism: The Religion of the Masses* (1935); E.B. HARPER (ed.), *Religion in South Asia* (1964), on the meaning of gods and rituals in the lives of common people; M.N. SRINIVAS, *Religion and Society Among the Coorgs of South India* (1952).

On modern Hinduism: J.N. FARQUHAR, *Modern Religious Movements in India* (1915); D.S. SARMA, *Studies in the Renaissance of Hinduism in the 19th and 20th Centuries* (1944); H. BHATTACHARYYA (ed.), *The Cultural Heritage of India*, vol. 4, 2nd ed. (1956).

On rituals and religious celebrations: P.V. KANE, *History of Dharmaśāstra*, 5 vol. (1930–62), the standard work of ancient and medieval religious and social law, customs, and rituals; C.G. DIEHL, *Instrument and Purpose: Studies on Rites and Rituals in South India* (1956).

On social practices and institutions: L.S.S. O'MALLEY, *Indian Caste Customs* (1932); J.H. HUTTON, *Caste in India: Its Nature, Function and Origins* (1946); G.S. GHURYE, *Indian Sadhus* (1953); S.N. DASGUPTA, *Yoga as Philosophy and Religion* (1924).

On symbols, plastic arts, theatre, and dance: H.R. ZIMMER, *Myths and Symbols in Indian Art and Civilization* (1946); G. TUCCI, *The Theory and Practice of the Maṇḍala* (Eng. trans. 1961); J.N. BANERJEA, *The Development of Hindu Iconography*, 2nd rev. ed. (1956).

Hinduism and world religions: D.E. SMITH (ed.), *South Asian Politics and Religion* (1966); A. KRAMER, *Christus und Christentum im Denken des modernen Hinduismus* (1958).

(Ed.)

Hinduism, History of

The history of Hinduism, a religion that in the 20th century claims more than 437,000,000 followers, began in India about 1500 BC. From its literature Hinduism can be traced back to before 1000 BC; evidence of its earlier antecedents is derived from archaeology, comparative philology, and comparative religion. This article is concerned only with the history of Hinduism; the manifold doctrines, practices and scriptures of the religion are described in greater detail elsewhere (see also HINDUISM; HINDU SACRED LITERATURE).

This article is divided into the following sections:

Sources of Hinduism
  Indo-European sources
  Non-Indo-European sources
The prehistoric period (3rd and 2nd millennia BC)
The Vedic period (2nd millennium–7th century BC)
  The beginning of text and ritual: the Ṛgveda
  Elaborations of text and ritual: the later Vedas and Brāhmaṇas
  The beginnings of philosophy: the *Upaniṣads*
Challenges to Brahminism (7th–2nd centuries BC)
Early Hinduism (2nd century BC–4th century AD)
  Transformations of Brahminism and popular religion
  The rise of the major sects: Vaiṣṇavism, Śaivism, and Śāktism
  Interplay of Indo-European and Dravidian cultures
  Establishment of basic doctrines and practices
  The spread of Hinduism in Southeast Asia and the Pacific
  Indian religious influence in the Mediterranean world
The Purāṇic period (4th–8th centuries)
The rise of devotional Hinduism (8th–11th centuries)
The age of Bhakti (11th–19th centuries)
  The challenge of Islām and popular religion
  Philosophical movements
  Bhakti movements
  The synthesis of Hinduism with Islām and folk religion
The modern period (19th–20th centuries)
  Hindu reform movements
  New religious movements
  The struggle for independence
  Hinduism outside India

SOURCES OF HINDUISM

Indo-European sources. The earliest literary source for the history of Hinduism is the Ṛgveda, the hymns of which were chiefly composed over the last two or three centuries of the 2nd millennium BC. The religious life reflected in this text is not that of Hinduism but of an earlier sacrificial religious system, generally known as Brahmanism or Vedism, which developed in India among the invaders who brought with them the horse and chariot and the Sanskrit language and who are generally known as Aryans, the name by which they referred to themselves. They were a branch of a related group of nomadic and seminomadic tribal peoples originally inhabiting the steppe country of southern Russia and Central Asia. Other branches of these peoples penetrated into Europe, bringing with them Indo-European languages that developed into the chief language groups now spoken there.

The religion of the Ṛgveda

Before they entered the Indian subcontinent (*c.* 1500 BC), the Aryans were in close contact with the ancestors of the Iranians, as evidenced by a near kinship between Sanskrit and the earliest surviving Iranian languages. Thus, the religion of the Ṛgveda contains elements from three evolutionary strata: an early element common to most of the Indo-European tribes; a later element held in common with the early Iranians; and an element acquired in the Indian subcontinent itself, after the main Aryan migrations. Hinduism arose from the continued accretion of further elements derived from the original non-Aryan inhabitants, from outside sources, and from the geniuses of individual reformers at all periods. The accretion was accompanied by an inverse process of dropping elements of beliefs and practices that had outlived their usefulness. This has been going on, often almost imperceptibly, throughout the history of Hinduism.

Indo-European influences. In Hinduism there are few direct survivals of the Indo-European heritage. Some of the rituals of the Hindu wedding ceremony, notably the circumambulation of the sacred fire and the cult of the domestic fire itself, look back to the remote Indo-European past. The same is probably true of the custom of cremation and some aspects of the ancestor cult. The Ṛgveda shows many other Indo-European elements, such as the worship of male sky gods with sacrifices, and the old heaven god Dyaus, whose name is cognate with those of the classical Zeus of Greece and Jupiter of Rome ("Father Jove"). The Vedic heaven, the "World of the fathers," resembled the Germanic Valhalla and seems also to be an Indo-European inheritance.

Cremation and the ancestor cult

Indo-Iranian influences. The Indo-Iranian element in later Hinduism is chiefly to be found in the initiatory ceremony (*upanayana*) performed with boys of the three upper classes, a rite both in Hinduism and in Zoroastrianism that involves the tying of a sacred cord. Vedic religion showed many such elements. The Vedic god Varuṇa, now an unimportant sea god, appears in the Ṛgveda as sharing many features of the Zoroastrian Ahura Mazdā ("Wise Lord"); the narcotic sacred drink *soma* corresponds to the sacred *haoma* of Zoroastrianism.

Indigenous influences. Even in the earlier parts of the Ṛgveda the religion had already acquired numerous specifically Indian features. Some of the chief gods, notably Indra (god of storms and battle), have no clear Indo-European or Indo-Iranian counterparts. Some of the new features may have evolved entirely within the Aryan framework, but it is generally presumed that many of them stem from the influence of the indigenous inhabitants. That the Vedic Aryans were in direct contact with the civilization of the Indus Valley in its prime is doubtful, but the religion of the valley's culture probably had some influence upon the Aryans.

Non-Indo-European sources. *The Dravidian hypothesis.* All features of Hinduism that cannot be traced back to the Ṛgveda are sometimes ascribed to the influence of the original inhabitants, who are often vaguely and incorrectly referred to as "Dravidians." The ruling classes of the Harappā culture (*c.* 2300–1700 BC), or the Indus civilization, may have spoken a Dravidian language, but the presence of Dravidian speakers throughout the whole subcontinent at any time in history is not attested. The Mediterranean racial type, to which most modern higher caste Dravidian speakers belong, is widespread throughout India; but it cannot be proved that all men of this type originally spoke Dravidian languages or that all followed the same culture. Equally or more widely spread in South and Southeast Asia is the Proto-Australoid racial type, the purest members of which in India are the tribal peoples of the centre and the south, many of whom speak languages of the Austric family. Thus, although many aspects of Hinduism are traceable to non-Aryan influence, not all of these aspects are borrowed from "Dravidians." In the 20th century, the term Dravidian is generally used to refer to a family of languages and not to an ethnic group.

Role of the Proto-Australoid racial type

Other influences. Further influences have affected Hinduism in historic times. The Central Asian nomads who entered India in the two centuries before and after the beginning of the Christian Era may have had some effect on the growth of devotional Hinduism out of Vedic religion. There was indirect influence of the classical western world on Hindu religious art, and several features of Hinduism may be traced to Zoroastrianism. The influence of later Taoism from China on Tantric Hinduism (an esoteric system of achieving release from transmigration) has been suggested, though this cannot be finally proved. In more recent centuries, the influence of Islām and Christianity on Hinduism can be seen.

The process of "Sanskritization" and its converse. The development of Hinduism may be interpreted as a constant interaction between the religion of the upper social groups, represented by the Brahmins (priests and teachers), and the cults of the masses. From the time of the Aryan invasion (*c.* 1500 BC) the indigenous inhabitants of the subcontinent have tended to adapt their religious and social life to Brahminic norms. This has developed from a desire of lower class groups to rise on the social ladder by adopting the ways and beliefs of the higher castes. This process, sometimes called "Sanskritization," was going on in Vedic times, when non-Aryan chieftains accepted the ministrations of Brahmins and thus achieved status in the social system for themselves and their subjects. It was probably the main means whereby Hinduism spread through the subcontinent and beyond it into Southeast Asia. Sanskritization still continues in the form of the conversion of tribal groups, and it is reflected in the persistent tendency of low-caste Hindus to try to raise their status by adopting high-caste customs, such as wearing the sacred cord and becoming vegetarians.

If "Sanskritization" has been the main means of spreading Hinduism throughout the subcontinent, its converse process, which has no convenient label, has been one of the means whereby Hinduism has changed and developed over the centuries. The Aryan conquerors had to live side by side with the Aboriginals, and many features of Hinduism, as distinct from Vedic religion, apparently developed from the cults of the non-Aryan lower orders. The doctrine of transmigration, which is the hallmark of Indian religions, may have developed in this kind of process. The phallic emblem of Śaivism, a major form of Hinduism centring on devotion to the god Śiva (originally a storm god), most likely was borrowed from early popular fertility cults. Many features of Hindu mythology and several of the lesser gods—such as Gaṇeśa, an elephant-headed god, and Hanumān, the monkey god—were incorporated into Hinduism by this means. The system of ecstatic devotional religion known as Bhakti seems to have begun in unorthodox circles on the fringes of Brahminic culture. Thus, the history of Hinduism may be interpreted on the one hand as the imposition of orthodox custom upon wider and wider ranges of people and on the other as the survival of features of non-Aryan cults, which gained strength steadily until they were taken over or adapted by the Brahmins.

Accretions from non-Aryan religions

THE PREHISTORIC PERIOD (3RD AND 2ND MILLENNIA BC)

Indigenous prehistoric religion. The prehistoric culture of the Indus Valley arose in the latter centuries of the 3rd millennium BC from the metal-using village cultures of the region. There is considerable evidence for the religious life of the Indus people, but in the absence of intelligible written records its interpretation is speculative. There is enough, however, to show that several features of later Hinduism had prehistoric prototypes.

In most of the village cultures, small terracotta figurines of women, found in large quantities, have been interpreted as icons of a fertility deity whose cult was widespread in the Mediterranean area and in West Asia from Neolithic times onward. This hypothesis is strengthened by the fact that the goddess was apparently in several instances associated with the bull—a feature also found in the ancient religions farther west.

Religion in the Indus Valley civilizations. The Harappā culture (often referred to as the Indus Valley civilization, located in modern Pakistan) has produced much evidence of the cult of the goddess and the bull. Figurines of both occur, the goddess being commoner than the bull. The bull, however, appears more frequently on the many seals. A horned deity, possibly with three faces, occurs on a few of these; in one specimen it is surrounded by animals. A few male figurines in hieratic (sacerdotal) poses and one apparently in a dancing posture suggest deities. No building has been discovered at any Harappan site that can with certainty be interpreted as a temple, but the famous "Great Bath" at Mohenjo-daro (also in Pakistan) was almost certainly used for ritual purposes, as were the *ghāṭs* (bathing steps on riverbanks) attached to later Hindu temples. The presence of bathrooms in most of the houses and the remarkable system of covered drains indicate a strong feeling for cleanliness that was probably connected with concepts of ritual purity and taboo, rather than with ideas of hygiene.

Many of the seals show religious and legendary themes that cannot be interpreted with certainty. There is clear evidence, however, of the worship of sacred trees or of the divinities believed to reside in them. The bull is often depicted standing before a sort of altar, which may be a cult object connected with a fertility ritual. The horned god has been interpreted, perhaps overconfidently, as a prototype of the Hindu god Śiva. Small conical objects appear to be phallic emblems, also connected with Śiva in later Hinduism, although they may have been pieces used in board games. Other interpretations of the remains of the Harappā culture are more speculative and, if accepted, would indicate that many features of later Hinduism were already in existence 4,000 years ago.

The fact that Harappans buried their dead with grave deposits, a practice not followed by the later Hindus, shows that they had some belief in an afterlife. At Lothal, a Harappan site in Gujarāt (in western India), a double grave containing a man and a woman has been found, possibly indicating the occasional sacrifice of a widow upon her husband's death, similar to the later self-sacrifice of the *satī* ("virtuous woman").

Harappan burials

Survival of archaic religious practices in modern folk and primitive religions. However much the more developed forms of Hinduism may owe to the Harappā culture, some elements of the religious life of that civilization may be paralleled in the current and past folk religions of India—notably sacred animals, sacred trees, especially the *pīpal* (*Ficus religiosa*), and the use of small figurines for cult purposes. Such features of primitive religion are found in all parts of India, among tribal peoples as well as among the less sophisticated Hindus.

THE VEDIC PERIOD (2ND MILLENNIUM–7TH CENTURY BC)

The beginning of text and ritual: the Ṛgveda. The Aryans of the early Vedic period left hardly any material remains, but they did leave a very important literary record in the Ṛgveda (see also HINDU SACRED LITERATURE).

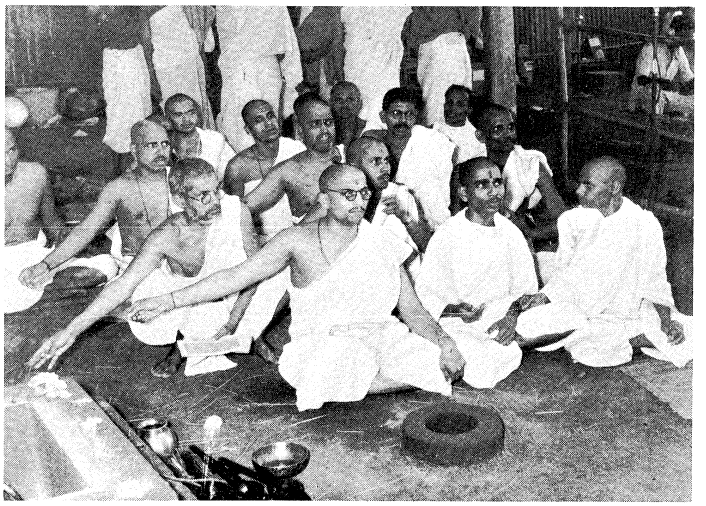
The Ṛgveda is not a unitary work, and may have been composed over several centuries. In its form at the time of its final edition it reflects a well-developed religious system and a culture considerably in advance of that depicted in the earlier hymns. The date commonly given for the recension of the Ṛgveda is 900 BC.

The religion reflected in the Ṛgveda is a polytheism mainly concerned with the propitiation of divinities associated with the sky and the atmosphere. Of these, the Indo-European sky father Dyaus, mentioned above, had by then become little regarded. More important were such gods as Indra, Varuṇa (guardian of the cosmic order), Agni (the sacrificial fire), and Sūrya (the sun; see also HINDU MYTHOLOGY).

The main ritual activity referred to in the Ṛgveda is the *soma* sacrifice. *Soma* was a narcotic beverage prepared from a now-unknown plant; recently it has been suggested that the plant was a mushroom. Later another plant was substituted for the agaric fungus, which had become difficult to obtain. The Ṛgveda contains few clear references to animal sacrifice, which probably became more widespread later. There is some doubt whether the priests formed a separate class of society at the beginning of the Ṛgvedic period. If they did so, the prevailingly loose boundaries of class made it possible for a man of nonpriestly parentage to become a priest. By the end of the period, however, they had become a separate class of specialists, the Brahmins (Brāhmaṇas), who claimed superiority over all the other social classes, including the Rājanyas, the tribal aristocrats.

The *soma* sacrifice in the Ṛgveda

The Ṛgveda throws little light on birth rituals, but the rites of marriage and disposal of the dead were basically the same as in later Hinduism. Marriage was an indissoluble bond cemented by a lengthy and solemn ritual





Scenes of a *soma* sacrifice performed in Pune (Poona), India, on behalf of a traditional Brahmin, following the same ritual used in 500 BC, an unusual example of the continuance of the Vedic tradition. (Though Vedic rites are still occasionally performed, photographic documentation of a complete and ritually correct ceremony is rare.) (Top) Group of 16 priests (four of each of the four Vedas) swear to the common interest of the sacrificer prior to the beginning of the ritual, their hands outstretched toward the fire (*āhavanīya*). (Bottom) Priests bring in fire to light the sacrificial fire in the newly constructed altar, taken from the household fire of the sacrificer, shown with his wife in the middle of the group.
C.M. Natu

centring on the domestic hearth. The funeral rites of the rich included cremation, though other funeral forms were also practiced. An interesting reference in one hymn shows that the wife of the dead man lay down beside him on the funeral pyre but was called upon to return to the land of the living before it was lighted. This may have been a survival from an earlier period when the wife was actually cremated with the husband, a custom that was revived in later times.

Among other features of Ṛgvedic religious life that were important for later generations were the groups or fraternities of *munis*. The *muni* was apparently a sort of wizard or shaman (religious personage having healing and psychic transformation powers), trained in various magic arts and believed to be capable of supernatural feats, such as levitation. He was particularly associated with the god Rudra, a deity connected with mountains and storm and more feared than loved. Rudra developed into the Hindu god Śiva, and his prestige increased steadily with time. The same is true of Viṣṇu, a minor solar deity in the Ṛgveda, who later became one of the most important and popular divinities of Hinduism.

Some of the hymns of the tenth book of the Ṛgveda show speculative tendencies, which were the beginnings of the persistent effort of the Indian philosopher to reduce all things to a single basic principle. The favourite myth of the Aryans apparently was one that attributed the origin of the cosmos to the god Indra, after he had slain the great dragon Vṛtra, a myth very similar to that of early Mesopotamia. With time, such tales were given up for speculative theories that are reflected in several hymns of the late tenth book of the Ṛgveda.

Even more important as a landmark in the history of Indian religion is the "Hymn of the Person" ("Puruṣa-sūkta," RV. x, 90), which is one of the latest hymns in the collection (see also HINDUISM). Here the four classes (*varṇa*) of Indian society are referred to: the priest (Brāhmaṇa) emerging from the head, the warrior (Rājanya, later Kṣatriya) from the arms, the peasant (Vaiśya) from the trunk, and the servant (Śūdra) from the legs of the primeval victim of a cosmic sacrifice. The "Puruṣa-sūkta" represents the beginning of a new phase, in which the sacrifice became even more important and elaborate, and the Brahmins even more exalted.

The exaltation of the Brahmins

Elaborations of text and ritual: the later Vedas and Brāhmaṇas. The chronology of later Vedic developments is extremely vague, but they can probably be related to the Painted Grayware strata in the archaeological sites of the western Ganges (Gaṅgā) Valley, dated from about 1000 to 500 BC. These reflect a culture still illiterate but showing considerable advances in civilization. Nothing, however, has yet been discovered from sites of this period that throws much light on the religious situation, and the historian is still chiefly dependent on texts.

The period was one in which the sacrifices were developed to become lengthy series of rituals—sometimes lasting more than a year—in which hundreds of animals were slaughtered. As the world was produced by sacrifice, according to the Vedas, so also a constant performance of sacrifice was required to maintain it, and this demanded numerous highly trained Brahmins to intone the hymns, to utter the ritual phrases, and to perform the ritual actions with scrupulous accuracy. Thus, the influence of the priesthood increased.

The class of Brahmins by then was divided into groups that married exogamously (only to members of other groups, or *gotras*), and the Brahmins formed various schools (*śākhā*) according to the recension of the sacred texts (*Brāhmaṇas*) that were specially studied. The Brahmin claimed supremacy over the ruler, although often he did not obtain it. One might suspect that the great elaboration of the sacrifice at this period and the appearance of long and expensive new sacrifices, especially intended for royal patrons, were deliberately devised to enhance the influence and wealth of the priests by flattering the kings. The lengthy series of rituals of the royal consecration, the *rājasūya*, emphasized the royal power and endowed the king with a divine charisma, raising him, at least for the duration of the ceremony, to the status of a god. Typical of this period is the elaborate *aśvamedha*, the horse sacrifice, in which a consecrated horse was left loose and allowed to wander at will for a year, followed by the king's troops, who defended it from all attack until it was brought back to the royal capital and sacrificed with a very complicated ritual. The ambition to perform an *aśvamedha* was encouraged in later times by the Brahminic traditions and was one of the causes that contributed to the instability of Hindu kingdoms.

The beginnings of philosophy: the "Upaniṣads." The next phase of Indian religious life, roughly between 700 and 500 BC, is the period of the beginnings of philosophy and mysticism marked by the *Upaniṣads* ("sittings near a teacher"). Historically, the most important of these are the two oldest, the *Bṛhadāraṇyaka* ("great forest text") and the *Chāndogya* (pertaining to the "Chandogas," a class of Brahmin specially connected with the intonation of hymns at sacrifices), both of which are compilations recording traditions of certain sages (*ṛṣi*) of the time, notably Yājñavalkya, who was a pioneer of new religious ideas.

The primary motive of the *Upaniṣads* is a desire for

mystical knowledge such as would ensure freedom from "re-death." Throughout the later Vedic period, the idea that the world of the fathers was not the end—and that even in heaven death was inevitable—had been growing. The age of the *Upaniṣads* was one of rapid material progress, and urban civilization was reviving. The geographical horizon was also expanding, as the "other worldly" aspects of Indian religion became prominent.

The development of asceticism and belief in rebirth

The new developments were connected with the growth of asceticism and the appearance of the doctrine of rebirth, generally called transmigration.

The Ṛgveda shows little evidence of asceticism, except among the *munis* (shamans). In the Atharvaveda appears another class of religious adepts, or specialists, the *vrātyas*, particularly connected with the region of Magadha (southern Bihār). The *vrātya* was a wandering hierophant (one who manifested the Holy), outside the regular system of Vedic religion. He travelled from place to place in a bullock cart with an apprentice and with a woman who appears to have been used for ritual prostitution. Flagellation and other forms of self-mortification seem to have been part of his routine. Efforts were made by the orthodox to bring the *vrātyas* into the Vedic system by special rituals of conversion, and it may be that these people helped to introduce non-Aryan beliefs and practices into Vedic religion. Meanwhile, the more complex sacrifices of the later Vedic period demanded purificatory rituals such as fasting and vigil as part of the preparations for the ceremony. Thus there was a growing tendency toward the mortification of the flesh.

The origin and the development of the belief in the transmigration of souls are very obscure. A few passages have been found to suggest that this doctrine was known even in the days of the Ṛgveda, but it is first clearly propounded in the earliest *Upaniṣad*—the *Bṛhadāraṇyaka*. Here it is stated that normally the soul returns to Earth and is reborn in human or animal form. This doctrine of *saṃsāra* (reincarnation) is attributed to the sage Uddālaka Āruṇi, who is said to have learned it from a Kṣatriya chief. In the same text, the doctrine of *karma* (works), that the soul achieves a happy or unhappy rebirth according to its works in the previous life, also occurs for the first time, attributed to Yājñavalkya. Both doctrines appear to have been new and strange ones, circulating among small groups of ascetics who were disinclined to make them public, perhaps for fear of the orthodox priests. These doctrines must have spread rapidly, for in the later *Upaniṣads* and in the earliest Buddhist and Jain scriptures they are common knowledge.

Dissatisfaction with the ritualism of Vedic religion and the increasing popularity of the life of the ascetic wanderer or the hermit encouraged mysticism and a search for the ultimate, union with which would bring complete bliss and release (*mokṣa*) from transmigration.

The concept of the Absolute as a neuter entity was not universal even in the earlier *Upaniṣads*. The oldest, the *Bṛhadāraṇyaka*, contains little-noticed passages of theistic import. The later *Upaniṣads* have a much larger theistic content and link up with theistic developments of later times. The *Upaniṣads* represent the beginnings of philosophy in India. Their authors usually make their points by analogical arguments, but the literature of the *Upaniṣads* also contains numerous attempts at logical reasoning and argument from observed facts.

CHALLENGES TO BRAHMINISM (7TH–2ND CENTURIES BC)

Domestic ritualism and reform

The elaboration of expensive and large-scale ritual, which is evident in the later Vedic literature, was accompanied by the first steps in the codification of the rules of Aryan conduct in the literature of the *sūtras* ("threads," or collections of aphoristic statements). From about 600 BC, texts appeared containing terse instructions on the ritual of lesser sacrifices (*Śrauta-sūtras*), on domestic rituals (*Gṛhya-sūtras*), and on the conduct of life (*Dharma-sūtras*). Society was ritually stratified in the four classes, each of which had its own *dharma* (law), or eternal norm of conduct. The ideal life was sacramentalized by numerous ceremonies incumbent upon the upper classes, carrying the individual from the cradle to the grave in a series of complex rites. The *Gṛhya-sūtras* show that in the popular religion of the time there were many minor divinities who are hardly mentioned in the literature of the large-scale sacrifices but who were probably far more influential on the lives of the lower orders than the great gods of Vedism (see also HINDUISM).

The century from about 550 BC onward was a period of great change in the religious life of India, for it saw the rise of breakaway sects of ascetics denying the authority of the Vedas and of the Brahmins and looking up to founders who claimed to have discovered the secret of obtaining release from transmigration. The most important of these were Siddhārtha Gautama, the Buddha, and Nātaputta Vardhamāna, called Mahāvīra ("Great Hero"), the great teacher of Jainism (see also BUDDHISM; JAINISM). There were many other heterodox teachers who organized bands of ascetic followers, each group following a specific code of conduct. They gained considerable support from ruling families and merchants. The latter were growing in wealth and influence, and many of them were in search of forms of religious activity that would give them a more significant part to play than did orthodox Brahminism and would be less expensive to support.

The new religious movements left scriptures behind them that throw some light on the popular religious life of the period. The god Brahmā was widely looked on as the highest god and the creator of the universe, with Indra, known chiefly as Śakra ("the mighty one"), second to him in importance. The Brahmins were very influential, but opposition had developed to their large-scale animal sacrifices—on both humanitarian and economic grounds—and their pretensions to superiority by virtue of their birth were questioned. The doctrine of transmigration was now almost universally accepted, although a group of outright materialists denied the survival of the soul after death. The ancestor cult, part of the Indo-European heritage, was kept up almost universally, at least by the higher castes. Popular religious life largely centred around the worship of local fertility divinities (*yakṣa*, *yakṣī*), snake-spirits (*nāga*), and other minor spirits in sacred places and groves (*caitya*). These sacred places were the main centres of popular religious life. There is no evidence of any buildings or images associated with them, and it appears that neither temples nor large icons existed at the time.

At this time (*c.* 500 BC), asceticism became widespread, and more and more intelligent young men "gave up the world" in search of release from transmigration in a state of psychic security. The orthodox Brahminical teachers reacted to these tendencies by devising the doctrine of the four *āśramas* (abodes), which divided the life of the twice-born after initiation into four stages: the *brahmacārin* (celibate religious student); the *gṛhastha* (married householder); the *vānaprastha* (forest hermit); and the *sannyāsin* (wandering ascetic). This attempt to keep asceticism in check and confine it to men of late middle age was never followed universally, but henceforth Hindu social theory centred round the concept of *varṇāśrama-dharma*, the duties of the four classes (*varṇa*) and the four stages of life (*āśrama*), which formed the ideal that the Hindu was encouraged to follow.

The doctrine of the four *āśramas*

The 3rd century BC was the period of the Mauryan Empire, the first great empire of India. Its early rulers were heterodox, and Aśoka (reigned *c.* 265–238 BC), third and most famous of the Mauryan rulers, was a professed Buddhist. There is no doubt that Aśoka's patronage of Buddhism did much to spread that religion, though in his inscriptions he recognizes the Brahmins as worthy of respect. Sentiments in favour of nonviolence (*ahiṃsā*) and vegetarianism, much encouraged by the heterodox sects, spread during the Mauryan period and were greatly encouraged by Aśoka, though such sentiments began long before Buddhism. A Brahminic revival appears to have occurred with the fall of the Mauryas. But the orthodox religion was itself undergoing change at this time, with a development of theistic tendencies that centred around the gods Viṣṇu and Śiva.

Inscriptional and iconographic evidence of devotional theism appears in the 2nd century BC, together with literary references. Several brief votive inscriptions refer to the god Vāsudeva, who by this time was widely worshipped in western India. At the end of the 2nd century, Heliodorus, a Greek ambassador from King Antialcidas of Taxila (in Pakistan), erected a large column in honour of Vāsudeva at Besnagar in Madhya Pradesh and recorded that he was a Bhāgavata, a term specially used for the devotees of Viṣṇu. It is possible that Vāsudeva originally may have been a popular god outside the orthodox pantheon, but his identification with the old Vedic god Viṣṇu and, later, with Viṣṇu's incarnation, Kṛṣṇa, was rapidly accepted.

From around the end of the Mauryan period the first surviving stone images of Hinduism appear, not representing any of the great gods but *yakṣa*s, or local chthonic (earth or fertility) divinities, of whom several large, rather crude figures survive. The original locations of these images are uncertain, but they were probably erected in the open air in sacred enclosures. Temples are not clearly attested in this period either by archaeology or literature. A very few fragmentary images of pre-Christian date are thought to be those of Vāsudeva, and Śiva, both in anthropomorphic form and in the form of a *liṅga*, or phallic emblem, is attested on coins of the 2nd and 1st centuries BC.

EARLY HINDUISM (2ND CENTURY BC–4TH CENTURY AD)

Transformations of Brahminism and popular religion. The development of the cult of Viṣṇu was much encouraged by the editors of the *Mahābhārata*, a great Hindu epic. Numerous interpolated episodes were introduced into the epic, among them the *Bhagavadgītā* ("Song of the Lord"). This marks the real beginning of Hinduism as distinct from Brahminism and the culmination of the literature of the *Upaniṣads*. Beside the impersonal, rather intellectual mysticism of the early *Upaniṣads*, a new theism developed, replacing sacrifice (*yajña*) by worship (*pūjā*) and the attitude of self-knowledge and introspection by one of respect and devotion (*bhakti*) to a personal god, who, in the case of the *Bhagavadgītā*, is conceived of as the Vedic god Viṣṇu incarnate as Kṛṣṇa. The *Bhagavadgītā* is significant also for its sturdy defense of the system of the four classes and for its doctrine that selfless action, done in order to carry out the duties of one's class and for the glory of God, is a more effective means of obtaining release from transmigration than sacrifice, asceticism, or meditation. It is chiefly this ethical doctrine that has made the *Bhagavadgītā* so influential in the Hinduism of the 19th and 20th centuries.

First distinction of Hinduism from Brahminism

The rise of the major sects: Vaiṣṇavism, Śaivism, and Śāktism. Alongside the cult of Viṣṇu, that of Śiva developed. The god Rudra gained in importance from the end of the Ṛgvedic period onward. In the *Śvetāśvatara Upaniṣad*, Rudra is for the first time called Śiva and is described as the creator, preserver, and destroyer of the universe. His followers are called on to worship him with the same devotion (*bhakti*) as that accorded to Kṛṣṇa-Viṣṇu in the *Bhagavadgītā*. The tendency for the laity to form themselves into religious guilds, or societies—evident in the case of the *yakṣa* cults, Buddhism, and Jainism—promoted the growth of devotional Vaiṣṇavism and Śaivism. These local associations of worshippers appear to have been among the main means of spreading the new cults. Theistic ascetics are less in evidence at this time; but a society of Śaivite monks, the Pāśupatas, was also in existence by the beginning of the Christian Era.

The period between the fall of the Mauryan Empire (*c.* 184 BC) and the rise of the Gupta Empire (*c.* AD 320) was one of great change, with the conquest of most of the area of Pakistan and parts of western India by a succession of invaders. India was opened to influence from the West as never before, not only by its invaders but by way of the sea through the flourishing trade with the Roman Empire. The most obvious effects of the new contacts were shown in art and architecture. The oldest free-standing stone temple in the subcontinent has been excavated at Taxila, near Rāwalpindi, Pakistan. Early in the Christian Era the Gandhāra school of sculpture grew up in the same region and made use of Hellenistic (Greek cultural) and Roman prototypes, mainly in the service of Buddhism. At that time Hindu temples probably were made of wood because no remains of them are known to exist, but literary evidence shows that they were not unknown.

By the time of the early Gupta Empire the new theism had been harmonized with the old Vedic religion, and two of the main branches of Hinduism were fully recognized. The Vaiṣṇavas had the support of the Gupta emperors, who took the title *paramabhāgavata* ("supreme devotee of Viṣṇu"). Viṣṇu temples were numerous and the doctrine of Viṣṇu's *avatāra*s (incarnations) was widely accepted. Of the ten incarnations of later Vaiṣṇavism, however, only two seem to have been much worshipped in the Gupta period. These were Kṛṣṇa, the hero of the *Mahābhārata*, who begins to appear also in his pastoral aspect as the cowherd and flute player, and the divine boar (Varāha), of whom several impressive images survive from the Gupta period.

The Śaivites were also a growing force in the religious life of India. The sect of Pāśupata ascetics, founded by Lakulīśa (or Nakulīśa), who lived early in the Christian Era, is attested by inscriptions from the 5th century and is among the earliest of the sectarian religious orders of Hinduism. The son of Śiva, Skanda (also called Kārttikeya, the war god), figures already on Kuṣāṇa coins around 100. Śiva's other son, the elephant-headed Gaṇeśa, patron deity of commercial and literary enterprises, does not appear until the 5th century. Very important in this period was Sūrya, the sun god, who had temples built in his honour, though in modern times he is little regarded by most Hindus. The popularity of the solar cult may have come from Iranian influence.

Several goddesses began to gain importance in this period. Although goddesses were no doubt always worshipped in local and popular cults, they play comparatively minor roles in Vedic religion. Lakṣmī or Śrī, goddess of fortune and consort of Viṣṇu, was worshipped before the beginning of the Christian Era, and several lesser goddesses are attested from the Gupta period. But the cult of Durgā, the consort of Śiva, was only beginning to gain importance in the 4th century, and the large-scale development of Śāktism (devotion to the active, creative principle, personified as the Mother Goddess) did not take place until medieval times.

Rising importance of goddesses

Interplay of Indo-European and Dravidian cultures. Aryan influence is already in evidence in the earliest Tamil (a principal Dravidian language) literature, perhaps dating from the 1st century AD. At this time in South India the orthodox cults were aristocratic in character, supported by kings and chiefs who gained in prestige by patronizing Brahmins and adopting Aryan ways. The Tamils were still in the main devoted to the old cults, some of which, however, were taking on an Aryan complexion. The pastoral god Murugaṉ was identified with Skanda and his mother, the fierce war goddess Korravai, with Durgā. Varuṇaṉ, a sea god who had adopted the name of the old Vedic god but otherwise had few Aryan features, and Māyoṉ, a black god who was a rural divinity with many of the characteristics of Kṛṣṇa in his pastoral aspect, also are depicted in Tamil literature. The final "Sanskritization" of the Tamils was brought about through the patronage of the Pallava kings of Kāñcīpuram, who began to rule in the 4th century AD and who financed the making of many temples and fine religious sculpture. Similar processes were taking place in the Deccan, Bengal, and other regions.

Establishment of basic doctrines and practices. The texts on conduct known as *Dharma-śāstra* show that by this time the caste system, as distinct from the four classes, was already in existence. The texts attribute the change to the miscegenation of the four original orders of society, a fallacy that was accepted by many earlier students of the Indian social order (see also CASTE SYSTEMS).

The *Smṛti* ("remembered," or traditional) literature, like the *sūtra* literature that preceded it, stresses the religious

merit of gifts to Brahmins. Kings transferred the revenues of villages or groups of villages to Brahmins, either singly or in corporate groups; and thus the status and wealth of the priestly class rose steadily. In the *agrahāras*, as the settlements of Brahmins were called, they were encouraged to devote themselves to the study of the Vedas and the subsidiary studies associated with them; but many Brahmins also developed the sciences of the period, such as mathematics, astronomy, and medicine, and others cultivated literature.

Family law and the *Smṛti* texts

The *Smṛti* texts are binding to this day on the orthodox Hindu, and until quite recently Hindu family law was based on them. By the Gupta period marriage, solemnized by lengthy sacred rites, was virtually indissoluble, though in earlier times there is evidence of divorce. Intercaste marriage was becoming rarer and more difficult, and the whole social life of India was hardening into a rigid pattern. Some of the less attractive features of Hinduism, such as child marriage and the rite of the *satī*, were already in existence, although less frequent than they later became. The earliest definite record of a *satī* burning herself on her husband's pyre is found in an inscription from Eran, Madhya Pradesh, dated 510, but the custom had been followed sporadically long before this. From the 6th century AD onward, such records become frequent.

The Gupta period (4th–6th centuries) saw the rapid development of temple architecture. Earlier temples had been of wood, but now free-standing stone and brick temples appeared in many parts of India. By the 7th century, stone temples, some of considerable dimensions, were to be found in the Aryanized parts of the country. In its origin, the design of the Hindu temple may have owed something to Buddhist precedent, for in some of the oldest temples the image was placed in the centre of the shrine, with an ambulatory path around it resembling the path surrounding the Buddhist *stūpa* (a religious building containing a relic). Nearly all surviving Gupta temples are comparatively small, and consist of a small cella (central chamber), constructed of very thick and solid masonry, with a verandah either at the entrance or on all sides of the building. The very earliest Gupta temples, such as one of the Buddhist temples at Sānchī, have flat roofs, but the *śikhara* (spire), typical of the north Indian temple, began in this period and with time steadily grew higher. The massive and tall tower of the Buddhist temple of Bodh Gayā, which was in existence in the 7th century, represents the culmination of Gupta temple architecture.

The Buddhists and Jains had made use of artificial caves for religious purposes, and these were adapted by the Hindus. Hindu cave-temples, however, are comparatively rare, and none are earlier than the Gupta period. In the Pallava site of Māmallapuram, to the south of Madras, a number of small temples were carved in the 7th century from outcroppings of rock and represent some of the oldest religious buildings in the Tamil country.

The spread of Hinduism in Southeast Asia and the Pacific. Hinduism and Buddhism made an immense impact on the civilizations of Southeast Asia and were the means whereby the peoples of that area emerged into literate civilization. Around the beginning of the Christian Era, Indian merchants in comparatively large numbers settled there, bringing Brahmins and Buddhist monks with them. These religious men were patronized by local chiefs, who became Hinduized. The earliest material evidence of Hinduism in Southeast Asia comes from Borneo, where late 4th-century inscriptions in good Sanskrit testify to the performance of Vedic sacrifices by Brahmins at the behest of local chiefs. Chinese chronicles attest an Indianized kingdom in Vietnam two centuries earlier. The dominant form of Hinduism exported to Southeast Asia was Śaivism, though some Vaiṣṇavism was also known there. Later, from the 9th century onward, Tantrism, both Hindu and Buddhist, spread in the region.

The civilizations of Southeast Asia developed forms of Hinduism and Buddhism that had distinctive local features and were attuned to the local cultures, but the framework of their religious life was essentially Indian. The *Rāmāyaṇa* (an epic centring on the deeds of Rāma) and the *Mahābhārata* stories became well-known in Southeast Asia and are still popular there in local versions. The people of Bali (in Indonesia) still follow a form of Hinduism much adapted to their own genius. Versions of the *Laws of Manu* (a mythical lawgiver) were taken to Southeast Asia and were translated and adapted to indigenous ways until they lost most of their original content.

Claims for early Hindu contacts farther east are more dubious. There is little evidence of the influence of Hinduism on China and Japan, except through Buddhism.

Indian religious influence in the Mediterranean world. Nearly as dubious as the question of Indian influence on the religious life of the Far East is its influence on that of the ancient Mediterranean world. Pythagoras, a Greek philosopher, may have obtained his doctrine of metempsychosis (transmigration, or passage of the soul from one body to another) from India, mediated by Achaemenian (6th–4th centuries BC) Persia, but similar ideas were known in Egypt and were certainly also present in Greece even before the time of Pythagoras. The Pythagorean doctrine of a cyclic universe may also be derived from India, but the only argument is one of resemblance, and the Indian theory of cosmic cycles is not certainly attested in the 6th century BC. Nevertheless, it is clear that Hindu ascetics did occasionally visit Europe at a later time, and the self-immolation of the ascetic Zarmanochegas in Athens (*c.* 20 BC) may have inspired St. Paul to write, "if I deliver my body to be burned, but have not love, I gain nothing" (I Cor.). The most striking resemblance between Greek and Indian thought is in the system of mystical *gnōsis* (esoteric knowledge) described in the *Enneads* of the Neoplatonic philosopher Plotinus (3rd century AD) and that of the *Yoga-sūtras* attributed to Patañjali, an Indian religious teacher sometimes dated in the 2nd century AD. Of these two texts, that of Patañjali is the older, and one cannot but suspect influence, though the problem of mediation remains difficult because Plotinus gives no direct evidence of having known anything about Indian mysticism. Several Greek (*e.g.*, Clement of Alexandria) and Latin writers show considerable knowledge of the externals of Indian religions, but none gives any intimation of understanding their more recondite aspects.

Resemblances between Hinduism and Greek and Christian thought

Certain Vaiṣṇava legends, especially those referring to the infant Kṛṣṇa, bear some resemblance to those of Christianity, and claims have been made by both Hinduism and Christianity that the one influenced the other. In most instances, however, the Christian legends (for example, the massacre of the innocents, with the divine child carried to safety) are attested earlier than their Hindu counterparts.

THE PURANIC PERIOD (4TH–8TH CENTURIES)

The Purāṇas. The period of the Guptas saw the production of the first of the series of lengthy versified texts known as *Purāṇas* ("ancient stories"). These are sectarian, chiefly devoted to the glory of one god or another. With the epics, with which they are closely linked in origin, they became the scriptures of the common man, since they were available to everybody, including women and members of the lowest order of society (Śūdras), and were not, like the Vedas, restricted to initiated men of the three higher orders. The origin of much of their contents may be non-Brahminic, but they were accepted and adapted by the Brahmins, who thus brought new elements into the orthodox religion.

The scriptures of the common man

Tantric traditions and Śāktism. At about this same time, toward the end of the 5th century, the cult of the mother goddess began to achieve a significant place in religious life. Śāktism, the worship of the Śākti, the active power of the godhead conceived in feminine terms, should be distinguished from Tantrism, or Tantricism, the search for spiritual power and ultimate release by means of the repetition of sacred syllables and phrases (*mantra*) and other secret rites. The latter, in a sense, is as old as the Ṛgveda, but it was given special prominence

by the sects that arose about this period. In Hinduism, Tantrism was chiefly associated with Śāktism, the cult of the Mother Goddess.

A temple was erected in honour of the Mothers, a vague group of goddesses related rather artificially to the gods of the Hindu pantheon by being described as their wives, at Gangdhār, Rājasthān, in AD 423. Here, magical rites of a terrifying kind were practiced, for the temple is described as "loud with the shouts of demonesses, crying in the thick darkness." The playwright Bhavabhūti's drama *Mālatī-Mādhava* (about the adventures of the hero Mādhava and his beloved Mālatī), composed in the early 8th century, contains a scene depicting secret rites with human sacrifice and ritual cannibalism. The goddess cults eventually centred around Durgā, the spouse of Śiva in her fiercer aspect. Surviving Hindu *Tantras*, the texts of the movement, are much later than many of those of Tantric Buddhism, and it may be that the Hindus derived much from the Buddhists in this respect. Although there is early evidence of Tantrism and Śāktism in other parts of India, the chief centre of both was modern Bengal, Bihār, and Assam.

Philosophical sūtras and the rise of the six schools of philosophy. Meanwhile, much religious activity of a more intellectual kind was taking place. From about the beginning of the Christian Era through the period of the Gupta Empire, the systems of the "Six Schools" (Ṣaḍdarśana) of orthodox philosophy were formulated in terse *sūtras*.

Śaṅkara and the Vedānta school

The most important of the Six Schools is the Vedānta ("End of the Vedas"), also called Uttara-mīmāṃsa, or later Mīmāmsā. The most renowned philosopher of this school and, indeed, of all Hinduism was Śaṅkara (traditionally dated *c.* 788–820, but he probably died about 20 years later). He was born at Kālaḍi in Kerala and is said to have spent most of his life travelling through India disputing with members of other sects. The Śaṅkaran system has sounded the keynote of intellectual Hinduism down to the present, but later teachers founded subschools of Vedānta, which are perhaps equally important.

Śaṅkara was also responsible for the growth of Hindu monasticism, which had been in existence for well over a millennium, in the form of hermit colonies. Also a few inscriptions from Gupta times onward contain references to orders of Śaivite ascetics, apparently living according to distinctive disciplines and with distinguishing garments and emblems. Śaṅkara founded a closely disciplined Śaivite order, perhaps modelled partly on the Buddhist *saṅgha* (order), known as Daśnāmī, which is still the most influential orthodox Hindu ascetic group. The Order is based in four main monasteries (*maṭha*) at the four corners of India, Śṛṅgerī in Mysore, Badrīnāth in the Himalayas, Dvārakā in Saurāshtra, and Purī in Orissa. The abbots of these monasteries control the spiritual lives of many millions of devout Śaivite laymen throughout India, and their establishments strive to maintain the traditional philosophical Hinduism of the strict Vedānta. In modern times, certain Daśnāmī leaders have incurred criticism for their firm opposition to social change (see also INDIAN PHILOSOPHY; MONASTICISM).

THE RISE OF DEVOTIONAL HINDUISM (8TH–11TH CENTURIES)

The medieval period, especially after the Muslim conquest, saw the growth of new devotional religious movements, centring round hymnodists who taught in the popular languages of the time.

The new movements began, as far as can be seen, with the appearance of hymns in Tamil associated with two groups of poets, the Nāyaṇārs, worshippers of Śiva, and the Ālvārs, devoted to Viṣṇu. The oldest of these can be dated in the early 7th century, though passages of devotional character can be found in earlier strata of Tamil literature. By that century also the religious life of the region had been influenced by the Bhāgavatas, the school of Vaiṣṇavism that based its teachings largely on the *Bhagavadgītā*.

This devotional poetry is characterized by a mystical fervour very different from that of the *Upaniṣads* and the *Bhagavadgītā*, in both of which, even when the object of meditation is conceived as a personal God, there is little expression of emotion. The Tamil "saints," however, felt an intense love (*aṉbu*) of a personal kind toward their god. They experienced overwhelming joy in his presence and deep sorrow when he did not reveal himself. Some of them had a profound sense of guilt or inadequacy in the face of the divine. But the dominant emotion in these poems is one of joy, often expressing itself in song and dance. The poems have a strong ethical content and encourage the virtues of love, humility, and brotherhood. Probably the ideas of these poets, spreading northward, were the origin of the growth of devotional *bhakti* in the centuries following the Muslim invasion of northern India.

The new forms of South Indian devotionalism also produced works in Sanskrit, the most important of which was the *Bhāgavata-Purāṇa* that soon was read all over India. Its tenth book is devoted entirely to Kṛṣṇa; and here, for the first time, the adventures of this god as a lovable child and as a youth sporting with the milkmaids are given great emphasis. The *Bhāgavata-Purāṇa* may have been written in the 10th century and is certainly a product of the Dravidian south. The doctrine of the *avatāras* (incarnations) of Viṣṇu was by now in full force, and the *Bhāgavata* recognizes 22 of them.

Antagonism between devotional cults and Buddhism

The devotional cults helped further to weaken Buddhism, which had long been on the decline. From time to time Hindus, especially Śaivites, took aggressive action against Buddhism. At least two Śaivite kings—the Hūṇa invader Mihirakula (early 6th century) and the Bengal King Śaśāṅka (early 7th century)—are reported to have been active persecutors, destroying monasteries and killing monks. The philosophers Kumārila and Śaṅkara were strongly antagonistic to Buddhism. In their journeys all over India they debated with Buddhists with special vehemence and tried to persuade kings and other influential people to withdraw their support from Buddhist monasteries. Thus, opposition from Hinduism must have had some effect in weakening the strength of the already declining Buddhism. Only in Bihār and Bengal, because of the patronage of the Pāla dynasty and some lesser kings and chiefs, did Buddhist monasteries continue to flourish. Buddhism in eastern India, however, was well on the way to reabsorption into Hinduism when the Muslims invaded the Ganges (Gaṅgā) Valley (12th century). Buddha was recognized as the ninth *avatāra* of Viṣṇu, and thus he could be worshipped by orthodox Hindus without compunction. The great Buddhist shrine of Bodh Gayā, the site of the Buddha's enlightenment, became a Hindu temple dedicated to the *Buddhāvatāra* and remained such until recent times.

At the very end of its existence in India Buddhism developed in a way that had some effect on Hinduism. Among the Buddhist Tantrists appeared a new school of preachers, often known as *siddha*s (those who have achieved), who sang their verses in the contemporary languages, early Maithili and Bengali. They taught that "giving up the world" was not necessary for release from transmigration and that by living a life of simplicity in one's own home one could achieve the highest state. This system, known as Sahajayāna ("the vehicle of the natural," or "the easy vehicle"), influenced on the one hand Bengali devotional Vaiṣṇavism, which produced sects called Sahajiyā with similar doctrines, and on the other the Nātha Yogīs (mentioned below), whose teachings influenced Kabīr and other later Bhakti teachers.

THE AGE OF BHAKTI (11TH–19TH CENTURIES)

The challenge of Islām and popular religion. The phase of Indian history marked by the domination of the Muslims in most of northern India saw great changes in Indian religion. The term *bhakti*, in the sense of devotion to a personal god, appears already in the *Bhagavadgītā* and the *śvetāśvātara Upaniṣad*. In these earlier sources, however, it represents a devotion still somewhat restrained and unemotional. The new form of *bhakti*, associated with hymn singing in the languages of the com-

mon people, on the other hand, was highly charged with emotion, and the relation of divinity and worshipper was often described on the analogy of that of lover and beloved. The advent of Islām in the Ganges (Gaṅgā) Basin at the end of the 12th century meant that royal patronage was withdrawn from Hinduism in much of the area. The attitude of the Muslim rulers toward Hinduism varied. Some, like Fīrūz Tughluq (ruled 1351–88) and Aurangzeb (ruled 1659–1707), were strongly anti-Hindu and enforced payment of *jizya*, the poll tax on unbelievers. Others, like the Bengali sultan Ḥusayn Shāh (reigned 1493–1519) and the great Akbar (reigned 1556–1605), were well-disposed toward their Hindu subjects and the Hindu faith. Numerous temples, however, were destroyed by the more fanatical rulers. Conversions to Islām were more numerous in areas where Buddhism had once been strongest—modern Pakistan, Bangladesh, and Kashmir.

Hinduism on the eve of the Muslim occupation

On the eve of the Muslim occupation, the Hindu religion was by no means sterile in northern India, but its vitality was centred in the southern, Dravidian-speaking areas rather than in the north. Over the centuries, the system of class and caste had become more rigid; in each region there was a complex hierarchy of castes strictly forbidden to intermarry and interdine, controlled and regulated by the secular power on the advice of the court Brahmins. The system was looked on as an integral part of religion and seems to have been accepted without question. Child marriage, polygamy, and widow burning were widespread. The large-scale Vedic sacrifices had practically vanished, but new, simple forms of animal, and sometimes human, sacrifice had appeared, especially connected with the cult of the Mother Goddess.

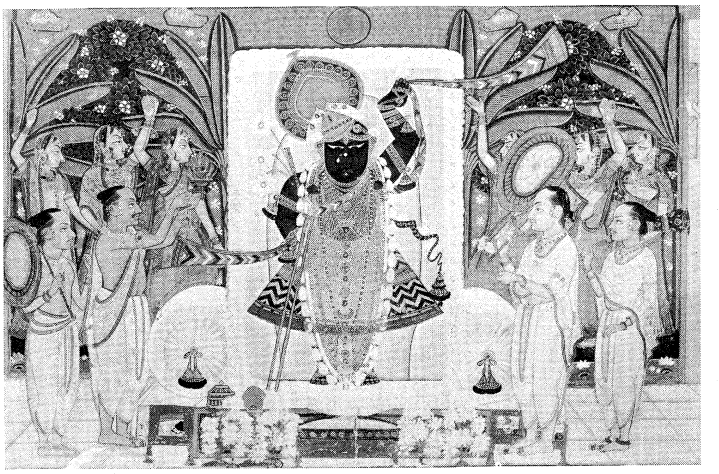
All the divinities of later Hinduism by that time were being worshipped. Rāma, the hero of the epic poem, had become the eighth *avatāra* of Viṣṇu, and his cult was growing, although it was much less prominent than it later became. Similarly, Rāma's monkey helper, Hanumān, now one of the most popular divinities of India and the most ready helper in time of need, was rising in importance. Strange syncretic gods had appeared, such as Harihara, a combination of Viṣṇu and Śiva, and Ardhanārīśvara, a synthesis of Śiva and his *śakti* Pārvatī or Durgā.

From the Gupta period onward temples tended to become larger and more prominent in Hinduism, and their architecture developed in distinctive regional styles. In northern India the finest remaining Hindu temples are to be found in Orissa, as well as at Khajurāho, in northern Madhya Pradesh. The finest example of Orissan temple architecture is the Liṅgarāja temple of Bhubaneswar, built in *c.* 1000. The largest temple of the region, however, is the famous Black Pagoda, the Suntemple of Konārak, built in the mid-13th century. Its 200-foot (60-metre) tower has long since collapsed, and only the assembly hall remains. The most important Khajurāho temples were built during the 11th century. Individual architectural styles also arose in Gujarāt and Rājasthān, but their surviving products are less impressive than those of Orissa and Khajurāho. By the end of the first millennium AD the South Indian style had reached its apogee in the great Rājarājeśvara temple of Thanjāvūr (Tanjore).

Temple design as cosmic symbolism

The designing of Hindu temples, like that of religious images, was codified in the *Śilpa-śāstras* (craft textbooks), and every aspect of the design was believed to be symbolic of some feature of the cosmos. The idea of microcosmic symbolism is very strong in Hinduism and comes from Vedic times; the *Brāhmaṇa* texts are replete with similar cosmic interpretations of the many features of the sacrifice. In this context, the erotic sculptures carved on the outer walls of the towers of some medieval temples, notably at Khajurāho and Konārak, may perhaps be interpreted.

In the temple the god was worshipped by the rites of *pūjā* (reverencing a sacred being or object) on the analogy of serving a great king. In the important temples, a large staff of trained officiants waited on the god. He was awakened in the morning along with his goddess, washed, clothed and fed, placed in his shrine to give audience to his subjects, praised and entertained throughout the day, ceremoniously fed, undressed, and put to bed at night. Worship consisted of hymn singing, burning of lamps, waving of lights before the divine image, and similar acts of homage. The god's dancing girls (*devadāsīs*) would perform before him at regular intervals, watched by the officiants and lay worshippers, who were his courtiers. These women, either the daughters of *devadāsīs* or girls dedicated in childhood, also served as prostitutes. The association of dedicated prostitutes with certain Hindu shrines can be traced back to before the Christian Era and may have existed in the Harappā culture. It became more widespread in post-Gupta times, especially in South India, and aroused the reprobation of 19th-century Europeans. Under the efforts of Hindu reformers the *devadāsīs* disappeared. The role of *devadāsīs* is understandable if the analogy of the temple and the royal court is remembered, for the Hindu king also had his dancing girls, who bestowed their favours on his courtiers.



"Four Priests Doing Puja Before the Image of Śrī Nāthajī," Rajasthani miniature painting from Nāthdwāra, *c.* 1830. In the Fogg Art Museum, Cambridge, Massachusetts.
By courtesy of a private collector, Cambridge, Mass.

The analogy of the royal palace was also in evidence in the *rathayātra*s (shrine processions). As on festival days the king would issue from his palace and parade around his city, escorted by courtiers, troops, and musicians, so also would the god parade around his city in a splendid procession, together with the lesser gods of the minor shrines. The god rode on a tremendous and ornate moving shrine (*ratha*). It was often pulled by large bands of devotees. *Rathayātra*s still take place in many cities of India. The most famous is the annual procession of Jagannātha, a form of Viṣṇu, at Purī, Orissa, which achieved notoriety from the fact that Westerners interpreted the frequent accidental deaths and the occasional sacrificial acts of worshippers, who would throw themselves under the gigantic wheels of his *ratha*, as common occurrences.

The wealth of the great temples

The great temples were (and still are) very wealthy institutions. They were supported by the transfer of the taxes levied by kings on specific areas of the nearby countryside, by donations of the pious, and by the fees of worshippers. Their immense hoarded wealth was one of the factors that encouraged the Ghaznavid and Ghūrid Turks to invade India in the 10th and 11th centuries. They were controlled by self-perpetuating committees—whose membership was usually a hereditary privilege—and by a large staff of priests and temple servants under a high priest who wielded tremendous power and influence. The great walled temple complexes of South India were (and still are) small cities, containing the central and numerous lesser shrines, bathing tanks, administrative offices, homes of the temple employees, workshops, bazaars, and public buildings of many kinds. Directly and indirectly they played an important part in the economy, as they were among the largest employers and greatest landowners in their areas. They performed very

definite social functions because they also served as schools, dispensaries, poorhouses, banks, concert halls, and, very often, even as brothels.

The Muslim occupation brought India into close contact with a different, more aggressive, faith. In such circumstances, the absence of a central religious authority in Hinduism was a source of strength rather than otherwise. The *purohita*s, the family priests who performed the domestic rituals and personal sacraments for the lay people, continued to function, as did the thousands of ascetics. In Muslim-occupied territory the temples were the chief sufferers. In the sacred cities of Vārāṇasī (Banaras) and Mathurā, no large temple remains from any period before the 17th century. The same is true of most of the main religious centres of northern India, but not of the regions where the Muslim hold was less firm, such as Orissa, Rājasthān, and South India.

Philosophical movements. *The qualified Advaita of Rāmānuja.* Before the time the Muslims invaded the subcontinent, the new forms of South Indian Bhakti had begun to flourish and develop and were spreading beyond the bounds of the Dravidian south. Certain Vaiṣṇava theologians of the Pāñcarātra and Bhāgavata schools had long opposed the Advaita Vedānta (nondualism) of Śaṅkara in favour of a more theistic interpretation of the *Upaniṣads* and the *Brahma-sūtras*. The greatest of these philosophers, Rāmānuja, a Tamil Brahmin who was for a time chief priest of the Vaiṣṇava temple of Śrīraṅgam, near Tirucchirāppalli (Trichinopoly), arose in the 11th century. He gave to the growing Vaiṣṇava Bhakti cults a philosophical framework that was developed by other Vaiṣṇava teachers and was not without influence on some Śaivite schools.

Influence of the philosopher Rāmānuja

Dualism and Śaiva philosophies. The most interesting development of Rāmānuja's doctrine of qualified monism is found in the philosophy of Madhva (died *c.* 1276), a Kanarese Brahmin who taught a doctrine of dualism according to which God and the soul are eternally distinct. Between Madhva's system and Christianity the parallels are so striking that it is generally believed that Madhva was in contact with the Syrian Christians of Kerala.

Two other Vaiṣṇava teachers deserve mention. Nimbārka, a Telugu Brahmin of the 12th century, spread the cult of the divine cowherd and his favourite *gopī* (cowherdess, especially associated with the legends of Kṛṣṇa's youth), Rādhā. His sect survives in the vicinity of Mathurā but has made little impact elsewhere. More important was Vallabha (Vallabhācārya; 1479–1531), who took up the Vaiṣṇava doctrine of grace and emphasized its erotic imagery. His sect is noteworthy for the stress it places on absolute obedience to the *guru* (teacher). Early in its existence it was organized with a hierarchy of senior monks (*gosvāmī*), many of whom became very rich. The Vallabhācārya sect was once very influential in the western half of North India, but it declined in the 19th century because of a number of lawsuits against the chief guru, the descendant of Vallabha.

The Śaiva sects also developed from the 10th century onward. In South India there emerged the school of Śaiva-siddhānta, still one of the most significant religious forces in that region, and one which, unlike the school of Śaṅkara, does not admit the full identity of the soul and God. A completely monistic school of Śaivism appeared in Kashmir in the early 9th century. Its doctrines differ from those of Śaṅkara chiefly in attributing personality to the absolute spirit, who is the god Śiva and not the impersonal *Brahman*.

The unorthodoxy of the Liṅgāyats

An important and interesting sect, founded in the 12th century in the Kannada-speaking area of the Deccan, was that of the Liṅgāyats, or Vīraśaivas ("Heroes of the Śaiva faith"). Its traditional founder, Basava, taught doctrines and practices of surprising unorthodoxy—he opposed all forms of image worship and accepted only the *liṅga* of Śiva as a sacred symbol. It rejected the Vedas, the Brahmin priesthood, and all caste distinction. Several Liṅgāyat practices, now largely abandoned, such as the remarriage of widows and the burial of the dead, seem very non-Hindu and indicate a possibility that Islām influenced the doctrines of the sect.

An important development of Śaivism in North India was brought about by Gorakhnāth (Gorakṣanātha), who in the 13th century became leader of a sect of Śaivite ascetics known as Nātha ("Lord") from the title of their chief teachers. The Gorakhnāthīs were particularly important as propagators of the practices of Haṭha Yoga, the form of Yoga that requires complex and difficult physical exercises and which has become well-known in the West. These yogis, who are still numerous, influenced the teaching of several of the Bhakti hymnodists.

Bhakti movements. The hymnodists and "saints" of medieval Bhakti appeared in all parts of India. The origin of the new forms of Hinduism has been attributed to the influence of Islām, but the proposition that the rise of popular emotional Bhakti was a response to Islām is chronologically impossible, for the practice of singing ecstatic hymns in the current local language was well-known in South India even before Muḥammad. All the features of this form of Bhakti are to be found in the *Bhāgavata-Purāṇa* and in the commentaries of Rāmānuja. The earliest Bhakti literature in a living Indo-Aryan language is from Mahārāshtra and was composed before Muslims occupied the area. Thus, emotional Bhakti existed long before the Muslim conquest. The presence of rulers of alien faith and the withdrawal of royal patronage from the temples and Brahminic colleges may have encouraged the spread of new, more popular forms of Hinduism. The psychological effect of the Muslim conquest also may have predisposed the masses to accept the simple teachings of the hymnodists, but Islām was only a contributory factor in the spread of the new movements.

Northern India. In the Hindi-speaking areas the development of the new religious forms was stimulated by Rāmānand (*c.* 1400–70). This teacher, in the line of Rāmānuja, made Vārāṇasī his headquarters, and there established a sect solely devoted to Rāma; the Rāmānandīs rejected caste and image worship. They admitted even untouchables as members and conveyed their message entirely in the contemporary language.

The most important spiritual descendant of Rāmānand was Tulsīdās, considered by many to be the greatest of Hindi poets. Tulsīdās, a Brahmin ascetic who spent most of his life at Vārāṇasī and died there in old age in 1623, produced the *Rāmcaritmānas* (or *Ramacaritmānas*, "The Lake of the Deeds of Rāma"), the most important text of Hinduism in the Hindi-speaking area. Tulsīdās' theology is simple, emphasizing divine love and compassion.

The schools of Nimbārka and Vallabhācārya produced a line of Kṛṣṇaite Bhakti singers in Hindi, the first of whom was the poetess Mīrā Bāī, a Rājput princess living in the earlier half of the 16th century. She composed hymns in honour of Kṛṣṇa in Rajasthani, and these were translated into various dialects of Hindi and Gujarati and are now sung all over western India.

Eastern India. In Bengal, the decline of Buddhism and the coming of the Muslims resulted in developments in folk religion. There seems to have been a certain loss of faith in the great gods in the period after 1200, and several local divinities, no doubt long worshipped by peasants and the poor, became popular with the richer and more educated. Among these are Manasā (the goddess of snakes); a divinity strangely called Dharma ("Law"); and a Śiva who, far from being the exalted ascetic deity of the myths of the *Purāṇas*, is sometimes depicted as a simple peasant farmer with a nagging wife.

Influence of Caitanya in Bengal

The various strands of Bengali Vaiṣṇavism were drawn together by the teacher Viśvambhara Miśra, generally known by his religious name, Caitanya (1485–1533), who established a sect that has given Bengali religious life much of its individual character. Caitanya gave up the world after an intense religious experience at the age of 22, and the band of friends and followers who surrounded him virtually became a religious order. The Caitanya sect is not particularly remarkable for its theology but rather for its practices. The focus of its religious life is the *saṃkīrtana*, a hymn-singing session (often accompanied by ecstatic dancing) conducted by a group of worshippers and sometimes continued in relays

for days on end. The *saṃkīrtana* also may take the form of a procession through the streets (*nagara-kīrtana*). With its discouragement of ritualism, its strong ethical emphasis, and its joyful and expressive method of worship, the Caitanya movement affected the whole life of Bengal and was not without influence on other parts of India.

Although the Caitanya movement may have tended to relax caste barriers, the complex Bengali system of subcastes, known as Kulīnism, persisted, as did the Śākta cults, often involving animal sacrifice and in some cases ritual sexual activity. Widow cremation was particularly prevalent in Bengal in the early 19th century.

Southern and western India. Mahārāshtra, closely in touch with the Dravidian south, was the first region where an Indo-Aryan language was spoken to produce literature inspired by the Bhakti movement. Its first important literary figure is Jñāneśvara (13th century), a teacher who was born in a Brahmin family that had lost status and who early came under the influence of the Nātha Yogīs and became an ascetic. His most important achievement is his long Marathi paraphrase and commentary on the *Bhagavadgītā*, commonly known as *Jñāneśvarī*, after its author. In other parts of India, until the end of the 19th century, the *Bhagavadgītā* was chiefly studied by the learned, with the commentaries of Śaṅkara, Rāmānuja, and others; but among Marathi speakers it was well known to all classes through Jñāneśvara's version.

The greatest medieval Maharashtrian teacher and, according to many critics, the greatest Bhakti poet of all India, was Tukārām (17th century). Tukārām was a Śūdra, the son of a poor village grain dealer, who took to a life of religion after bereavement and failure in business.

Bhakti teachings in all of India's main languages

All the main languages of India are represented among important Bhakti teachers. Much of the life of Caitanya was spent at the sacred city of Purī in Orissa, and for this reason most Oriya Bhakti shows the influence of his teaching. The greatest of the Assamese Bhakti reformers, Śaṅkardeb (Sanskrit, Śaṅkaradeva, traditional dates 1449–1569), spread Vaiṣṇavism of a type similar to that of Caitanya in the Brahmaputra Valley and produced an Assamese paraphrase of the *Bhāgavata-Purāṇa* as well as shorter devotional poems. Kashmir produced an important religious poet in Lallā, or Lāl Ded, a woman ascetic devotee of Śiva whose verses are still very popular among the Kashmir peasantry and are repeated even by Muslims. There is much Bhakti literature in Gujarati, the most famous being the hymns of Narsiṃha Mehtā (1415–81).

In the literature of the Dravidian languages other than Tamil, the Bhakti movement was largely Śaiva. Vīraśaiva devotional literature in Kannada goes back to Basava, the founder of the sect. In Telugu the most interesting figure is Vemana (15th century), a Śaivite teacher of low birth, who attacked the caste system and the Brahmins in particular with a vehemence rare in all Indian literature and who was almost equally antagonistic to iconolatry (worship of images), pilgrimage, and other aspects of ritualism.

The synthesis of Hinduism with Islām and folk religion. Approximately contemporary with Vemana was Kabīr (1440–1518), who was probably born a Muslim. Kabīr rejected many of the external forms and doctrines of both Hinduism and Islām—caste, image worship, asceticism, divine incarnation, sacred texts, and pilgrimages. He believed, however, in transmigration and the doctrine of *karma;* he regularly referred to God as Rāma and to the phenomenal world as *māyā* (illusion); and he admitted the qualified reality of the lesser gods of Hinduism. Kabīr probably knew something about Ṣūfism (Islāmic mysticism), which had itself been influenced by Hinduism, but his teaching is essentially Hindu and it is hard to recognize any specifically Islāmic element in it, except perhaps his opposition to image worship and to strict asceticism, although both had also been part of the teaching of some earlier Hindu reformers. Kabīr's doctrine was not a syncretism of Hinduism and Islām but rather a simplified form of Hinduism, by which he hoped to foster brotherly love between the two communities. His poems are popular all over the Hindi-speaking area, among Hindus and Muslims alike.

A small sect, the Kabīrpanthīs, looks back to Kabīr as its founder, but its importance is less than that of the vigorous new religion (Sikhism) founded by one of Kabīr's disciples, Nānak. In its final form, Sikhism contains elements taken from Islām (equality in the faith, opposition to iconolatry, extreme reverence for the sacred book) and probably also from Christianity (the Sikh Baptism and Communion meal), but its theology is still essentially Hindu.

Much has been said about the synthesis of Hinduism and Islām in the period of Muslim dominance, but, as far as the Hindus were concerned, this was generally in the matter of superficial observances. Thus, *parda*, the strict seclusion of women, became usual among the Hindu upper classes of northern India, and numerous Muslim social customs were adopted, as were Persian and Arabic words in the vocabularies of most Indian languages. The fundamental theology of Hinduism, however, was unaffected by Islām, even in the teachings of such men as Basava and Kabīr, who themselves may have been somewhat influenced by Muslim observances and social customs.

The Hindu influence on Islām

What synthesis did take place came from the Muslims, most of whom were Indian by blood. In Hindi, Bengali, Gujarati, Punjabi, and Marathi there is much poetic literature, written by Muslims and commencing with the Islāmic invocation of Allāh, which nevertheless betrays strong Hindu influence. Thus, there are texts that proclaim Kṛṣṇa as being in the line of the prophets of Islām and as the teacher of the unity of God. Much mystical poetry, though written by authors with Muslim names, uses Hindu imagery and Hindu terminology. The origin of this literature is to be found in the accommodating character of early Indian Ṣūfism, which well before Kabīr proclaimed that Muslim, Christian, Jew, Zoroastrian, and Hindu were all striving toward the same goal and that the outward observances that kept them apart were false. Some of the Indian Ṣūfīs were much influenced by Hindu customs. A school of Kashmir Ṣūfīs, whose members call themselves *rishis*, after the legendary Hindu sages (*ṛṣi*), respect and repeat the verses of Lāl Ded and are strict vegetarians.

Syncretic tendencies were encouraged by tolerant Muslim rulers, and these tendencies reached their zenith in the reign of Akbar (1556–1605), who took a great interest in the religion of his Hindu subjects, favoured vegetarianism, and tried to establish a single, all-embracing religion for his empire. Although the efforts of Akbar failed, they influenced India for more than 50 years after his death. But the orthodox Muslim theologians had long been complaining about the growth of heresy, and the emperor Aurangzeb (reigned 1659–1707) did all in his power to discourage it. Popular Muslim preachers throughout the 18th and 19th centuries worked to restore orthodoxy. Thus, syncretic tendencies virtually came to an end before the imposition of British power in the mid-18th century. Furthermore, British rule, emphasizing the distinctions between Hindu and Muslim, was not calculated to encourage the revival of efforts to harmonize the two religions.

THE MODERN PERIOD (19TH–20TH CENTURIES)

Western Christian missionary activities

From their small coastal settlements in southern India, the Portuguese had promoted Catholic missionary activity and had made converts, most of whom were of low caste; the mass of caste Hindus was unaffected. Small Protestant missions operated from the Danish factories of Tranquebar in Tamil nadu and Serampore in Bengal, but they were even less influential. The British East India Company, conscious of the disadvantages of unnecessarily antagonizing its Indian subjects, excluded all Christian missionary activity from its territories. Indeed, the Company continued the patronage accorded by indigenous rulers to many Hindu temples and positively forbade its Indian troops to embrace Christianity. The growing evangelical conscience in England brought this

policy to an end with the renewal of the Company's charter in 1813. The Company's policy then became one of strict impartiality in matters of religion, and missionaries were allowed to work throughout its territory. Thus, Christian ideas began to spread.

Hindu reform movements. *Brahmo Samaj.* The pioneer of reform was Rammohan Ray. His intense belief in strict monotheism and in the evils of image worship began early and was probably derived from Islām, because at this time he had no knowledge of Christianity. He later learned English and in 1814 settled in Calcutta, where he was prominent in the movement for encouraging education of a Western type. His final achievement was the foundation of the Brahma Sabha, later known as Brahmo Samaj ("Society of God"), in 1828.

Rammohan outwardly remained a Hindu, wearing the sacred cord and keeping up most of the taboos of the orthodox Brahmin; but his theology was surprisingly un-Indian. He was chiefly inspired by 18th-century Deism (rational belief in a transcendent creator god) and Unitarianism (belief in God's essential oneness), but some of his writing suggests that he was aware of the religious ideas of the Freemasons (a secret fraternity that espouses some deistic concepts). Several of his friends were members of the active Masonic lodge in Calcutta. His ideas of the future life are obscure, and it is possible that he did not believe in the doctrine of transmigration. Rammohan was one of the first higher class Hindus to visit Europe, where he was much admired by the intelligentsia of Britain and France. He died in Bristol, England.

Debendranath Tagore and Keshab Chunder Sen

After Rammohan Ray's death, Debendranath Tagore became leader of the Brahmo Samaj, and under his guidance a more mystical note was sounded by the society, nearer to the main stream of Indian religious life. The third great leader of the Brahmo Samaj, Keshab Chunder Sen, was a more vigorous reformer and introduced the complete abolition of caste in the *samāj* and the admission of women. As his theology became more and more syncretistic and eclectic, a schism developed, with the more conservative faction remaining under the leadership of Tagore. Keshab's faction, the Brahmo Samaj of India, adopted as its scripture a selection of theistic texts gathered from all the main religions; at the same time, it became more Hindu in its worship, employing the *saṃkīrtana* (hymn-singing session) and *nagara-kīrtana* (street procession) of the Caitanya sect. In 1881 Keshab founded the Church of the New Dispensation (Naba Bidhan) with the purpose of establishing the truth of all the great religions in an institution that, he believed, was to replace them all. When he died in 1884, the Brahmo Samaj began to decline, but it produced the greatest poet of modern India, Rabindranath Tagore, son of the second of its great leaders, Debendranath Tagore.

Arya Samaj. A reformer of different character was Dayanand Sarasvati, who was trained as a yogi but steadily lost faith in yoga and many other aspects of Hinduism. After travelling widely as an itinerant preacher, he founded the Arya Samaj in 1875, and it rapidly gained ground in the west of India. Dayanand rejected image worship, sacrifice, and polytheism and based his doctrines firmly on the four Vedas as the eternal word of God. Later Hindu scriptures were to be judged on their merits, and many of them were believed to be positively evil. The Arya Samaj did much to encourage Hindu nationalism, and among its members was the revolutionary Lala Lajpat Rai. It did not disparage the knowledge to be found in the West, and it established many schools and colleges.

New religious movements. *Ramakrishna Mission.* The most important developments in Hinduism, however, did not arise primarily from the new *Samaj*es. The mystic Ramakrishna, who was a devotee at a temple of Kālī called Dakṣiṇeśvar to the north of Calcutta, attracted a band of educated lay followers who spread his simple doctrines. As a result of his psychological experiments, he came to the conclusion that "all religions are true." Nevertheless, the religion of a man's own time and place was for him the best expression of the truth. Even idolatry met the needs of simple people and was not to be disparaged. Ramakrishna thus gave the educated Hindu a basis on which he could accept the less rational aspects of his religion without believing in them literally.

Influence of the mystic Ramakrishna

Among the followers of Ramakrishna was Narendranath Datta, who became an ascetic after his master's death, assuming the religious name Vivekananda. In 1893 he attended the World's Parliament of Religions at Chicago, where his powerful personality and stirring oratory made a very deep impression. After lecturing in the United States and England, he returned with a small band of Western disciples to India in 1897. There he founded the Ramakrishna Mission, the most important modern organization of reformed Hinduism. Vivekananda, more than any earlier Hindu reformer, encouraged social service and the uplift of the masses. Influenced by progressive Western political ideas, he set himself firmly against all forms of caste distinction and fostered a spirit of self-reliance in his followers. The Ramakrishna Mission has done much to spread a knowledge of Hinduism outside India and now has branches in many parts of the world.

Theosophical Society. Another neo-Hindu movement, which at one time exerted considerable influence, is the Theosophical Society. Founded in New York City in 1875 by Helena Blavatsky of Russia, its original inspiration was Kabbala (Jewish esoteric mysticism), Gnosticism (esoteric salvatory knowledge), and other forms of Western occultism. When Madame Blavatsky came to India in 1879, her doctrines quickly took on an Indian character, and from her headquarters at Adyar she and her followers established branches in many cities of India.

The society survived serious accusations of charlatanry levelled against its founder and certain other leaders, and it reached the peak of its influence under its next important leader, Annie Besant, a reform-minded English woman. Under her guidance, many Theosophical lodges were founded in Europe and the United States, and these helped to acquaint the West with the principles of Hinduism, if in a rather garbled form.

Aurobindo Ashram. Another modern teacher whose doctrines have had some influence outside India was Sri Aurobindo, who began his career as a revolutionary. He withdrew from politics, however, and settled in Pondicherry, then a French possession. There he established an *āśrama* (a retreat) and achieved a high reputation as a sage. His followers looked on him as the first incarnate manifestation of superbeings whose evolution he prophesied, and apparently he did not discourage this belief. After his death, the leadership of the Aurobindo Ashram was taken over by "the Mother," Mme Mira Richard, a Frenchwoman who had been one of his leading disciples.

Other reform movements. Numerous other teachers have affected the religious life of modern India. Among them was the great Bengali poet, Rabindranath Tagore, who was influenced by many currents of earlier religious thought, both Indian and otherwise. Tagore was particularly popular in Europe and America around the time of World War I, and he did much to enlist sympathy for Hindu religious thought in the West.

Less important outside India, but much respected in India itself, especially in the Dravidian south, was Ramana Maharishi, a Tamil mystic (died 1950) who maintained almost complete silence. His personality was powerful and he attracted a large band of devotees.

Swami Sivananda, who had been a doctor, established an *āśrama* and an organization called the Divine Life Society near the sacred site of Rishikesh in the Himalayas. This organization has numerous branches in India and some elsewhere. His movement teaches more or less orthodox Vedānta, combined with both yoga and *bhakti*, but rejects caste and lays stress on social service.

Jiddu Krishnamurti represents the most attenuated form of Hinduism. He rejects all religious organization and makes no claim to special revelations or exceptional spiritual development but teaches self-realization through introspection and the abandonment of personal ambition.

The struggle for independence. The Hindu revival and reform movements of the 19th and early 20th centuries were closely linked with the growth of Indian nationalism and the struggle for independence. The Arya Samaj strongly encouraged nationalism, and even though Swami

Religion and the growth of nationalism

Vivekananda and the Ramakrishna Mission were always uncompromisingly nonpolitical, their effect in promoting the movement for self-government is quite evident.

Religion and politics were joined in the career of B.G. Tilak, an orthodox Maharashtrian Brahmin who realized that the masses of India could only be aroused by appeals couched in religious terms. Tilak used the annual festival of the god Gaṇapati (Gaṇeśa) for nationalist propaganda. His view of the *Bhagavadgītā* as a call to action was also a reflection of his nationalism, and through his mediation the *Bhagavadgītā* became a stimulus to later leaders, including Mahatma Gandhi.

In Bengal, also, Hindu religious concepts were enlisted in the nationalist cause. In his historical novel *Ānandamaṭh*, the Bengali novelist Bankim Chandra Chatterjee described a band of martial ascetics at the time of the decline of the Mughal Empire, who were pledged to free India from Muslim domination. These had as their anthem a stirring devotional song written in simple Sanskrit—"Bande Mātaram" ("I revere the Mother"). The Mother referred to is at once the stern demon-destroying goddess Kālī and a personification of India. This song was soon taken up by the more extreme nationalists. Vivekananda emphasized the need to turn the emotion of *bhakti* toward the suffering poor of India. During his short career as a young revolutionary leader, Sri Aurobindo made much use of "Bande Mātaram," and called on his countrymen to strive for the freedom of India in a spirit of devotion. The *bhakti* of the medieval hymnodists was thus enlisted in the cause of modern independence.

Mahatma Gandhi. Much influenced by the traditional *bhakti* of his native Gujarāt and fortified by Christian and other religious literature that encouraged similar attitudes, Mahatma Gandhi, the most important leader for independence, appeared to his simpler followers as the quintessence of the Hindu tradition. His austere celibate life was one that the Indian layman had learned to respect implicitly. Thus, Gandhi's message reached a wider public than that of any of the earlier reformers.

The Western element in Gandhi's ideology has often been exaggerated. His doctrine of nonviolence can be found in many Hindu sources, although his beliefs were much strengthened by Christian ethical literature and especially by the later writings of Leo Tolstoy. His political technique of passive resistance, *satyāgraha*, also has Indian precedents, although in this he was much influenced by Western writers such as the American Henry David Thoreau. The chief innovations in Gandhi's philosophy were the dignity of manual labour and the equality of women. Precedents for both of these are to be found in the writings of some 19th-century reformers, but they have little basis in earlier Indian thought. In many ways Gandhi was a traditionalist. His respect for the sacred cow—which, as some other educated Indians did, he rationalized as the representative of Mother Earth—was a contributory factor in the failure of his movement to attract large-scale Muslim support. His insistence on strict vegetarianism and celibacy among his disciples, in keeping with the traditions of Vaiṣṇava ascetic ethics, also caused difficulty among some of his followers. Still, the success of Gandhi represented a political culmination of the movement of popular Bhakti begun in South India early in the Christian era.

Gandhi as innovator and traditionalist

The religious situation after independence. The rise of nationalism, after the division of India into India and Pakistan in 1947, led to a widening of the gulf between Hindu and Muslim. In the early 1970s it was fashionable in Indian circles to paint the relations of the two religions in earlier centuries as friendly, blaming alien rule for the division of India. In Pakistan the tendency has been to insist that Hindus and Muslims have always been "two nations," even though the Hindus were happy under their Muslim rulers. Neither position is wholly correct. In earlier times there was much mutual influence. But the conservative and rigid moralistic element in Indian Islām was gaining the upper hand long before British power was consolidated in India, and Islāmic influence on Hinduism was largely in externals.

Among the pioneers of nationalism, Tilak glorified the Maharashtrian hero Śivajī as the liberator of his country from the alien yoke of the Mughals; and Bankim Chandra Chatterjee's militant ascetics, pledged to conquer and expel the Muslims, sang a battle hymn that no orthodox Muslim could repeat. British rulers of India did little or nothing to lessen Hindu–Muslim tension, and their policy of separate electorates for the two communities worsened the situation. Many leaders of the Indian National Congress movement, such as Jawaharlal Nehru, carried their Hinduism lightly and favoured a secular approach to politics. The majority, however, followed the lead of Gandhi, whose insistence on Hindu values discouraged Muslims from joining his movement, despite the fact that he recited passages from the Qur'ān as well as from the Hindu and Christian scriptures at his prayer meetings. To the right of the Congress politically, the Hindu Mahasabha was equally nationalistic, but its explicitly Hindu nationalism was not opposed to nonviolence in its drive to establish a Hindu state in India.

The transfer of power in 1947 was accompanied by slaughter and pillage of huge proportions. Literally millions of Hindus left their homes in Pakistan for India, and as many Muslims migrated in the opposite direction. The tension culminated in the assassination of Gandhi by a Hindu fanatic in January 1948.

India as a secular state

The policy of the Indian government was to establish a secular state, and the successive Congress governments have broadly kept to this policy. The governments of the Indian states, however, have not been so restricted by constitutional niceties. Some state governments have introduced legislation of a specifically Hindu character.

On the other hand, the Congress government has passed legislation offensive to Hindu traditional prejudices such as no British Indian government would have dared to enact. All forms of discrimination against "untouchables" are now forbidden, although it has been impossible to enforce this law in every case. A great blow to conservatism was dealt by legislation of 1955–56 which gave full rights of inheritance to widows and daughters, enforced monogamy, and permitted divorce on quite easy terms. A further enactment that tended to undermine traditional Hinduism was the law forbidding dowries (1961). The dowry has long been a tremendous burden to the parents of daughters, but the strength of social custom is such that this law cannot be fully enforced.

The mantle of Mahatma Gandhi fell on Vinoba Bhave, one of his most devoted Maharashtrian supporters. For some years after independence Vinoba led a campaign of social service that culminated in the *bhūdān* (land-giving) movement, which persuaded many landowners and wealthy peasants to give fields to landless labourers. This movement had some small success in rural areas, but it gradually lost momentum.

Although the memory of Gandhi is revered by most Indians in the latter part of the 20th century, his policies and principles carry little weight. The great bulk of social service is performed by government agencies rather than by voluntary bodies, whether Gandhian or otherwise. The social structure of traditional Hinduism is slowly crumbling in the cities. Intercaste and interreligious marriages are becoming more frequent among the educated, though some aspects of the caste system show remarkable vitality, especially in the matter of appointments and elections. The bonds of the tightly knit Hindu joint family are also weakening, a process helped by recent legislation and the emancipation of women. The professional priests, who perform rituals for laymen in the home or at temples and sacred sites, complain of lack of custom, and their numbers are diminishing. Mythological films, once the most popular form of entertainment, are now less often made.

Nevertheless, Hinduism is far from dying. Organizations such as the Ramakrishna Mission flourish and expand their activities. New teachers appear from time to time and attract considerable followings. The Jan Sangh, a right-wing party devoted to maintaining Hindu values, gains a considerable number of seats in every election, especially in Uttar Pradesh. The adaptability of Hinduism to changing conditions is illustrated by the

The appearance of a new deity: Santoṣī Mātā

appearance in the Hindu pantheon of a new divinity, of special utility in an acquisitive society. This is the goddess Santoṣī Mātā, now worshipped widely by women in many cities of Uttar Pradesh. The new goddess was unheard of a few years ago and has no basis in any Purāṇic myth. Propitiated by comparatively simple and inexpensive rites performed in the home without the intervention of a priest, Santoṣī, it is believed, will grant practical and obvious blessings, such as a promotion for a needy, overworked husband, a new radio, or even a refrigerator. News of Santoṣī's blessings is passed from housewife to housewife, and, against all reason, even moderately well-educated women have become her devotees.

On both the intellectual and the popular level, Hinduism is thus in the process of adapting itself to new values and new conditions that have been brought about by mass education and industrialization and is responding to 20th-century challenges.

Hinduism outside India. During the latter part of the 19th and through the 20th centuries, large colonies of Hindu migrants have been established in East Africa, Malaysia, the islands of the Pacific and the Indian Ocean, and some of the islands of the West Indies. These migrants have taken their religion with them and have adhered to it faithfully over several generations. In recent years they have been aided by Hindu missionaries, chiefly of the Arya Samaj or the Ramakrishna Mission. Since World War II many Hindus have also settled in the United Kingdom. Most of these migrants, however, have been comparatively uneducated. Their religion has made little impression on the people among whom they live, and they themselves have made no serious attempts to gain converts. Yet, one of the most striking aspects of contemporary Western culture is its readiness to accept Eastern religious ideas in a way unprecedented since the days of the Roman Empire. The most recent manifestation of the spread of Indian religious attitudes in the Western world is the Hare Krishna cult, officially known as the International Society of Krishna Consciousness Inc., with its head office in Los Angeles. This is essentially a *bhakti* movement, broadly following the precedents of Caitanya. Since its foundation by a Hindu *sannyāsī*, Swami Prabhupada, in 1965, its growth has been surprising, and *nagar-kīrtans* (devotional singing and dancing) may now be seen in the streets of New York and London, performed by young men and women from Christian or Jewish homes wearing *dhotī*s and *sarī*s. These manifestations of the spreading influence of Hinduism are part of a process that began in 1784 with the first English translation of a Hindu religious text, Charles Wilkins' version of the *Bhagavadgītā*. Even if modern forms of Hinduism have only an ephemeral existence, new manifestations most likely will be seen in the West. Hinduism is now a part of the cultural heritage of the world.

BIBLIOGRAPHY. Among the many general studies of Hinduism treating the subject briefly in historical perspective, perhaps the best is R.C. ZAEHNER, *Hinduism*, 2nd ed. (1966). The monumental R.C. MAJUMDAR (ed.), *The History and Culture of the Indian People*, 10 vol. (1951–), contains sections on religion and philosophy in each volume, giving useful general surveys of the subject at different periods. BENJAMIN WALKER, *The Hindu World*, 2 vol. (1968), is an alphabetically arranged encyclopaedia of Hinduism that pays due attention to its historical aspects. On the archaeology and anthropology of religion in India, B. and R. ALLCHIN, *The Birth of Indian Civilization* (1968), contains a valuable chapter on the subject and useful bibliographical references. D.D. KOSAMBI, *Myth and Reality*, ch. 2–4 (1962), and *The Culture and Civilization of Ancient India in Historical Outline* (1965), are perhaps in places overimaginative but stimulating. The best general survey of Hinduism as practiced by the masses is probably still L.S.S. O'MALLEY, *Popular Hinduism: The Religion of the Masses* (1935). The classical case history of the process of Sanskritization is M.N. SRINIVAS, *Religion and Society Among the Coorgs of South India* (1952). The standard survey of Vedic religion is A.B. KEITH, *The Religion and Philosophy of the Veda and Upanishads*, 2 vol. (1925). More recent are the works of J. GONDA, especially *The Vision of the Vedic Poets* (1963) and *Change and Continuity in Indian Religion* (1965). For the evolution of theistic Hinduism, the classic survey of NICOL MACNICOL, *Indian Theism from the Vedic to the Muhammadan Period* (1915), is still important. For a more recent interpretation, see S. CHATTERJI, *The Evolution of the Theistic Sects in Ancient India* (1962). E.W. HOPKINS, *Epic Mythology* (1915), is still the best general survey of the subject. Recent more specialized historical studies are SUVIRA JAISWAL, *The Origin and Development of Vaiṣṇavism* (1967); and V.S. PATHAK, *History of the Śaiva Cults in Northern India* (1960). On the history of South Indian Śaivism, see C.V. AYYAR, *Origin and Early History of Śaivism in South India* (1936). Many studies of the Hindu temple have been published, but most of these are mainly concerned with its architecture and sculpture, though a few also give accounts of temple ritual. Interesting and important is the article of B. STEIN, "The Economic Function of a Medieval South Indian Temple," *Journal of Asian Studies*, 19:163–175 (1960). On medieval devotional Hinduism, see J.E. CARPENTER, *Theism in Medieval India* (1921); LILIANE SILBURN, *La Bhakti* (1964); and M.B. SINGER (ed.), *Krishna: Myths, Rites and Attitudes* (1966). A remarkable survey by the early 19th-century Catholic missionary J.A. DUBOIS, *Hindu Manners, Customs and Ceremonies*, 3rd ed. (1906, reprinted 1959), though betraying prejudice here and there, is one of the very few detailed and reliable studies of living Hinduism as it existed before it came under any Western or other reforming influence. A good overall survey of the developments in Hinduism in the last 150 years is yet to be written. J.N. FARQUHAR, *Modern Religious Movements in India* (1915), still forms a useful survey of developments down to the date of publication, though written from a missionary viewpoint. See also D.S. SARMA, *Studies in the Renaissance of Hinduism in the Nineteenth and Twentieth Centuries* (1944). On 19th-century Bengali reformers, see N.S. BOSE, *The Indian Awakening and Bengal*, 2nd ed. rev. (1969). The EARL OF RONALDSHAY, *The Heart of Aryâvarta: A Study of the Psychology of Indian Unrest* (1925), is a contemporary study of the religious bases of Hindu nationalism, written with considerable understanding, though by one of the British rulers of India. The short work of C.E.M. JOAD, *Counter Attack from the East: The Philosophy of Radhakrishnan* (1933, reprinted 1951), was perhaps among the first to recognize the extent of the influence of Hinduism on the West. The collection of articles edited by M. SINGER, *Traditional India: Structure and Change* (1958), contains several dealing with the recent religious history of Hinduism. R.W. SCOTT, *Social Ethics in Modern Hinduism* (1953), examines recent trends.

(A.L.B.)

Hindu Kush (Mountains)

The Hindu Kush is one of the great watersheds of Central Asia, forming part of the vast alpine zone that stretches across the continent from east to west. Broadly defined, it is a mountain system nearly 1,000 miles (1,600 kilometres) long and possibly 200 miles (320 kilometres) wide, running northeast to southwest, and dividing the valley of the Amu Darya (the ancient Oxus River) to the north from the Indus River Valley to the south. There is no agreement among geographers as to its precise boundaries. To the east, the Hindu Kush buttresses the Pamir Plateau near the Chinese border, after which it runs southwest through Pakistan and into Afghanistan, finally merging into minor ranges in western Afghanistan. The highest peak is Tirich Mīr, which is 25,263 feet (7,700 metres) high.

Historically, the passes across the Hindu Kush have been of great military significance, providing access to the northern plains of India. Alexander the Great (*q.v.*), king of Macedonia, who passed over the Hindu Kush in the 4th century BC, was among those who invaded India by this route. During the period of British rule in India, the Indian government was keenly concerned with the security of these passes, and more especially with their own control of an associated physical feature to the south, the Khyber Pass. The Hindu Kush range has rarely constituted the frontier between major powers, but has usually formed part of an intermediate buffer zone.

The name Hindu Kush first appears in 1333 in the writings of Ibn Baṭṭūṭah, the medieval Berber traveller, who said that the name meant "Hindu killer," a meaning still given by Afghan mountain dwellers, who are traditional enemies of Indian plainsmen. More likely the name is a corruption of the classical term Hindu-Caucasus, or else Hindu-Koh, meaning "Indian Mountains."

The mountains. The eastern limit of the Hindu Kush is difficult to determine because of a locally complex

topography, although the Karambar Pass (14,225 feet, or 4,345 metres) between the valleys of Chitrāl and Gilgit may be tentatively accepted as marking the boundary. The western limit is still more uncertain, as the mountains lose height and fan out into minor ranges. The Kirmu Pass (10,879 feet, or 3,316 metres) to the west of Kābul may, nevertheless, be regarded as indicating the approximate boundary. Geologists, however, consider the Hindu Kush range to extend much further west not only into Afghanistan but also into Iran.

Three main regions

Three main sections of the Hindu Kush may be defined. These are the eastern Hindu Kush, which runs from the Karambar Pass in the east to the Dorah Pass (14,940 feet, or 4,554 metres) not far from Tirich Mīr; the central Hindu Kush, which then continues to the Khawak Pass (11,640 feet, or 3,548 metres) to the north of Kābul; and the western Hindu Kush, also known as the Koh-i-Bāba, which gradually descends to the Kirmu Pass.

The Eastern Hindu Kush. In its extreme eastern section, between the passes of Karambar and Baroghil (12,480 feet, or 3,804 metres), this region is not very high and has mountains that often take the form of rounded domes. Further to the west the main ridge rises rapidly to Baba Tangi (21,368 feet, or 6,513 metres) and becomes rugged, after which, within the space of about 100 miles, are concentrated the highest mountains of the entire region—about two dozen summits of more than 23,000 feet, or 7,000 metres, in height. A first cluster of high peaks around Urgend (23,094 feet, or 7,039 metres) is followed further south by the massif (principal mountain mass) of Sara Ghar (24,111 feet, or 7,349 metres). Another line of imposing mountains, which includes Koh-i-Langar (23,162 feet, or 7,060 metres), Shachaur (23,346 feet, or 7,116 metres), Udrem Zom (23,376 feet, or 7,125 metres), and Nader Shah (23,376 feet, or 7,125 metres), leads to the three giant mountains of the Hindu Kush, which are Noshaq (24,580 feet), Istor-o-Nal (24,242 feet), and Tirich Mīr (25,263 feet). Most major glaciers of the Hindu Kush—among them Kotgaz, Ushko, Nirogi, Atrak, and Tirich—are in the valleys of this section.

The Central Hindu Kush. The central region lies almost entirely within Afghanistan. According to the report of a British expedition in 1967, this region "has no nice, easily definable east–west ridge but rather a tortuous twisting watershed with massive off-shoots running north towards the Soviet Union and south into Nuristan (a region of Afghanistan)." Maximum heights, which are lower than those in the eastern section, include Koh-i-Bandakor (22,451 feet, or 6,843 metres), Koh-i-Mondi (20,498 feet, or 6,248 metres), and Mīr Samir (19,878 feet, or 6,059 metres). These peaks are surrounded by a host of lesser mountains. Glaciers are poorly developed, but the mountain passes—which include Putsigram (13,450 feet, or 4,100 metres), Weran (15,400 feet, or 4,700 metres), Ramgul (15,400 feet, or 4,700 metres), and Anjuman (13,850 feet, or 4,225 metres)—are high, thus making transmontane communications difficult.

Watershed of the Central Hindu Kush

The Western Hindu Kush. The mountains of the western region fan out gradually toward the Afghan town of Herāt, near the Iranian border, declining into hills of lesser importance. Communication is easier in this region, as the passes, such as the Shebar Pass (9,800 feet, or 2,987 metres), have long since been crossed by roads.

A wider definition of the limits of the Hindu Kush would lead to the inclusion of a fourth region known as Hindu Rāj. This is formed by a long, winding chain of mountains—with some lofty peaks, such as Darkot (22,447 feet, or 6,842 metres), and Būni Zom (21,499 feet, or 6,553 metres)—which strikes southward from the Lupsuk Peak (18,853 feet, or 5,746 metres) in the eastern region, then continues to the Lowarai Pass (12,100 feet, or 3,700 metres) and beyond to the Kābul River. If this chain were to be considered as part of the Hindu Kush, then the outlying mountains of Swāt Kohistān to the south would also form part of the complex. For most purposes of this article, however, the Hindu Rāj and its associated ranges are excluded from consideration.

International boundaries. International boundaries running through the Hindu Kush are primarily those of Pakistan and Afghanistan. The Karambar Pass lies about 40 miles west of the Chinese borders, while to the west the Hindu Kush, strictly considered, approaches the border between Afghanistan and Iran without extending into Iranian territory. Between these extremes the Pakistan–Afghanistan border follows the main watershed of the Hindu Kush throughout its eastern region, from Lupsuk Peak just north of the Karambar Pass to the Dorah Pass just south of Tirich Mīr. Not far from the Dorah Pass the boundary leaves the main watershed and follows minor spurs until it crosses the Kābul River, continuing along the crest of the Safed Koh Range toward the south. The Khyber Pass constitutes an important strategic gateway because it cuts through the Safed Koh range instead of through the Hindu Kush thus offering a comparatively easy route between the valley of the Kābul and the plains of Punjab.

The erratic boundary line is the result of a series of compromises reached at the end of the 19th century between the British and the ruler of Afghanistan; called the Durand Line, after the British negotiator, it has been inherited by the modern states of Pakistan and Afghanistan. Another curious configuration established about the same time and as yet unchanged is the Vākhān corridor, a panhandle of Afghan territory designed to act as a buffer between British India and tsarist Russia.

Geology. In many of its features the Hindu Kush resembles its eastern neighbour, the Karakoram Range that extends from Tibet into Pakistan. The Hindu Kush, which some authorities consider to be a continuation of the Karakoram, has the same core of igneous metamorphic rock (*i.e.*, rock formed by heat and pressure that has solidified from a molten state) flanked, especially toward the north, by sedimentary material. In the Hindu Kush, however, there appears to be a greater prevalence of sedimentary rocks—a fact that may explain the softer and more rounded forms of many of the mountains. In the Afghan section the core of the chain is formed by a complex sequence of metamorphic rocks containing marble, together with intrusions of granodiorites (rocks formed deep down by heat and pressure, and containing a certain admixture of both dark-coloured and light-coloured minerals). In Nūrestān the hills consist mainly of schists (medium- or coarse-grained metamorphic rocks), gneiss (a coarse-grained rock in which bands containing granular materials alternate with bands of schistose materials), and intrusions of granite with zones of migmatite (rock consisting of alternate layers of granite and schist). The Hindu Kush differs markedly from the Karakoram, however, in the winding direction of its strike (*i.e.*, its course, or bearing).

Drainage. The Eastern Hindu Kush appears to be formed of two parallel chains, consisting of a lower one to the north, which acts as a watershed, and a higher southern one that carries the main peaks. Drainage is comparatively simple on the northern side but highly complex on the southern one, where valleys follow two contrasting directions—northeast to southwest and roughly east to west. Most of the rivers, such as the Panjshir, the Alingar, the Kunar, and the Pānjkora, follow the northeast to southwest direction and are then suddenly deflected toward the east–west axis by the Kābul River, into which they flow. The Karumbar and Gizar valleys also take the same east to west direction. The Indus River, however, swerves in its course from one direction to another as it makes its roundabout descent toward the lower plains.

Climate. Since the range separates one important zone of Asia from another, the climate shows great variations. The mountains of Swāt Kohistān are virtually within the area of the rain-bearing monsoon winds, and most of the Eastern Hindu Kush, as well as the Hindu Rāj, rises up at the extreme limit of monsoonal Asia. The Central and Western Hindu Kush, however, border the Mediterranean climatic zone. Thus, moving from the southeast to the northwest and west, one moves from a region of rainy or snowy summer (from July to September) and of dry winters into a region of hot dry summer and cold and rainy or snowy winter (from December to early March).

Climatic contrasts

Climatic variations between these opposites also occur, producing often striking local contrasts.

A graphic image of climatic conditions is presented by the glaciers. The mantle of snow and ice is heaviest at the extreme eastern end of the Hindu Kush, where the Chiāntar Glacier is situated, and is also heavy in the higher section around Tirich Mīr and Saraghrar and in parts of the Hindu Rāj. Toward the west, however, glaciation is more sporadic. In the Central Hindu Kush, mountains 12,000 feet high are often bare almost to the summit. Most of the glaciers of the Hindu Kush appear to be retreating. A striking feature of some glacial regions are the so-called nieves penitentes, which are protruding spikes of frozen snow forming the illusion of kneeling human figures, sometimes two or three feet high, which are especially noticeable in the early morning; they are caused by the alternation of fierce sun and rapid evaporation during the day and of severe cold at night.

Vegetation. Differences in latitude and the variety of climates make it difficult to generalize about vegetation. Compared to the Himalayas of Nepal, and still more to those of Sikkim and Bhutan, the mountains of the Hindu Kush appear bare, stony, and poor in vegetation, although there are local exceptions. Some rich forests and pastures are found in the extreme southeast of the extended Hindu Kush region, as well as in the hills of Swāt Kohistān and in the Pānjkora Valley of the Dīr District. Parts of the valleys of the Gupis and Yasin rivers in the Gilgit area enjoy sufficient summer precipitation to be partially covered with vegetation. In the valleys of the Swāt and Dīr districts, as well as in some parts of the Chitrāl area, rice is cultivated.

The highlands of the eastern extremity of the Hindu Kush with their rolling pastures, bear some resemblance to Tibet, but further south valleys become arid and stony. A typical view in parts of Chitrāl, for example, would include snowy peaks in the distance, dry, barren, brick-red or ochre-coloured mountains all around, and bright emerald-green islands of vegetation near the villages where springs and irrigation furnish abundant water. In such oases, poplar trees are a distinct feature, often being accompanied by old and gigantic plane trees.

Much moisture carried by the monsoon winds penetrates the lower part of Chitrāl across the Lowarai Pass, so that stands of coniferous trees occur in the surrounding districts. The valleys of Nūrestān receive sufficient precipitation to have some pastureland and forests.

To the west of Afghanistan, and to the north of the Hindu Kush, extremely dry summers prevail; much of the country is stony, or sparsely covered with thorny and spiny plants, or with poor grass.

The meagre vegetation of most parts of the Hindu Kush does not favour wild animal life. The snow leopard manages barely to survive in the most remote valleys. Bears formerly roamed Nuristan, but few remain today. The markhor (a kind of wild goat) was once abundant, but hunters have now thinned the stock. Birdlife is rich, however, and eagles are occasionally to be seen.

The inhabitants. A long and tormented history, together with fragmented topography, has produced a veritable mosaic of ethnic units in the region. Kirgiz nomads, about 30,000 in number, graze their herds on the Vākhān uplands. The lower parts of the Vākhān corridor and the higher parts of the Sanglich and Anjuman valleys, all on the northwest slopes of the Hindu Kush, are sparsely inhabited by the so-called Pamir Tadzhiks, most of whom are Shīʿah Muslims. Other Tadzhiks (who are Sunnī Muslims), Uzbeks, and some Ḥazāra (Persian-speaking Mongols) live in the valleys of the central and western parts of the Hindu Kush. Afghans are found in the major towns, in Kābul, and in many districts to the south of the Hindu Kush, with the exception of Nūrestān.

On the southeast (Pakistan) side of the Hindu Kush, most people are Chitrāli, a racially mixed ethnic group that shows a marked cultural unity.

The Kafirs of Nūrestān

The Kafirs of Nūrestān, numbering 60,000 or 70,000, and of Chitrāl, numbering 2,000 or 3,000, are an exceptionally interesting people. Their name means "infidel" or "non-Muslim" and seems to have been used since the 11th century. Traditionally, they are divided into two groups—the *kalash* ("black") Kafirs of Chitrāl, and the *kati* ("red") Kafirs of Nūrestān. In the remote past, the Kafirs possibly inhabited a much larger area. The Kafirs of Nūrestān were forcibly converted to Islām in 1896.

Physically, the Kafirs do not seem to differ much from their neighbours; they speak a language classed by some as Dardic. It is in their religion and culture that their ethnic individuality is most strikingly expressed. In religion they practise a form of polytheism; worship consists mainly in the sacrifice of animals. Dancing is important, and divination through shamans is practiced. The dead are disposed of, unburied, in heavy wooden coffins. Large wooden statues of ancestors, often on horseback, stand near graveyards; many are works of vigorous and elemental beauty. Kafir homes are strong rectangular wooden buildings. Their economy is based on agriculture, hunting, and the raising of goats and oxen.

Exploration. The West took note of the Hindu Kush when Alexander and his armies crossed the mountains, which according to some authorities, he did twice. The Hindu Kush was well known to Arab geographers, as well as to the Chinese, who occupied Chitrāl in the 8th century AD. Marco Polo (*q.v.*), the Venetian traveller, and his group passed along the Hindu Kush through the Vākhān corridor. The mountains were also traversed by Timur (Tamerlane), the Mongol conqueror, in the 14th century and by Bābur, the Turkish conqueror who founded the Mughal empire, in the 15th century, in their expeditions against India. A number of explorers visited the region in the 19th century, and much knowledge was gained by the British during the two wars that they waged against Afghanistan from 1838 to 1842 and from 1878 to 1879. Further detailed knowledge of the area has been gained in the 20th century. Between 1960 and 1970 over 150 mountaineering and exploring expeditions visited the Hindu Kush.

Communications. The Hindu Kush offers a formidable barrier to communication. There are, however, some important passes. The Baroghil Pass (12,480 feet, or 3,804 metres high), at the head of the Chitrāl Valley, is one of the lowest and easiest openings in the 1,500 miles of forbidding mountains that border the north of India and Pakistan. Further west the highest section of the Hindu Kush offers, for about 100 miles, a wall quite unpassable by normal means of communication. In the Central Hindu Kush the passes are also high; only in the western section do more accessible passes occur. In 1964 a tunnel was completed under the Salang Pass (12,008 feet, or 3,660 metres) north of Kābul; consequently, for the first time in history, the great mountain wall has been pierced, making the north of Afghanistan accessible to the south at all seasons.

The Kakal Tunnel

The Hindu Kush can now be approached by motor transport from many directions. Chitrāl in the south is accessible, via the Lowarai Pass from Peshāwar, while the Kakal Tunnel has made Khānābād in the north accessible from Kābul. Lesser roads lead on to Faizābād and, from there, to the Vākhān corridor, or to Zebak, both situated in the heart of the mountain ranges. Many major peaks are now barely three or four days march from the last village accessible to motor transport.

Economic resources. The economic resources of the region remain virtually undeveloped. It is thought that the northern slopes may contain coal and perhaps petroleum or natural gas; the presence of gold, silver, copper, zinc, and lead also is suspected. Quantities of salt are known to exist, and lapis lazuli (a deep blue mineral, used as a gem), rubies, and beryl (a mineral, usually green, one variety of which is emerald) have been reported. There is some antimony (a white metallic element used in alloys of medicine) in Chitrāl. There is a hydroelectric potential, especially in the Daryā-ye Qondūz and Kokcha river basins on the northern side. Forests are still partially standing, but much merciless lumbering is rapidly reducing this resource. While many local peoples, especially the Chitrāli, make ingenious use of springs and rivers by building small aqueducts, agri-

culture would benefit greatly from more modern methods of irrigation. The local economy is at the bare subsistence level. Some dried fruit, a little lumber, mineral salt, charcoal, fodder grass, mats, and ropes are exported, mainly from Chitrāl. Sheep and goats are also raised.

Future prospects. Prospects for future development, as in most remote mountain districts, are not bright; it is even possible that improved communications, combined with the attractions of cities and industries elsewhere, may cause depopulation. While tourism can become an important source of income, tourist facilities—roads, trails, hotels, huts, and camping sites—have yet to be planned and built. Apart from this, however, there appears no reason why the Hindu Kush Mountains, with their wild valleys and rugged peaks, waterfalls and forests, and in some regions the prevailing blue skies and bracing air, should not become one of the playgrounds of the world.

BIBLIOGRAPHY. General geographic descriptions of the region appear in D.N. WILBER (ed.), *Afghanistan* (1956); and J.C. GRIFFITHS, *Afghanistan* (1967). M. KLIMBURG, *Afghanistan* (1966), is an authoritative survey (in German) rich in detailed information. S.G. BURRARD and H.H. HAYDEN, *A Sketch of the Geography and Geology of the Himalaya Mountains and Tibet*, 2nd ed. (1933); and A. GANSSER, *Geology of the Himalayas* (1964), deal briefly and technically with the eastern section of the chain. The results of alpine exploration and climbing up to 1965 are recorded by B. CHWASCINSKI, "The Exploration of the Hindu Kush," *Alpine Journal*, 2:199–214 (1966).

Invaluable material on Kafir culture, previous to its dissolution, is contained in *The Kafirs of the Hindu-Kush* by G.S. ROBERTSON (1896). R.C.F. SCHOMBERG, *Between the Oxus and the Indus* (1935) and *Kafirs and Glaciers* (1938), provide much information on the Hindu Kush and its people. *The Pathans, 550 B.C.–A.D. 1957* by O.K. CAROE (1958); and *Anthropological Research in Chitral* by P. GRAZIOSI (1964), offer further important contributions.

Aspects of Hindu Kush geology are treated in S. MATSUSHITA, *Geology of the Karakoram and Hindu Kush* (1965); vegetation in S. KITAMURA, *Flora of Afghanistan* (1960); and insect life in M. UENO (ed.), *Insect Fauna of Afghanistan and Hindukush* (1963). For early history of the region, see W.W. TARN, *The Greeks in Bactria and India* (1938) and *Alexander the Great* (1948). For later periods, consult *Afghanistan: Highway of Conquest* by A. FLETCHER (1965); and *Afghanistan: A Study of Political Developments in Central and Southern Asia* by W.K. FRASER-TYTLER, 3rd ed. (1967).

Two eminently readable travel accounts are: *Between Oxus and Jumna* by A.L. TOYNBEE (1961); and *A Short Walk in the Hindu Kush* by E. NEWBY (1958).

(F.M.)

Hindu Mysticism

Mysticism is usually defined as union, or desire for union, of the self with something greater than the self, whether that be defined as a principle that pervades the universe or as a personal God. Hindu mysticism includes both these definitions and a great many that lie in between. At one extreme is the realization of the identity of the individual self with the impersonal principle called Brahman, the position of the Vedānta school of Indian philosophy; and at the other is the intensive devotionalism to a personal God, called by a variety of names, which is found in the Bhakti (devotional) sects.

TYPES OF MYSTICISM

It is usual for writers on the subject, following Surendranath Dasgupta, a historian of Indian philosophy, to consider that there are five major varieties of Hindu mysticism, the five having arisen in historical order as follows:

1. The sacrificial, based on the texts called Vedas, the earliest Hindu scriptures, and *Brāhmaṇas*, ritual texts (*c.* 1500–1000 BC); in this variety the sacrifice, a complex ritual involving recitation of sacred formulas and the offering of sacred materials to the fire, recreates the original creation; the forces of nature and the gods are controlled by the sacrificer through his knowledge of the texts and rituals.

2. The *Upaniṣadic*, exemplified by early religio-philosophic texts called *Upaniṣads* (1000–500 BC), which are filled with dense speculations on the relationship of man to the universe, and in which are found the beginnings of both monistic (concerned with a unitary principle of reality, immanent in the world) and theistic (concerned with a personal or suprapersonal God) systems.

3. The Yogic, relating to physical and mental discipline; the earliest known text of this school is the *Yoga-sūtra* of Patañjali, dated variously between the 2nd century BC and the 4th century AD. According to Yogic mysticism, man realizes union by means of physical and mental control of himself, which leads to control of both natural and divine forces.

4. The Buddhistic, in which enlightenment is the realization of the four Truths—the fact of pain, the cause of pain, the cessation of pain, and the means of arriving at these three truths: the Eightfold Path. The ultimate state, the culmination of one path of the Eightfold Path, is Nirvāṇa, "the blowing out," the extinction of desire (see BUDDHIST MYSTICISM).

5. The devotional or Bhakti type of mysticism comprises a range of theistic systems, with a conception of absolute dualism between man and God on the one extreme, and a conception of qualified nondualism on the other. Although there are traces of this devotionalism throughout the history of Indian religion, it began to develop in earnest in South India in the 9th and 10th centuries AD with the hymns of the poet saints called Āḻvārs. Bhakti is the primary characteristic of medieval and modern Hinduism.

Such a historical division as the one just given obscures the fact that several of these five "types" have much in common; for example, the technique of meditation is similar in both Buddhism and Yoga. It may be that a different type of categorical analysis will be more fruitful.

Four elements common to all Hindu mystical thought

There are four things common to all Hindu mystical thought. The first is that it is all based on experience: the state of realization, whatever it is called, is both knowable and communicable, and the systems are all designed to teach men how to reach it. It is not, in other words, pure speculation. Second, it has as its goal the release of the spirit-substance of man from its prison in matter, whether matter be considered real or illusory. Matter is the cause of the pain of which Buddhism speaks. Third, all the systems recognize the importance or the necessity of the control of the mind and body as a means of realization; sometimes this takes the form of extreme asceticism and mortification, and sometimes, at the other extreme, it takes the form of the cultivation of mind and body in order that their energies may be properly channelled. And, finally, at the core of Hindu mystical thought is the functional principle that knowing is being. Thus, knowledge is something more than analytical categorizing; it is total understanding. This understanding can be purely intellectual, and some schools can equate the final goal with omniscience, as does Yoga. Knowing can also mean total transformation: if one truly knows something, he is that thing. Thus, in the devotional schools, the goal of the devotee is to transform himself into a being who, in eternity, is in immediate and loving relationship to the deity. But despite the fact that these are both ways of knowing, the difference between them is significant. In the first instance, the individual has the responsibility to train and use his own intellect. The love relationship of the second, on the other hand, is one of dependence, and the deity assists the devotee through grace. The distinction is generally made by the analogy of the cat and the monkey: the cat carries her young in her mouth, the kitten has no responsibility. But the young monkey must cling by its own strength to its mother's back.

A BRIEF HISTORY

Early mysticism. The earliest books of the Hindus, the Vedas (*c.* 1500 BC), are not deeply concerned with speculative thought. They are, rather, hymns to a series of gods who have control over natural forces and therefore over the fortunes of men. These gods are categorically separate from men, however, and, although they enter into human life, there is no union between men and gods. The attitude of men, hardly loving, was that propitiation

Mysticism of the Vedas and *Brāhmaṇas*

was necessary. Nature, to a wandering people, was often ruthless, sometimes dangerous, and only by sacrifice could natural forces be controlled. As time went on, the sacrifices became more and more elaborate, so elaborate that it took days to complete them. And with increasing complexity came increasing specialization in performance. Thus, around 1000 BC, a series of texts developed, called *Brāhmaṇas*, which were handbooks to these rituals. From these books it is clear that ritual had become an end unto itself. Its propitiatory function had been lost; the rituals had life and meaning all their own. They themselves were the re-enactment of the creation; except in form, they were no longer offerings to deities. They were themselves the deities. The ritual was identified with the cosmos, and by control of the ritual the cosmos was controlled. The gods that had populated the universe during the Vedic period had all but disappeared. The sacrifice was identified with the world process itself: the creation had been through the sacrifice of primordial man. The Vedic god of fire, Agni, had now become the messenger from man to the powers that control the universe. The sacrificer spoke the esoteric formulas that themselves contain awful power as he poured the sacrificial substance into the mouth of the fire; because he knew the words that contain power, the sacrificer was that power. The formulas, called *mantra*s, are important in much of later mystical thought and practice. They are used, often in the form of "seed *mantra*s" (*bīja-mantra*), both as aids to meditation and as direct means of realization. For, it is said, "the seed is in the tree, and the tree in the seed"; the universe is in the *bīja-mantra*, and knowledge of that *mantra* means that the knower is the universe.

This was the legacy that was carried into the texts called *Upaniṣads*, usually dated between 1000 and 500 BC, though there are very recent books that call themselves by this name. The early *Upaniṣads* are usually cited as the beginnings of Hindu speculative thought, and indeed they are filled with tortuous musings. But in no way do they present a unified system, or even a related series of systems; there are variations, inconsistencies, and contradictions, not only between one text and another, but often internally within a single text. Some of the *Upaniṣads*, such as the *Īśā* and *Kena*, are theistic and are used as authority for later theistic thought. Others, such as the *Bṛhadāraṇyaka* and *Chāndogya*, are largely pantheistic, though even these early texts are woven of various strands. Despite their inconsistencies, the *Upaniṣads* were all considered authoritative, for they are developments of the absolute authority, the Vedas themselves. This provided later thinkers with justification for a variety of positions.

There is, says the *Chāndogya Upaniṣad* (7.25), a spiritual principle called Brahman, which creates the universe, pervades the universe, and is the universe:

> That [*i.e.*, the spiritual principle], indeed, is below. It is above. It is to the west. It is to the east. It is to the south. It is to the north. It, indeed, is this whole world.

The self and the soul (called Ātman) also pervade the universe:

> Next, the instruction with regard to the Ego. I, indeed, am below. I am above, [etc.]. Next, the instruction with regard to the soul. The Soul, indeed, is below: The Soul is above, [etc.].

Mysticism of Śaṅkara

Such pantheistic statements can be interpreted in many ways. The great teacher Śaṅkara (*c.* AD 788–820) chose to have them mean that the soul, the *Ātman*, is identical with the all-pervasive spiritual principle, the Brahman, and that release is the experience of that identity. In his commentaries on the *Upaniṣads* and on the *Brahma-sūtra* of Bādarāyaṇa, Śaṅkara established what came to be called the "End of the Veda," the Vedānta. The *Upaniṣads*, he held, demonstrated that the pure self as pure being, pure intelligence, and pure bliss is itself the ultimate truth, and that the world and all else that is contingent is imperfect and false (*māyā*). *Mokṣa*, or release, to the Vedānta

> meant the dissociation of the self from the subjective psychosis and the world. . . . [The] Vedānta . . . held that the world as such has no real existence at all, but is only an illusory imagination which lasts till the moment when true knowledge is acquired. (From S.N. Dasgupta, A History of Indian Philosophy.)

To attain this knowledge, to separate the contingent from the eternal, to realize oneness with Brahman, no action or worship is necessary. Knowledge is what pierces the veil of ignorance that hides the truth. It is only necessary that a man lead a disciplined life, that he be no longer concerned with the things of this world, that he be able to discern what is transitory and what is permanent, that he be desirous and capable of peace, and that he exercise restraint and faith and be capable of deep concentration (the latter being a stipulated technique in many systems). Śaṅkara and the Vedānta do not deny the concept of a personal God (Īśvara). But since the definition of Brahman must be in terms of negatives ("it is not this, nor is it that"), there are those who are not satisfied by this. Brahman is by them called Īśvara. But when the ultimate state of identity is realized, the necessity for the notion of Īśvara, like other imperfect things, vanishes.

There are other possibilities, said the schools of Yoga. These also do not deny God, but union with God is not the goal of Yoga. God is a particular soul, not essentially different from other particular souls, which are, like God, eternal. The bondage of the soul, say the schools of Yoga, is due to lack of discrimination. Mental activity, which is involved with the things of the world, must be controlled. The means to this control is an eightfold path, of which the three most important elements are fixed attention, contemplation, and concentration (*samādhi*). In *samādhi*, the perfect life of the spirit, the soul is omniscient, self-illumined, and eternal. *Mokṣa* is the soul's realization of itself as an eternal monad.

Neither of these systems, then, has any real place for a personal God. The process of release is initiated by the self, pursued by the self, and attained by the self. And yet, these systems were confronted with such *Upaniṣadic* statements as this one, from *Śvetāśvatara Upaniṣad* 4:11:

> The One who rules over every single source,
> In whom this whole world comes together and dissolves,
> The Lord, the blessing-giver, God adorable—
> By revering him one goes forever to this peace.

Mokṣa, then, to the theistic *Upaniṣads*, is not in becoming one with the One but is in worship and reverence, which results in knowledge of the peace that is in God but that is not the totality of God. By *bhakti*, by devotion and faith, one gains that peace and joy; God is a God of love and grace who, like the god *Kṛṣṇa* (Krishna), hero of the *Bhagavadgītā* ("The Lord's Song"), can be known, at least in part, as a person. Such is the basis for the dualistic systems.

Mysticism of Madhva and Rāmānuja

The philosopher Madhva, born in the 12th century in South India, took a position of absolute dualism, in opposition to the Vedānta. To Madhva and his followers all texts and schools that speak of the unity of the self with the Brahman are false. The world is real, and the essential principle is that of the fivefold difference: difference between the self and God, between selves, between matter and God, between self and matter, and between matter and matter. Reality is of two kinds, independent and dependent; only God is independent reality, matter and souls being dependent on and controlled by God. The self, however, though dependent, is active and is responsible for its own release, that release coming through *bhakti* or devotion. Since it is the individual's responsibility, the ultimate state is gained by comparatively few (only in Madhva, of all Hindu religious and philosophical systems, is there a notion of eternal salvation). That ultimate state, as in other systems, is one in which there is neither pain nor sorrow. But unlike many other systems, in Madhva, since every soul is unique, the ultimate condition varies from individual to individual.

Rāmānuja, who also was born in South India about AD 1050, stopped well short of this absolute dualism and in fact represents a middle ground between the two extremes. To Rāmānuja, God is not the unqualified absolute principle of Śaṅkara but is a personal God who can be known, though imperfectly. God is internally differ-

entiated, and souls and matter are aspects of that differentiation. Though they are contained within God and are attributes of him, souls and matter possess qualities of their own. They are, therefore, both the same as and different from God. Identity means to Rāmānuja only that souls and matter do not exist independently; hence, identity is not the ultimate state. Matter, since it is one of the attributes of God, cannot be unreal. Furthermore, the sacred texts speak of the creation of the world by God, and if the world is a creation of God, it must be real; because matter changes and does not persist in a single form, it cannot be concluded to be unreal. So while for Śaṅkara liberation is the removal of the barrier of *māyā* (illusion), for Rāmānuja it is communion with God through devotion. When the goal is reached, the individual is not obliterated, but, as in Madhva, he retains his identity and enjoys forever the blissful fruits of his faith. He does not become one with God. This is the first of two new ideas that were to have great effect on later developments. The second is that the worshipper proceeds along the path of devotion because of the grace of God, which instills in him a desire to approach the godhead and share in the bliss of nearness. These ideas are vital to the later syncretistic systems, of which that of the Vaiṣṇavas of Bengal might be taken as an example.

Later mystical thought and method. Although the Vedic gods had largely disappeared from the monistic philosophies, they persisted, in changed form, in popular Hinduism, and eventually a select few found their way into the theism of Rāmānuja and Madhva. The most popular were Śiva, "the Auspicious One," and Viṣṇu, or Nārāyaṇa, and his incarnations Rāma and especially Kṛṣṇa. In the later schools, the personal nature of the deity was accepted: these gods had names and qualities. The questions became those of the proper relationship of man to God and the nature of a godhead that allows the simultaneous immanence and transcendence that Rāmānuja proposed.

The answer to the first of these questions seems to have been found—as it was by St. John of the Cross, the 16th-century Christian mystic, and others in very different traditions—in the symbol of human love. The Lord is the lover, the worshipper is the woman whom he loves. The human soul is feminine and yearns for and sometimes attains the Lord. This is clearly poetic expression rather than systematic doctrine. But it is a poetic expression that found voice through many of the languages of India. To the woman poet Mahādevīyakka, who wrote in Kannada in the 12th century, the Lord Śiva (Shiva) is the object of a very real and intensely passionate love. And yet he is beyond passion, attributes, death, and form:

Love mysticism of Mahādevīyakka

I've fallen in love with the splendid one:
death decay form
place and part
has He none,
and no birthmarks.
O mother, I love him.
. . .
I fell for the Handsome One
who is not of this world,
has no fear:
therefore, Arjuna of the Jasmines
is my husband.
Take the rest of them these husbands—
who die, who rot—
and throw them
into kitchen fires.

(Translation by A.K. Ramanujan; used by permission.)

Like St. Catherine of Siena, the 14th-century Christian mystic, who felt herself married to Christ, Mahādevīyakka felt herself married to Śiva, who was to her real and immediate and yet beyond all material qualities. Male devotees also, like the earlier Āḻvars, who wrote in Tamil, felt themselves to be female, and the Lord Kṛṣṇa (Krishna) the sole male of the universe. Nammāḻvār, of the 8th century, wrote such lines as these:

All places, shining like great lotus pools
On a blue mountain broad, to me are but
The beauties of his eye—the lord of earth
Girt by the roaring sea, heaven's lord, the lord
of other good souls, black-hued lord—and mine!

Kṛṣṇa, black in colour, pervades all. The Vaiṣṇavas of Bengal, too, in the 15th and 16th centuries, used the same image to express their passionate love for Kṛṣṇa:

Let the earth of my body be mixed with the earth
my beloved walks on.
Let the fire of my body be the brightness
in the mirror that reflects his face.
Let the water of my body join the waters
of the lotus pool he bathes in.
Let the breath of my body be air
lapping his tired limbs.
Let me be sky, and moving through me
that cloud-dark Shyāma, my beloved.

(Trans. by Edward C. Dimock, Jr., and Denise Levertov, *In Praise of Krishna;* Copyright © 1967 by Asia Society, Inc. Reprinted by permission of Doubleday & Company, Inc., and Jonathan Cape Ltd.)

In such poetry the images are immediate and intimate, but they are not images of union. The longed-for condition of the worshipper is near, not identical with, the divine lover. For love is the ultimate state, and if there were identity, the dynamic process of love could not exist. Yet the intimacy is extreme; Kṛṣṇa's loves are in some way extensions of himself. The relationship, say the Vaiṣṇavas, is both the same and not the same: it is as the flame to the fire, or the flower to its scent.

THE VAISNAVAS OF BENGAL

Unlike many other schools, the Vaiṣṇavas tried to explain this relationship systematically. The main text of the doctrine called *acintya-bhedābheda*, "unknowable and simultaneous difference and nondifference" (between the soul and God) is the *Ṣaṭ-saṃdarbha* of the 17th-century philosopher Jīva Gosvāmin. According to this text, the relationships are defined as follows:

Kṛṣṇa mysticism

The deity, Kṛṣṇa, is of three levels of reality. These are described by an analogy also used by Rāmānuja, that of the sun. The sun itself is of course the highest aspect; Brahman, the One of the Vedānta, is merely the undifferentiated radiance of this orb. But knowledge of the sun comes by two of its characteristics: sunbeams, which transmit the qualities of the sun to the earth, and the reflection of the sun in a mirror or a body of water. The three are interdependent: without the solar disk, neither sunbeams nor reflection exist. Without a reflection, the sun does not exist, for what is real is reflected. And the sunbeams are intermediate between the two. But the point of the image is that the reflection of the sun is a function of the sun itself; it represents the reality of Kṛṣṇa in the world. Not only is the world therefore real, but through it Kṛṣṇa can be known.

The text called *Bhāgavata-Purāṇa*, from South India in the 9th or 10th century, is considered to be revelation and truth. It is enigmatic and poetic, telling of the various aspects of the god Kṛṣṇa—as king, mighty God, child, and, most important to the Vaiṣṇavas of Bengal, as lover of the cowherding village girls called *gopīs*. This *Purāṇa* is taken to be the commentary by the legendary sage Vyāsa on the Vedas themselves. It is accepted as the ultimate authority.

The love between Kṛṣṇa and the *gopīs* has all the varied aspects of a fully human love, with anger and satisfaction, separation and union. Yet it is, of course, more than human love, for Kṛṣṇa is the true form of the godhead. Although he walked the earth, as the story is told in the *Purāṇa*, he was untouched by the material qualities of it; he could not in fact be touched by them, for his true form, while a form as if made up of those qualities, is in fact immaterial. The worshipper, therefore, in order to know the pure love that existed between Kṛṣṇa and the *gopīs*, must transform himself; he must also gain a divine, immaterial form; he must in fact become a *gopī*. As with Madhva, there is a difference between matter and spirit, a difference between man and God, and a difference between *gopī* and Kṛṣṇa. There can be union of the *gopīs* with Kṛṣṇa, as in the analogy of human sexual love. But there can never be unity.



Krsna (Krishna) dancing with the gopīs, painting from western Rājasthān, *c.* 1610. In the N.C. Mehta Collection of the Gujarāt Museum Society, Ahmadābād, India.
By courtesy of the Gujarat Museum Society, Ahmadabad, India

But there are others in the *Bhāgavata* story besides the *gopīs*. There are Kṛṣṇa's foster-parents, his friends, his servants, and others. All of these love him, in different ways. Each man, say the Vaiṣṇavas, is distinct and individual; therefore paths of devotion will differ, and, as the *Gītā* says, what is gained depends upon the path followed. The common quality of the ultimate experience is that all—*gopīs*, parents, and the others—are near Kṛṣṇa and love him. And not only are there many people in the story whose paths might be followed, but the story itself is a representation, in time, of what is happening eternally in heaven. Thus each moment of each emotion of each of the characters is also eternal. When the worshipper is transformed, he knows love for Kṛṣṇa through one who loves him in the story, and he experiences forever what on earth would be a momentary passion.

Mystical techniques

The mode of transformation involves characteristics that are common to all later schools (and many earlier schools as well) of mysticism. Among these are discipline of the body (though not in the way of Haṭha, or physical, Yoga), the help of the guru (religious teacher), and the use of the name of the deity.

There are said to be 64 acts of devotion, and these are undertaken for several purposes. They concentrate the whole attention on the religious object, the attainment of transformation; they discipline the body so that religious activity becomes normal activity; and they demonstrate purpose, which disposes the deity to grant his grace. The acts themselves include singing the names and praises of Kṛṣṇa, going to the temple, listening to the reading of the *Bhāgavata-Purāṇa*, putting the Vaiṣṇava signs upon the body, associating with holy men, and concentrating on or "remembering" the episodes of the *Bhāgavata-Purāṇa*, especially those appropriate to one's own goals. With constant performance of such acts, if the effort is sufficiently sincere, grace is granted, and the worshipper passes beyond the state in which external acts are necessary. Sometimes while still in this life one attains the state of true and eternal love.

Like the sun, the true form of the deity is knowable on earth only indirectly, until transformation is reached. A guru is one who has already reached the ultimate state but who, like the *bodhisattva* ("Buddha-To-Be") of Mahāyāna Buddhism, remains in the body in order to help others reach the ultimate state. The guru is a channel of the grace of the deity; he is the sunbeam. Unlike the deity himself, the guru is related to the earth, for he retains his material human form, and he has memories of the pain of life on the earth. He knows man and therefore can be sympathetic; he also knows Kṛṣṇa and can demonstrate his love.

There are two types of gurus—the guru who gives the *mantra*, or formula, that the devotee will meditate on and repeat to assist himself on the path of *bhakti;* and the guru who instructs and guides the devotee step by step along the way. The *mantra*—often a name of the deity—is crucial in the process of transformation, and it is said that taking the name of Kṛṣṇa only once will cleanse the heart and dispose Kṛṣṇa to grant his grace. The *mantra* in many ways contains the same power as do the formulas of the sacrifice; the *mantra* is in fact said to be the sacrifice of this, the Kali, age, the last and most degenerate of the traditional Indian ages of the world. The *Viṣṇu-Purāṇa* points out that men, in this age of degradation, are incapable of performing the Vedic sacrifice properly; the gracious deity, therefore, has given his name as an easier means to salvation.

This concept of the power of the word is familiar to many systems of Indian thought. But to the Bhakti systems, and to the *Tantras*, esoteric and magical sacred texts, it is crucial.

THE TANTRAS

The Tantric tradition is a very ancient one but one that has become, through time, so interwoven with more orthodox Hinduism that it is difficult to define precisely. Like the Vedānta, it is monistic; but it differs from the Vedānta in that, while it sees an identity between the soul and the cosmos, it speaks of the internalization of the cosmos rather than of the release of the soul to its natural state of unity. The body is the microcosm, and the ultimate state is not only omniscience but total realization of all universal and eternal forces. The body is real, not because it is the function or creation of a real deity but because it contains the deity, together with the rest of the universe. The individual soul does not unite with the One; it is the One, and the body is its function.

Tantrism, though not always in its full esoteric form, is a feature of much modern mystical thought. In Tantrism, the consciousness is spoken of as moving—driven by repetition of the *mantra* and by other disciplines—from gross awareness of the material world to realization of the ultimate unity. The image is of a serpent, coiled and dormant, awakened and driven upward in the body through various stages of enlightenment until it reaches the brain, the highest awareness. The modern mystic Ramakrishna describes the process, which also describes the experience that all Hindu mystical process seeks:

Mysticism of Ramakrishna

> When [the serpent] is awakened, it passes gradually through [various stages], and comes to rest in the heart. Then the mind moves away from [the gross physical senses]; there is perception, and a great brilliance is seen. The worshiper, when he sees this brilliance, is struck with wonder. The [serpent] moves thus through six stages, and coming to [the highest one], is united with it. Then there is *samādhi* . . . When [the serpent] rises to the sixth stage, the form of God is seen. But a slight veil remains; it is as if one sees a light within a lantern, and thinks that the light itself can be touched, but the glass intervenes. . . . In *samādhi*, nothing external remains. One cannot even take care of his body any more; if milk is put into his mouth, he cannot swallow. If he remains for twenty-one days in this condition, he is dead. The ship puts out to sea, and returns no more.
>
> (Translation by Edward C. Dimock, Jr. Source is *Śrīśrīrāmakṛṣṇa-kathāmṛta;* Calcutta: Ramakrsna Mission.)

It is axiomatic that no religious idea in India ever dies or is superseded. All of these systems exist still, and are practiced still. As are so many areas of Indian life, Hindu mystical thought is of wondrous complexity and richness, uninhibited by dogma, a laboratory for the understanding of these special strivings of the religious mind.

BIBLIOGRAPHY. In most cases, Hindu mysticism is treated as an integral part of philosophy as a whole; the standard works on mysticism, then, are the standard histories of philosophy. The classic work among these is S.N. DASGUPTA, *A History of Indian Philosophy*, 5 vol. (1922–55). Excerpts of the relevant ideas from Dasgupta's larger study is *Hindu Mysticism* (1927, reprinted 1959). Other standard histories include S. RADHAKRISHNAN, *Indian Philosophy*, 2nd ed., 2 vol. (1956). More recent and more specific is R.C. ZAEHNER, *Hindu and Muslim Mysticism* (1960). In addition there are many sources on specific aspects, such as (for Yoga) MIRCEA ELIADE, *Le Yoga: immortalité et liberté* (1954; Eng. trans., *Yoga: Immortality and Freedom*, 2nd ed., 1969); or for the Bhedābheda school of the Vaiṣṇavas of Bengal, S.K. DE, *The Early History of the Vaisnava Faith and Movement in Bengal*, 2nd ed. (1961).

(E.C.D.)

Hindu Mythology

Hindu mythology consists of the stories of the several varieties of gods, the myths, and the tales of Hinduism, one of the oldest continuous religious traditions in the world and the predominant religion of India. This mythology has inspired a monumental literature, of theatre, song, dance, sculpture, painting, and architecture. In accepting the fearful as well as the benevolent, the sublime as well as the grotesque, it mirrors life, as viewed by the Hindus, to the fullest. Hindu mythology has a continuous documented history dating from *c.* 1400 BC. It is rich and varied, ranging from vague personifications of natural phenomena to an intricate mythical architecture of space and time. It can provide a grand cosmogonic design whereby the Hindu can understand the origin of the universe and man's place in it, or it can tell a simple story to explain the holiness of a river ford.

Myth in Hindu life and worship

Myth constantly enriches the Hindu's life, born as he is with the hereditary deities of his family and his village, to which he also adds a deity that he chooses for his own devotions. A wide variety of festivals, in which he shares with his fellow Hindus dramatizations of cosmic and divine events, punctuate his year. As a caste Hindu he communes with his caste fellows in a common caste myth. He will go on a pilgrimage and at the great centres of pilgrimage share in all-India myths. In the large temples he will visit and he will worship his gods, often represented in each place by a different icon, illustrating the variety of his gods' feasts. At festivals he will listen to the recitations of professionals, which glorify the myth of the day. Artistic performances of dance and music will continually remind him of the myths of which they are expressions and which also provide the themes of devotional songs. There is no aspect of Hindu life that is not entwined in mythology.

SOURCES AND VARIETIES OF MYTHS

Distinction must be made between the so-called Vedic mythology and classical Hindu mythology, though they are in part continuous. Classical mythology in turn can be contrasted with folk myth, though it is not easy to distinguish the differences. This difficulty is due to historical processes. The oldest sources for Hindu mythology are the Vedic texts composed in Vedic, an early form of Sanskrit. These texts are the literature of an Aryan people who were in fact the last to arrive in India. These Aryans, who spoke an Indo-European language akin to the languages of Europe and Persia, began their invasions prior to 1500 BC. They found in the Indian subcontinent peoples of different stock, and the ruins of a culture in the Indus Valley (in what is now Pakistan), known as the Harappan civilization. At first the Aryans were slow to assimilate the indigenous, non-Aryan inhabitants, who were styled *dasyus* (people to be tamed) and were profoundly hated. Vedic literature is thus mostly conservative, the expression of the ruling Aryan warrior and priest classes. It is not until 500 BC that there is found in the Hindu literature impressive evidence of a body of beliefs that incorporates, side by side with the Vedic one, myths that were non-Vedic as well as non-Aryan. But the degree of assimilation is so great that it is hard, if not impossible, to isolate satisfactorily the Aryan elements from the non-Aryan ones. Yet assimilation was by no means complete, and today there are still folk mythologies that have not been absorbed into official Hindu mythology.

The great works of Vedic literature

Vedic. The Vedic literature, which is the earliest source of Hindu mythology, ranges from the Ṛgveda (Rigveda) (*c.* 1400 BC) to the *Upaniṣads* (*c.* 500 BC). This literature provides the sole documentation for all Indian mythology before Buddhism and the early texts of classical Hinduism. Since it is the literature of a ruling class, it does not represent all the myths and cults of the early Indo-Aryans, let alone those of the non-Aryans.

The most important texts are the four collections (saṃhitā) known as the Veda or Vedas (*i.e.*, "Book[s] of knowledge"): the Ṛgveda (Wisdom of the Verses), the Yajurveda (Wisdom of the Sacrificial Formulas), the Sāmaveda (Wisdom of the Chants) and the Atharvaveda (Wisdom of the Atharvan Priests). Of these, the Ṛgveda, the oldest, is also the richest in myth. The text consists essentially of hymns in praise of certain gods as powers of nature and has become correlated with the Vedic liturgy. The gods celebrated are mostly the more important gods, and the text presents the official pantheon of the ritual. The Yajurveda mainly contains the formulas spoken by certain priests at the ritual. The Sāmaveda is an anthology of Ṛgvedic verses and thus adds nothing to the Ṛgveda. Different is the Atharvaveda, which is concerned with magic, and mentions spirits and demons of the lower cults.

In the Vedic texts following these earliest compilations, viz., the *Brāhmaṇas* (discussions of the ritual) *Āraṇyakas* (books studied in the forest) and *Upaniṣads* (secret teachings concerning cosmic equations), the interest in the early Ṛgvedic gods wanes and they become little more than accessories to the Vedic rite. Polytheism begins to make way for a sacrificial pantheism of a "god sacrifice," Prajāpati (lord of creatures), who is the "All." In the *Upaniṣads* Prajāpati merges with the concept of Brahman, the supreme reality and substance of the universe (not to be confused with the Hindu god Brahmā), replacing any specific personification, and mythology is transformed into rudimentary philosophy (see also HINDU SACRED LITERATURE).

Classical. Classical Hindu mythology is documented by the Sanskrit texts of the classical period (*c.* 500 BC–*c.* AD 1000), art and architecture, and by the vernacular, largely epigonic, literature after AD 1000. The earliest sources of the classical period are the texts of the two colossal epics, the *Mahābhārata* (Great Epic of the Bharata Dynasty) and *Rāmāyaṇa* (Romance of Rāma), and the no less voluminous encyclopaedias of Hinduism, the *Purāṇas* (sacred lore). At first sight the discontinuity between Vedic and classical mythology appears to be so sharp that one might wish to consider them as being of altogether different traditions. Yet it soon becomes clear that they are in part continuous, and that what appears to be discrepancy is but a difference between the narrow, liturgically minded emphasis of the Vedas, and the wide, catholic acceptance of the epics and *Purāṇas*. For example, the great god of the Ṛgveda is Indra, god of war and monsoon, prototype of the warrior; but for the population as a whole he was more important as the rain god than the war god, and it is as such that he survives in early classical mythology. Little is learned in the Veda of goddesses, yet they must have existed for they rose steadily in recognition in classical mythology.

The literature of classical Hinduism

While in these works some of the Vedic gods have an afterlife in which their importance is reduced, other gods, previously of less official significance, arise. The two principal gods of classical Hinduism are Viṣṇu (Vishnu) and Rudra-Śiva. Both are known in the Vedas, Viṣṇu as the strider who with his three strides establishes the three worlds of heaven, atmosphere, and earth, and thus is present in all three orders; Rudra-Śiva as a mysterious god who is to be propitiated.

The classical literature documents the stages of the rise of the two gods, who become as it were god-types, attracting to themselves the identities of other popular gods and heroes: Viṣṇu those gods who protect the world and its order, Śiva (Shiva) the powers that are outside and beyond them. To these two is often added Brahmā, personification of the Brahman of the *Upaniṣads*, creator of the world and teacher of the gods. Though still somewhat important in the *Mahābhārata*, Brahmā later dwindles away.

In the Purānic literature (*c.* AD 500–1000) sectarianism creeps into mythology and one god is extolled above the others. Of prime interest are cosmology, myths of the great ascetics (who in some respects eclipse the old gods), and myths of sacred places, usually rivers and fords, whose powers to reward the pilgrim are often cited and related to a local legend.

Folk and tribal. The sources of folk and tribal mythology are vernacular literature, oral tradition, folklore, and folk and tribal arts. Folk mythology derives from the most ancient times and has influenced both Vedic

and classical mythology. Classical mythology became what it is by continuously assimilating myths not previously known or accepted, so that the line between classical and folk (and tribal) mythology is apt to be arbitrary. The great (Sanskrit) tradition of classical mythology (as contrasted with the local or little traditions of folk mythology) may include within its own scheme a god who continues an independent existence on a folk level. Conversely, an incident of the great tradition may be adopted and adapted on the folk level. For example, the local Mahārāstrian god Viṭhobā is identified with a manifestation of Viṣṇu and thus assured a place in the great tradition; on the other hand, the widely celebrated festival of Navarātrī (nine nights) is associated in northern India with a village goddess, Naurthā.

The concepts of *avatāra* and *vāhana*

Certain concepts that evolved in classical mythology have facilitated the absorption of folk elements; two of these should be singled out: *avatāra* (incarnation), and *vāhana* (vehicle).

The concept of *avatāra* (lit. descent) issues from the belief that in times of trouble a god, notably Viṣṇu, descends and incarnates himself as a man or hero to set matters aright. Such a concept provides the opportunity of identifying a local deity (like Viṭhobā, above) with an all-Indian god like Viṣṇu. The concept may also extend to the worship of very local hierophanies (manifestations of the sacred; *e.g.*, South Indian Vaiṣṇavism accepts "icon-incarnation" [*arcāvatāra*], in which Viṣṇu "descends" in a local icon).

According to the concept of a *vāhana* (lit. mount), every god has an entourage of his own, which includes a favourite riding animal; this facilitates many folk associations. Viṣṇu's mount is the bird Garuḍa, an old solar symbol; Śiva's is the bull Nandi, whose worship may go back to the ancient Harappan civilization. There are other mythological patterns, *e.g.*, adoption in a family; thus the folk god Gaṇeśa, an elephant-headed god, is made the son of Śiva, as is Kumāra Kārttikeya, the war god, who arose from the South Indian war god Murugaṉ. Hanumān, the monkey god, becomes an all-Indian god as a helper of Rāma, who is an *avatāra* of Viṣṇu.

Other spirits and godlings of folk provenance are not absorbed to the same degree and so retain their folk character. Important are the snakes (*nāgaṣ*), objects of folk cults to which great power is attributed; the *yakṣas*, kobold-like keepers of wealth, with a king, Kubera; *vetāla*s, ghoulish pranksters who haunt corpses; and spirits of the restless dead (*bhūta*s, *preta*s), who must be warded off. Though the existence of these spirits is fully recognized in the classical mythology, they remain on a folk level.

Thus, it is often impossible to draw a precise line between gods of the great (Sanskrit) tradition and those of numerous local or little traditions, since the same god may be universally honoured as a great all-powerful god and still participate in a local tradition as a god of limited function and cult. Conversely, sacred manifestations of purely local interest are associated with the higher mythology by becoming a little manifestation of a great god, *e.g.*, the footstep of Rāma and the bathing place of Sītā (Rāma's spouse).

It is therefore hard to find folk myths and cults that are not in one form or another elevated into the Sanskrit tradition by a specific association with a major god. Only where there has been no appreciable cultural contact between Indian tribal people and Hindus (or "hinduized" folk groups) can a significant distinction be drawn between classical Hindu mythology and "folk" mythology. The latter has been little studied, but the myths of the tribes of Chota Nāgpur (Bihār state), the Santāls (West Bengal, Bihār state), the Todas in the Nīlgiri Hills (Tamil Nadu state), and others are examples of such mythology. Much material lies hidden in folktales, which have been investigated only sporadically.

MODES OF REPRESENTATION

Recitation. Little relevance attaches to a distinction between "written" and "oral" traditions, since both have been in constant interplay. A myth is essentially *told*. Sanskrit uses words like *purāṇa* (ancient story) and *ākhyāna* (illustrative narrative). In the oldest source, the Ṛgveda, myths are not so much told as alluded to; it is in the later Vedic literature of the *Brāhmaṇas* that narratives are found, and these are often prejudiced by the authors' liturgical concerns.

The many forms of Hindu myths

The recitation of certain myths was prescribed for specific rituals. The epic *Mahābhārata* has it that Vedic stories were narrated "in the pauses of the ritual," probably by Brahmins (priests). The warrior class (Kṣatriyas) had their own mythographers in their *sūta*s (charioteers and panegyrists), who celebrated the feats of great rulers. These *sūta*s, who became popular narrators of myth and legend, had their own bardic repertoire, which was extended soon to higher mythology. They and their likes—often wanderers who found a ready audience at a sacrifice or a place of pilgrimage—disseminated the lore.

Such narrators still continue to repeat and embroider their ancient stories of gods, sages, and kings. At an early stage their narratives were dramatized and gave rise to the Sanskrit theatre, where epic mythic themes preponderate, and to the closely related dance, surviving in the now largely South Indian schools of bhārata-nāṭya (traditional dance) and the kathakali (narrative dance) of Kerala.

Depiction in the visual arts. Like literature and the performing arts, the visual arts also contributed to the perpetuation of myths. Hindu sculpture tends to be less narrative than Buddhist, which delights in scenes from the Buddha's lives. In Hindu sculpture the tendency is toward hieratic poses of a god in a particular conventional stance (*mūrti*), which, once fixed, perpetuates itself. An icon is a frozen incident of a myth (*e.g.*, *mūrti* [image] of Śiva is the "destruction of the elephant," in which Śiva appears dancing before and below a bloody elephant skin that he holds up with eight arms before the image of his horrified consort; the stance is the summary and reminder of his triumph over a certain demon). A god may also appear in a characteristic pose while holding in his multitudinous hands his various emblems, by each of which hangs a story. Carvings such as those that appear on temple chariots tend to be more narrative; even more so are the miniature paintings of the Middle Ages. A favourite theme in the latter is the myth of the herd god Kṛṣṇa (Krishna) and his loves for the cowherd wives (*gopīs*), narrated in a series of paintings.

RECURRENT MYTHIC THEMES

Myths of origin. *Cosmogony.* In the Vedic literature there are different but not exclusive accounts of the origin of the universe. The simplest is that the creator built the universe with timber, as a carpenter does a house. Hence there are many references to gods measuring out the different worlds as parts of one edifice, atmosphere upon earth, heaven upon atmosphere. Creation may be viewed as procreation: the personified Heaven, Dyaus (the word is related to the Greek Zeus) impregnates the Earth goddess, Pṛthivī, with rain, so that crops grow on her and are harvested. Quite another myth is recorded in the last (tenth) book of the Ṛgveda: in the "Hymn of the Primeval Person" ("Puruṣa-sūkta," 10.90) it is said that the universe was created out of parts of the body of a single primeval person (*puruṣa*), when his body was immolated and dismembered at the primordial sacrifice.

Creation and eschatology

In the same book of the Ṛgveda, mythology begins to be transformed into philosophy; *e.g.*, "in the beginning was the non-existent, from which the existent arose" (Ṛgveda 10.72.2). Even the reality of the nonexistent is questioned: "then there was neither the nonexistent nor the existent" (Ṛgveda 10.129). Such cosmogenic speculations continue, particularly in the older *Upaniṣads*. Originally there was nothing at all, or Hunger, which then, to sate itself, creates the world as its food. Alternatively, the creator creates himself in the universe by an act of self-recognition, or self-formulation, or self-formation. Or the one creator grows "as big as a man and a woman embracing" (*Bṛhadāraṇyaka Upaniṣad* I.4.3.), splits into man and woman, and in various transformations the couple create other creatures. In one of the last stages of this line of thought (*Chāndogya Upaniṣad* 6.2), there is an ac-

count that was to become fundamental to the ontology of the philosophical schools of Vedānta: in the beginning was the Existent, or Brahman, which through heaven, earth, and atmosphere, the triadic space, and the three seasons of summer, rains, and harvest, the triadic time, produced the entire universe.

Later cosmogony embroiders the ancient notion that the original creation lasted a year and is closely bound up with conceptions of time (see below *Myths of time and eternity*). In the beginning the god Nārāyāṇa (identified with Viṣṇu) floated on the snake Ananta (endless) on the primeval waters. A lotus grows from his navel, in which the god Brahmā is born, who simultaneously recites the four Vedas with his four mouths and creates the "Egg of Brahmā," which contains all the worlds. There are numerous other accounts that refer to demiurges, or creators, like Manu (the primordial ancestor of mankind).

Though the Vedas do not seem to conceive of an end to the world, classical cosmogony accounts for the destruction of the world at the close of an eon, when the Fire of Time puts an end to the universe. Elsewhere the destruction is specifically attributed to the god Śiva, who dances the *tāṇḍava* dance and destroys the world. Yet this end is not an absolute end but a temporary suspension (*pralaya*), after which creation begins again in the same fashion.

Cosmology. The Vedic texts do not show great concern with cosmology, nor are the views expressed consistent. Generally the universe is regarded as three layers of "worlds" (*loka*): heaven, atmosphere, and earth. Heaven is that part of the universe where the sun shines, and is correlated with sun, fire, and ether; the atmosphere is that part of the sky between heaven and earth where the clouds insert themselves in the rainy season, and is correlated with water and wind; earth, a flat disk, like a wheel, is here below, "holder of treasure" (*vasuṃdharā*) and giver of food. Besides this tripartite pattern, there is on the one hand an ancient dual notion of heaven as masculine and father, and earth as feminine and mother; on the other hand, in later stages there is a conception of five elements (ether-space [*ākāśa*], wind [*vāyu*], fire [*agni*], water [*āpas*], and earth [*bhūmi*]), by combinations and permutations of which the universe is formed.

The three worlds of heaven, atmosphere, and earth

In classical Hindu mythology the *Purāṇa* texts present a mythical cosmography of great elaboration. The old tripartite universe persists, but is modified; there are three levels, heaven, earth, and the nether world; the first and last are further subdivided. Earth consists of seven circular continents, the central one surrounded by the salty ocean and each of the other concentric continents by oceans of other liquids. In the centre of the central mainland stands the cosmic mountain Meru; the southernmost portion of this mainland is Bhāratavarṣa, the old name for India. Vertically, above earth there are seven layers in heaven, at the summit of which is the world of Brahmā (Brahmā-loka), and seven layers below earth, the location of hells inhabited by demons of various kinds.

Myths of time and eternity. The oldest texts speak little of time and eternity. It is taken for granted that the gods, though born, are immortal; they are called "sons of Immortality." In the Atharvaveda, Time appears personified as creator and ruler of everything. In the *Brāhmaṇas* and later Vedic texts there are repeated esoteric speculations concerning the year, which is the unit of creation, and thus identified with the creative and regenerative sacrifice and with Prajāpati (lord of creatures), the god of the sacrifice. Time is an endless repetition of the year, and thus of creation, the completion of which took a year. This is the starting point of later notions of repeated creations.

The four *yuga*s of the world

Classical myths develop around the notion of *yuga* (world age), of which there are four. These four *yuga*s, Kṛta, Tretā, Dvāpara, and Kali—they are called after the four throws, from best to worst, in a dice game—comprise a *mahāyuga* (large *yuga*), and, like the comparable ages of the world depicted by the Greek poet Hesiod, are periods of increasing deterioration. Time itself deteriorates, for the ages are successively shorter. Each *yuga* is preceded by an intermediate "dawn" and "dusk." The Kṛtayuga lasts 4,000 years, with a dawn and dusk of 400 years each, or a total of 4,800 years; Tretā a total of 3,600 years; Dvāpara 2,400 years; and Kali (the current one), 1,200 years. A *mahāyuga* thus lasts 12,000 years, which observes the usual coefficient of 12, derived from the 12-month year, the unit of creation. These years are "years of the gods," each lasting 360 human years, 360 being the days in a year. Two thousand *mahāyuga*s form one *kalpa* (aeon), which is itself but one day in the life of Brahmā, whose full life lasts 100 years; the present is the midpoint of his life. Each aeon is followed by an equally long period of abeyance (*pralaya*), in which the universe is asleep. Seemingly the universe will come to its definitive end at the end of Brahmā's life.

Another myth lays particular stress on the destructive aspect of time. Everything dies in time: "Time ripens the creatures, Time rots them" (*Mahābhārata* 1.1.188). So "Time" (*kāla*) is another name for the god of death, Yama. In classical mythology the name is associated especially with Śiva in his destructive aspect as Mahākāla and is extended to his consort, who may be known as the goddess Kālī or Mahākālī. On a mythological level the speculations on time reflect the doctrine of the eternal return in the philosophy of transmigration. The universe returns just as, after death, a soul returns to be born again. In the oldest description of the process (*Chāndogya Upaniṣad* 5.3.1.-5.3.10), the account is still mythic, but with tendencies to naturalism. The soul on departing may go either of two ways: the Way of the Gods, which brings it through days, bright fortnights, the half year of the northern course of the sun, to the full year, and eventually to Brahmā; or the Way of the Manes (spirits), through nights, dark fortnights, the half year of the southern course of the sun, and, failing to reach the full year, eventually back to earth clinging to raindrops. If the soul happens to light on a plant that is subsequently eaten by a man, the man may impregnate his woman and thus the soul is reborn. Once more the significance of the year as a symbol of complete time is clear.

Myths of the gods. According to the epic *Mahābhārata* (1. 1.39), there are 33,333 Hindu deities. In other, later sources, that number is multiplied a thousandfold.

Vedic myths. Generally speaking, Vedic gods lack the clear delineation of the gods in the Greek pantheon. Some major gods were clearly personifications of natural phenomena, and where natural phenomena were still visible behind the deity, no clearly delineated divine personality was perceived.

Principal Vedic deities

The three most frequently invoked gods are Indra, Agni, and Soma. Indra, the foremost god of the Vedic pantheon, a god of war and rain, is the most anthropomorphic. Agni (a cognate of the Latin *ignis*) is the deified fire, particularly the fire of sacrifice; Soma is the inebriating drink of the sacrifice, or the plant from which it is pressed; neither is greatly personified.

The principal focus of Vedic literature is the sacrifice, which in its simplest form can be viewed as a ritualized banquet to which a god is invited to partake of a meal shared by the sacrificer and his priest. The invocations give occasion to mention, often casually, the past exploits of the deity. The offered meal gives strength to the deity to repeat his feat and to aid the sacrificer.

The central myth of the Ṛgveda is that of Indra killing the dragon Vṛtra, who prevents the monsoon rains from breaking. As the monsoon is the greatest single factor in Indian agriculture, the event celebrated in this myth impinges on everyone's life. In the social circles represented in the Ṛgveda, however, the myth is cast in a baronial mold and the breaking of the monsoon is viewed as a cosmic battle. The entire monsoon complex is involved: Indra is the Lord of the Winds, the gales that accompany the monsoon; his weapons are lightning and thunderbolt, with which he lays Vṛtra low. To accomplish this feat he needs strengthening with soma. Simultaneously, he is the god of war, invoked to defeat the non-Āryan *dasyu*s, the indigenous peoples referred to in the Vedas. These important concerns—the promptness and abundance of the rains, success in warfare, and the Āryan conquest of the land—all find their focus in Indra.

While morality is not at issue in Indra's myth, it is in those of other principal Vedic deities. Central to ancient morality was the notion of *ṛta*, the basic meaning of which appears to have been the truthfulness with which the alliance between men, and between men and gods, was observed, and so insured the physical and moral order of the universe. Varuṇa is an older sovereign god, who with Mitra (related to the Persian god Mithra) presides over the observance of the *ṛta*. Thus Varuṇa is a judge before whom a man can stand guilty, while Indra is a king by warlike conquest. Typical requests that are made of Varuṇa are for forgiveness, deliverance from evil committed by oneself or others, and for protection; Indra is prayed to for bounty, aid against enemies, and leadership against demons and *dasyus*.

The notion of *ṛta*

Distinct from both is Agni, the fire, who is observed in all his multifarious manifestations: in the sacrificial fire, in lightning, or hidden in the logs from which fire may be drilled. As the fire of sacrifice, he is the mouth of the gods and the carrier of the oblation, the mediator between the human and the divine orders. Agni is above all the good friend of the Aryans and is prayed to to strike down and to burn their enemies and to mediate between gods and men.

Among other Vedic gods, only few stand out. One is Viṣṇu, important perhaps more in retrospect than in fact. He is famous for his "three strides," with which he traversed the universe and took possession of it. In his later mythology this pervasiveness, which invites identification with other gods, remains characteristic. His function as helper to the conqueror-god Indra is important.

Impersonality is increased by the prevalence of god pairs and groups of gods. Thus Varuṇa and Mitra are members of the group of Ādityas (sons of Aditi, an old progenitrix), who generally are celestial gods. They are also combined in the double god Mitrā-Varuṇa. Indra and Viṣṇu are combined as Indrā-Viṣṇu. There is also Rudra, an ambivalent god who is dreaded for his unpredictable attacks but is simultaneously benign insofar as he can restrain his attacks. Although there are many demons (*rākṣasas*), no one god embodies the evil spirit; rather, many gods have their devil within, inspiring fear as well as trust. Among the perpetually beneficient gods are the Aśvins (horsemen), who present themselves as helpers and healers and are frequent visitors to the lowly. Almost otiose is the personified Heaven, Dyaus, who most often appears literally as the sky, and often as day. As a person, he is coupled with Earth (in the god pair Dyāvā-Pṛthivī) as a father; Earth by herself is more predominantly known as Mother. Apart from Earth, the other goddess of importance in the text of the Ṛgveda is Dawn (Uṣas), who brings in the day and thus is said to bring forth the sun; some of the more felicitous hymns of the Veda are devoted to her.

In the later Vedic period the significance of the Ṛgvedic gods and their myths began to wane. The peculiar theism of the Ṛgveda, in which any one of several different gods might be hailed as supreme and attributes of one god transferred to another (called "kathenotheism" by scholars), stressed godhead more than individual gods. In the end this led to a pantheism of Prajāpati, the deified sacrifice or ritualized deity; with his consort Vāc (*i.e.*, the speech of ritual recitation), he is said to have begotten the world.

Vedic kathenotheism

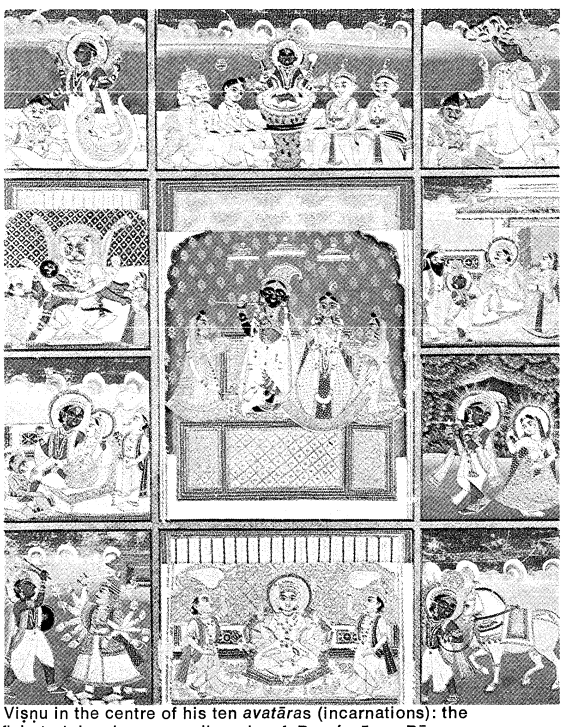
Classical myths. The tendency toward some sort of pantheism (even theism of the moment, as in Vedic kathenotheism) increased in classical Hinduism and led to a kind of polytheistic monotheism in the exaltation of several supreme gods who are not prominently represented in the Vedic corpus, while many of the Vedic gods disappeared or were greatly diminished in stature. New patterns became apparent: the notion of *ṛta*, the basis of the conception of cosmic order, was reshaped into that of *dharma*, the sum total of the religious-social tasks and obligations of social man, which upholds order in the universe; there was a broader vision of the universe and the place of divinity.

In classical mythology three principal moments are envisioned in the life of the cosmos: creation, maintenance, and destruction. Important myths about the gods are tied to these moments. Traditionally, Brahmā is the creator, emanating the universe and simultaneously promulgating the four Vedas from his four mouths. The conception of time as almost endlessly repeating itself in aeons detracts, however, from the uniqueness of the first creation, and Brahmā becomes little more than a demiurge. Far more attention is given to the maintenance and to the destruction of the universe.

Maintenance and destruction are symptomatic of order and disorder, and order and disorder in turn are closely associated with society and the realm outside society. The god Viṣṇu, who became fixed as the god of maintenance, is thus also the social god *par excellence;* while Śiva, partly fixed as the agent of destruction, is in many respects an asocial god. Viṣṇu is the saviour from lawlessness, destroyer of those who threaten the good order, and king of the harmonious realm. Śiva represents untamed wildness, he is the lone hunter and dancer, the *yogin* (the accomplished practitioner of Yoga) withdrawn from society, and the ash-covered ascetic. The distinction between the gods is not between good and evil, but rather between two ways in which the divine manifests itself in this world—as both benevolent and fearful, both harmonious and disharmonious.

What little mythology is attached to Viṣṇu consists largely of the mythology of his incarnations (*avatāra*). Though the notion of "incarnation" is found elsewhere in Hinduism, it is basic to Vaiṣṇavism. The concept is particularly geared to the social role of Viṣṇu; whenever *dharma* (universal law and order) is in danger, Viṣṇu departs from his heaven, Vaikuṇṭha, and incarnates himself in an earthly form in order to restore the good order. In each of these incarnations there is a particular mythology.

By courtesy of the Victoria and Albert Museum, London



Viṣṇu in the centre of his ten *avatāras* (incarnations): the fish, tortoise, boar, man-lion, dwarf, Paraśurāma, Rāma, Kṛṣṇa, Buddha, and Kalkin. Painting from Jaipur, India, 18th century; in the Victoria and Albert Museum, London.

The classical number of these incarnations is ten, ascending from theriomorphic (animal form) to fully anthropomorphic (human form) manifestations. These are: Fish (Matsya), Tortoise (Kūrma), Boar (Varāha), Man-Lion (Narasiṃha), Dwarf (Vāmana), Rāma-with-the-Ax (Paraśurāma), King Rāma, Kṛṣṇa (Krishna), Buddha, and the future incarnation, Kalkin. Among these, Rāma and Kṛṣṇa are singled out here to convey the kinds of myths that are typical of the god-type of Viṣṇu.

The ten *avatāras* of Viṣṇu

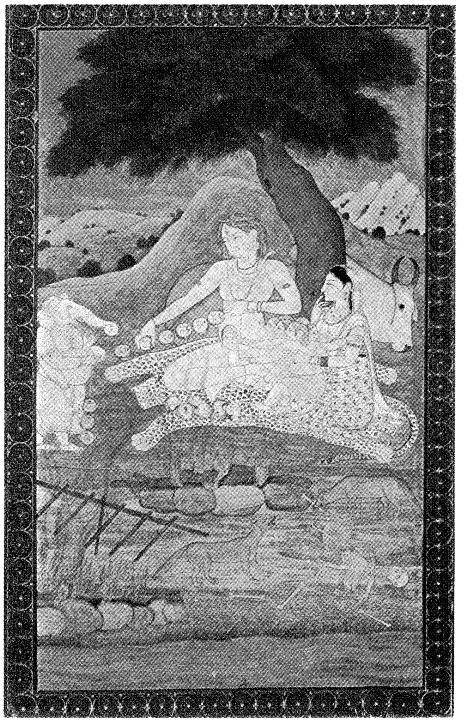
The classical narrative of Rāma is recounted in the Sanskrit epic *Rāmāyaṇa* by the sage Vālmīki, the traditional author of the epic. Rāma is deprived of the kingdom to which he is heir and exiled to the forest with his wife Sītā and his brother Lakṣmaṇa. While there, Sītā is abducted by Rāvaṇa, the demon king of Laṅkā. In their search for Sītā, the brothers ally themselves with a monkey king whose general, Hanumān (later to become a monkey deity), finds Sītā in Laṅkā. In a cosmic battle, Rāvaṇa is defeated and Sītā rescued. When Rāma is restored to his kingdom, the populace casts doubt on Sītā's chastity while a captive. To reassure them, Rāma banishes Sītā to a hermitage where she bears him two sons and eventually dies by re-entering the earth from which she had been born. Rāma's reign becomes the prototype of the harmonious and just kingdom, to which all kings should aspire. Rāma and Sītā set the ideal of conjugal love; Rāma in his relationship to his father, that of filial love; Rāma and Lakṣmaṇa that of fraternal love. Everything in the myth is designed for harmony, which after being disrupted is at last regained.

Kṛṣṇa (Krishna), the most important of the incarnations of Viṣṇu, went through a complex development as a god-hero. In the *Mahābhārata* he is primarily a hero, a chieftain of a tribe, and an ally of the Pāṇḍavas, the heroes of the *Mahābhārata*. He accomplishes heroic feats with the Pāṇḍava prince Arjuna. Typically he helps the Pāṇḍava brothers to settle in their kingdom, and, when the kingdom is taken from them, to regain it. In the process he emerges as a great teacher who reveals the *Bhagavadgītā* (part of book six of the *Mahābhārata*), the most important religious text of Hinduism, which, again, argues for *dharma*. In the further development of the Kṛṣṇa myth, this dharmic aspect recedes and makes way for an idyllic myth about Kṛṣṇa's boyhood, when he plays with and loves young cowherd women in the pasture while hiding from an uncle who threatens to kill him. The influence of this idyll on art has been profound.

A god thus active for the good of society and the individual inspires love. Viṣṇu has indeed been the object of devotional religion (*bhakti*) to a marked degree, but again mainly in his incarnations, and among them especially as Kṛṣṇa and Rāma. The god returns devotion with his grace, through which the votary may be lifted from transmigration to release. Like most other gods, Viṣṇu has his especial entourage: his wife is Lakṣmī or Śrī, the lotus goddess, granter of beauty, wealth, and good luck. She came forth, not unlike Venus, from the ocean when gods and demons churned it in order to recover from its depths the ambrosia, *amṛta*. At the beginning of the commercial year special worship is paid to her for success in one's affairs. Viṣṇu's mount is the bird Garuḍa, arch enemy of snakes, and his emblems are the lotus, club, discus (as a weapon), and a conch shell, which he carries in his four hands.

Śiva and his consorts

In contrast to this helpful, social god-type, Śiva represents the unpredictability of divinity. In him the Vedic Rudra is partly continued, but his mythology has become exceedingly complex: he is the hunter who slays and skins his prey and dances a wild dance while covered with the bloody hide. Far from society and the ordered world, he sits on the inaccessible Himalayan plateau of Mt. Kailāsa, an austere ascetic, averse to love, who burns Kāma, the god of love, to ashes with a glance from his third eye—the eye of insight beyond duality—in the middle of his forehead. Yet another epiphany is that of the *liṅga*, an upright rounded post, usually of stone, a formalized phallic symbol, in which form he is worshipped all over India. And at the end of the aeon, he dances the universe to destruction. He is, nevertheless, invoked as Śiva, Śambhu, Śaṅkara (all meaning: "the Auspicious One"), for the god that can strike can also spare. Snakes seek his company and twine themselves around his body. He wears a necklace of shells. He sits in meditation, with his hair braided like a hermit's, his body smeared white with ashes. These ashes recall the burning pyres, on which the *saṃnyāsin*s ("renouncers") take leave of the social order of the world and set out on a lonely course toward release, carrying with them a human skull.



Śiva and his family at the burning ground. Pārvatī, Śiva's wife, holds Skanda while watching Gaṇeśa (left) and Śiva string together the skulls of the dead. The bull, Nandi, rests behind the tree. Kangra painting, 18th century; in the Victoria and Albert Museum, London.

By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper

Like so many ascetics—often irascible and dangerous—Śiva demands to be seduced. His consort is Pārvatī (Daughter of the Mountain), a goddess most unlike the consorts of Viṣṇu in his various incarnations. She is also personified as the Goddess (Devī "goddess"), Mother (Ambā), black and destructive (Kālī), fierce (Caṇḍikā), and well-nigh inaccessible (Durgā). As Śiva's female counterpart, she inherits some of Śiva's more fearful aspects. She comes to be regarded as the power (Śakti) of Śiva, without which Śiva is literally powerless. Śakti is in turn personified in the form of many different goddesses, often said to be aspects of her.

The spheres of the Viṣṇu complex and the Śiva complex are thus very different ones. In important respects they represent the two different ethics of Hinduism: the *dharma* ethic, which aims at upholding the *dharma* and the cosmic and social order based on it, and the *mokṣa* (liberation) ethic, which searches for release from an order which perpetuates transmigration.

Folk and tribal myths. There is a tremendous diversity in folk and tribal mythology throughout the whole Indo-Pakistan subcontinent, but these myths have neither been fully collected nor systematically studied. Among locally important deities, Manasā, a snake goddess, worshipped in Assam and Bengal to ward off snake bites and secure prosperity, has an enormous mythology of her own. In South India one finds popular cobra cults with a variety of myths and lore. In Mahārāshtra, a form of Viṣṇu, known as Viṭṭhal or Viṭhobā, who has an important temple at Pandharpur, has also spawned a rich mythology.

Cultural heroes and god-types

Myths of cultural heroes. A cultural hero can easily be assimilated to a god-type by identifying him with an incarnation of a god. Thus great religious teachers are considered manifestations of the god of their devotional preaching, and their lives become part of mythology. There is another mythology concerning great ascetics that is very rich. These ascetics are practically gods on earth, who have amassed tremendous powers that they are not loathe to use. The sage, Kapila, meditating in the nether world, burned to ashes 16,000 princes who had dug their

way to him. Another sage, Bhagīratha, brought the River Ganges down from heaven to sanctify their ashes, in the process creating the ocean. Agastya, revered as the Brahmin who brought Sanskrit civilization to South India, drank up and digested the ocean. When the Vindhya mountain range would not stop growing, Agastya scaled it and commanded it to cease growing until his return; he still has not returned. Viśvāmitra, a king who became a Brahmin, created a new universe with its own galaxies to spite the gods. It is in such myths that the mythopoeic imagination exults in its sensitivity to the awesome, mysterious, and marvelous.

In myths concerning kings and princes, a prevailing theme is the trial of the son by the father. For example, the ancient king Yayāti had five sons, to whom he wanted to transfer his own senescence for a stipulated period. All refused except the youngest, Pūru, who agreed and as a reward became his father's successor; from him were descended the Pauravas, the line of succession or dynasty in which the heroes of the *Mahābhārata* were later born. The latter heroes too underwent a trial when they were exiled from their newly won kingdom; similarly, Rāma underwent his ordeal in exile. Heroines undergo their own trials, which usually threaten their chastity, as in the case of Sītā in the *Rāmāyaṇa* and Draupadī, the one wife of all five Pāṇḍava brothers, whose sari became endless when a lustful villain attempted to pull it off.

Moving from myth to legend, there are also stories told of the great teachers, and every founder of a sect is soon deified as an "incarnation" of a god: the philosopher Śaṅkara (*c.* 788–820) as an incarnation of Śiva, the religious leader Rāmānuja (d. AD 1137) as that of Nārāyaṇa-Viṣṇu, the Bengal teacher Caitanya (1486–1533) as that of Kṛṣṇa and his beloved Radhā at one and the same time.

Myths of holy rivers and places. Of particular sanctity in India are the perennial rivers, among which the Ganges stands first. This river, personified as a goddess, originally flowed only in heaven until she was brought down by the king, Bhagīratha, in order to purify the ashes of his ancestors. She came down reluctantly, cascading first on the head of Śiva, in order to break her fall, which would have shattered the earth. The Ganges' confluence with the Yamunā at Allāhābād is the most sacred spot in India. Another river of importance is the Sarasvatī, which loses itself in desert; it was personified as a goddess of eloquence and learning.

Every major and many a minor temple and sanctuary has its own myth; how it was founded and what miracles were wrought there. The same is true of famous places of pilgrimage, usually at sacred spots near and in rivers; important among these are Vrindāvan (Brindāban) on the Yamunā, which is held to be the scene of the youthful adventures of Kṛṣṇa and the cowherd wives. Another such centre with its own myths is Gayā, especially sacred for the funerary rites that are held there. And there is no spot in Vārānasi (Banaras) along the Ganges that is without its own mythical history.

HINDU MYTHOLOGY IN CONTEMPORARY INDIA

Myth and modernity

Much of the classical mythology persists today, and the Hindu is exposed to it in his yearly round. Meanwhile, the mass media have made their contributions: the type of motion picture called "mythological" is extremely popular, perpetuating the ancient stories down to the village level, and so are "devotionals" in which an example of *bhakti* (devotion to a god) is illustrated. The radio regularly carries *bhajan*s (devotional songs) and classical South Indian songs, the themes of which are often mythic. Every orthodox Hindu's home has at least a corner set aside as a domestic sanctuary where representations of a chosen deity are placed and *pūjā* (worship) is done with prayers, hymns, flowers, and incense. In richer establishments there will be shrine rooms. Mythic illustrations are favourites in Indian calendar art.

Mythology has adjusted itself effortlessly to modernity. The *āśram* (*āśrama*, retreat) of the mystic and religious leader Aurobindo Ghose (1872–1950) in Pondicherry, dedicated to the Mother Goddess (personified by this group as a single principle), is an extremely modern establishment complete with tennis courts. There arise new temples, like a recent one in Vārānasi, in which the entire *Rāmacāritmānas* ("Holy Lake of the Acts of Rama") of the poet Tulsīdās is written out and important scenes etched on mirrors. This same favourite poem is the basis of the annual celebration of the Rām Līlā (the play of Rāma) in northern India, in which the entire community participates. The same Rāma story was evoked by Mahatma Gandhi when he set the *Rām Rāj* (Kingdom of Rāma) as India's governmental ideal. On occasion, social protest arms itself with myth to make its point. For example, the personality of Karṇa, an antagonist in the *Mahābhārata* who is berated for his low birth, is extolled in intellectual circles as a truer champion than the aristocratic heroes. A Kannada-language play of the 1960s based on the life of King Yayāti enjoyed great popular and critical success. Antinorthern groups in Tamil Nadu State used the story of Rāma, whose expedition against Rāvaṇa is believed by some to be the Aryan invasion of South India, reversing it by abusing Rāma and glorifying Rāvaṇa.

On a popular level, people at temples and fairs are continually reacquainted with their mythological heritage by *pauraṇika*s, tellers of the ancient stories, heirs of the *sūta*s of 3,000 years ago, and no festival ground is complete without tents where the religious and pious are reminded of their myths by saintly speakers, modestly rewarded by fees, but richly by the honour in which they are held.

BIBLIOGRAPHY. Although some of his interpretations are no longer held, A.A. MACDONELL, *Vedic Mythology*, 2nd ed. (1963), is the most convenient collection of all relevant material for the mythology of Vedic times. In his *Religions of Ancient India* (1953), the master of Vedic studies LOUIS RENOU gives a fine account of the Vedic religion, while L.D. BARNETT in *Hindu Gods and Heroes* (1922), treats the Vedic and Hindu gods in a more narrative and selected form than does Macdonell. J. GONDA, *Aspects of Early Viṣṇuism* (1954), spans Vedic and epic mythology in an investigation into the development of Viṣṇu. E.W. HOPKINS, *Epic Mythology* (1915), is a basic inventory of the mythology of the Mahābhārata and *Rāmāyana*. There is no single comprehensive study of Purāṇic mythology, but a representative selection, particularly with regard to the cosmology, may be found in *The Vishńu Puráńa: A System of Hindu Mythology and Tradition*, trans. from the original Sanskrit by H.H. WILSON (1864–77).

(J.A.B.v.B.)

Hindu Sacred Literature

Hindu sacred literature consists of numerous sacred oral traditions and written texts, which in the course of their gradual development and composition have become the sources of Hindu religion and the foundations of Indian society. The scope of what may be sacred is wider to a Hindu than it is to members of most other religions. The sacred for the Hindu is not only that which bears upon the divine but also that which bears upon man's place in the family and in society and upon his relations with others; consequently Hindu sacred literature includes much that might be considered secular in other religious traditions.

The article is divided into the following sections:

I. The concept of a sacred book in Hinduism
II. Nature and varieties of Hindu sacred literature in Sanskrit and Indian regional languages
  Oral traditions
  Written texts
III. Major religious texts: their contents and significance
  Vedic texts
  Mahābhārata and *Rāmāyaṇa*
  The *Bhagavadgītā*
  Purāṇas
  Dharma-śāstra
  Tantras
  Other texts
  Sacred literature in regional languages
IV. Religious uses of the sacred texts

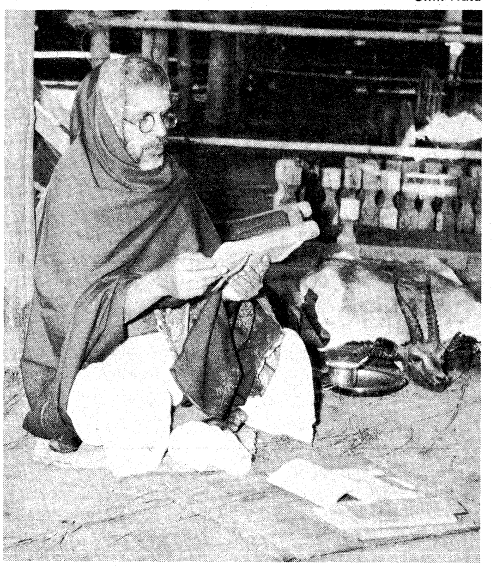
I. The concept of a sacred book in Hinduism

Orthodox Hindu authors commonly divide their sacred literature into two classes, *Śruti* and *Smṛti; Śruti* (literally

Śruti and *Smṛti*

"learning by hearing") is the primary revelation, which stands revealed at the beginning of creation. This revelation was "seen" by the primeval seers (*ṛṣi*) who set in motion an oral transmission that has continued from generation to generation until today. The seers were the founders of the lineages of Brahmins (Hindu priestly elite) through which the texts have been, and continue to be, transmitted. From this heritage the Brahmins derive their function as sacred specialists and teachers. *Smṛti* (literally "recollection") is the collective term for all other sacred literature, principally in Sanskrit, which is considered to be secondary to *Śruti*, bringing out the hidden meanings of the revelation, restating it for a wider audience, providing more precise instructions concerning moral conduct, and complementing *Śruti* in matters of religion. While the distinction between *Śruti* and *Smṛti* is a useful one, in practice the Hindu acquires his knowledge of religion almost exclusively through *Smṛti*.

C.M. Natu



Brahmin priest reading the unbound folio manuscript of a sacred text, a *śrauta-sūtra*, at a Vedic *yajña* (sacrifice). The deerskin is the sacrificer's seat, the wooden cups are for drinking the *soma*.

For the present purposes *Śruti* may simply be identified with the entire Vedic literature, especially the Veda proper, or Saṃhitās (collections of the earliest hymns and ritual formulas), and the appended works: the *Brāhmaṇas* (discussions of the rite), the *Āraṇyakas* ("Books of the Forest"), and the *Upaniṣads* (doctrines concerning cosmic equations). Since the transmission of the Veda was oral, there was no notion that a written copy of it was by itself a sacred object, as for example the Torah scroll in Judaism or the Qur'ān in Islām; instead, it was the sound of Veda that was sacred, and very great weight was placed on the correct pronunciation and recitation of the texts. Also, since transmission took place in Brahmin families, it was largely out of the reach of other classes, who, if they were qualified to hear the Veda at all, would hear it only at the performance of rituals.

Orthodox Hindu thought has addressed itself to the question of where the authoritativeness of the Veda is vested. It does not accept the position that the Veda is God's word, since that would involve possible divine fallibility. It concludes that the Veda is self-authenticating, and only in matters that are completely suprasensual. It is accepted that this authoritativeness applies in two areas: (1) injunctions to perform certain rituals that lead to results not ascertainable by other means of knowledge and (2) statements on the nature of the self.

Accordingly, the Veda is divided into two parts: the "part of ritual" (*karma-kāṇḍa*) and the "part of knowledge" (*jñāna-kāṇḍa*), the former comprising all the liturgical statements of the Veda, the latter encompassing principally the *Upaniṣads*.

While the Veda is sacred of itself, so that mere recitation of it is meritorious, in the *Smṛti* texts it is necessary to distinguish between books that are similarly sacred, and books about sacred matters. There is also a distinction between texts that are sacred as a whole or only in some of their parts. For example, in regard to cultural epics, the *Rāmāyaṇa* ("Romance of Rāma") can be recited for great personal merit in its entirety as a religious observance; this is not true of the *Mahābhārata* ("Great Epic of the Bhārata Dynasty"). On the other hand, the *Mahābhārata* contains the *Bhagavadgītā* ("The Lord's Song"), which is the single most important *Smṛti* text.

Criteria of sacredness

Generally speaking, a book is sacred in Hinduism if it is widely believed or believed by particular religious groups that recitation and listening bestow a special merit. Thus many texts mention the "reward of listening" (*śravanaphala*), a specific merit proper to them. This qualification applies not only to the Sanskrit tradition but also to works in India's regional languages. In addition, a book composed by or hallowed by the use of the founder of a particular sect will be sacred to that sect; for example, the Granth Sāhib (the sacred scripture, Ādi Granth, personified) of the Sikhs (a religious and social group often regarded as being outside Hinduism), which is little known outside the sect, is not typical in that it is physically the object of cult veneration, not unlike the Qur'ān.

Another criterion for the relative sacredness of a text is the ritual use to which it is put. For example, Tulsīdās' entire *Rāmacaritmānas* ("Holy Lake of Rāma's Feats," a Hindī version of the *Rāmāyaṇa* romance) is recited and simultaneously acted out as part of the annual Rām Līlā festival. Similarly, popular litanies might become sacred by their constant use in devotional singing. Likewise certain sacred places of pilgrimage may be celebrated in "glorifications" (*māhātmyas*), which in turn can be adopted into semi-epical and encyclopaedic works called *Purāṇas* ("ancient stories"). Also of considerable importance in the daily life of the Hindu is astrology (see ASTROLOGY).

II. Nature and varieties of Hindu sacred literature in Sanskrit and Indian regional languages

ORAL TRADITIONS

The most sacred of the texts of Hinduism, the Veda, was transmitted orally. Many other Hindu sacred texts were also transmitted orally before, and after, they were written down. For example, the varieties of both oral and written versions of the Rāma story in Indian regional languages and Javanese side by side with the written Sanskrit *Rāmāyaṇa* of Vālmīki show that the oral tradition was not closed with the written Sanskrit composition.

For many religious sects, the manner of the Vedic transmission was the prototype of their own transmission. Followers of sects will speak of the "handing down" (*sampradāya*) or of a "succession of gurus or teachers" (*guruparamparā*), and it is well known that written texts contain only part of the doctrines actually handed down in a sect. This situation is not merely to be ascribed to secretiveness or even to sectarian exclusiveness but also to the nature of Hindu culture. Thus only one who is an initiate in a particular sect can have full knowledge of the beliefs and practices that pertain to it. A novice apprentices himself to a guru (religious teacher) and only after many years of association does he become fully versed in the sacred doctrines of his particular tradition.

Mantra

Secrecy will often attach to what are considered the fundamentals of a sect; for example, the esoteric significance of a particular *mantra* (spell or sound-symbol), or mystical interpretations of events in the biography of a certain god, the precise qualities of his virtues, are rarely made public. Likewise, in temple ceremonies, where every temple has its own style and tradition, the procedures and their significance are almost exclusively preserved orally. Oral tradition was probably strongest on the more popular levels of Hinduism, partly because some of the membership might have been illiterate, and partly to conceal what might be considered aberrant by other groups. Consequently, some of the texts of Tan-

trism (a syncretistic religious movement from *c.* 6th century AD on) are practically unintelligible without the aid of informants who are conversant with its unwritten tradition.

WRITTEN TEXTS

Vedic texts. The Veda (see Table) for a long period of its existence was an "oral tradition," but its transmission was so well organized and its preservation so perfect that these oral "texts" surpass the written texts in accuracy. Since its content was precisely fixed at an early date and henceforth maintained, it can legitimately be treated as a written text.

While scholars traditionally speak of the four Vedas when they mean the four oldest compilations (Saṃhitās), it is better to follow the Indian tradition and describe as Veda the entire corpus of texts that are built on these Saṃhitās. The following principles of classification are operative in the Vedic corpus: schools, branches of the schools, *Saṃhitās, Brāhmaṇas, Āraṇyakas, sūtras, Vedāṇgas.*

Schools. The Veda as a whole is divided into four complementary parts that are each transmitted by a separate school. There are four Vedic schools: Ṛgveda, Yajurveda, Sāmaveda, and Atharvaveda; often the first three are mentioned together as the "triple Veda." These four Schools derive their distinctions from their respective places in the liturgy. In those rituals in which professional priests were employed, a division of labour took place early: one did the main recitation, another the main manipulations, another chanted hymns, a fourth supervised and did acts of expiation; from these separate functions grew the four schools. Those expert in the Ṛgveda-Saṃhitā recited the hymns (*ṛc*, hence Ṛgveda) contained in that collection; the experts in the actual ritual operations accompanied by sacrificial formulas (*yajus*) formed the Yajurveda-Saṃhitā; the priests who chanted compiled their own Sāmaveda-Saṃhitā (*sāman*, melody); while the priests of the fourth category had their own Atharvaveda-Saṃhitā (*atharvan* is the name of a special priest).

Relation of priestly division of labour to formation of Vedic schools

Branches of the schools. Although the ritual practices of the Vedic priests remained remarkably similar even though scattered in widespread settlements, inevitably local variations occurred. Note was taken of these variations and the varying tradition was recognized as a "branch" (*śākhā*) of its Veda school. In every school of the Veda there are such branches, which are most numerous in the Yajurveda. These branches usually have a somewhat different version of their Saṃhitā, and always somewhat different rules concerning the rites themselves. No superiority is claimed by any one branch in a school over another. In the Yajurveda tradition, for example, one set of branches is described as black, the other as white (the colours serve merely to classify).

Saṃhitā. As pointed out, the Saṃhitā (collection) contains the basic texts of a school. It is on the material incorporated in the Saṃhitā that the four groups of priests, each belonging to a different Veda school, draw in the performance of the ritual. They provide only the hymns, formulas, and chants and give no instruction in the procedures of the ritual.

Brāhmaṇa. To each Saṃhitā are attached a number of prose texts called *Brāhmaṇa* (discussion concerning *Brahman*, or the power of word and rite), usually voluminous texts which enlarge on the ritual, setting forth its procedures and at times explaining its provenance in mythical terms. These are distributed according to the various branches; thus there are two *Brāhmaṇas* belonging to the Ṛgveda.

Āraṇyakas. More recent, but continuing the *Brāhmaṇa* tradition, are the *Āraṇyakas*, "Books of the Forest," in which the speculative element grows more important. They take their name from discussions and practices that were too important, or dangerous, to be held in the village. They may contain the *mantras* (spells) and descriptions of rituals not treated in the *Brāhmaṇas.* For example, the *Taittirīya Āraṇyaka* (Taittirīya is the name of a branch of the Yajurveda) gives the material for the mystical *Pravargya* rite, which in another branch is treated in the *Brāhmaṇa.*

Upaniṣads. In the *Upaniṣads*, which form an advanced stage of the *Āraṇyakas*, the speculative element predominates and the rite is simply the starting point of the speculation, if it is mentioned at all. They are of comparatively limited size, the few larger ones not exceeding 2,000 lines, and some consisting of only a few pages. Again they divide over the branches, but there is far more flexibility here: the same speculations may occur in texts belonging not only to different branches but to different schools.

Vedāṅgas (literally "studies accessory to the Veda"). The intense preoccupation with the liturgy gave rise to scholarly disciplines that were part of the Vedic erudition. There were six such "fields": (1) *śikṣā* (instruction), viz., in the proper articulation and pronunciation of the Vedic texts. Different branches may have different ways of pronouncing the texts, which are written up in *prātiśākhyas* (literally, "instructions for the *śākhās*"—branches), four of which are extant; (2) *chandas* (metre), of which there remains only a late representative; (3) *vyākaraṇa* (analysis and derivation), in which the language is grammatically described—Pāṇini's famous grammar (*c.* 600 BC) and the *prātiśākhyas* are the oldest examples of this discipline; (4) *nirukta* (lexicon), which discusses and gives meanings for difficult words, represented by the *Nirukta* of Yāska (*c.* 600 BC); (5) *jyotiṣa* (luminaries), an inchoate astronomy and astrology for fixing the right times of a ritual; and (6) *kalpa* (mode of performance), which studies the correct ways of performing the ritual.

Early scholarly disciplines

Sūtras. It is under the rubric of *kalpa* or procedures that a large number of texts called *sūtras* ("brief manuals") belong. They are manuals, written with great economy of language, explaining in detail how the office of the priests of the various schools and branches should be conducted. There are three types of *sūtras:* (1) *śrauta* (scriptural), which describe the kind of rites called *śrauta*, for which a sacrificer engages priests belonging to the different schools; (2) *gṛhya* (domestic), which describe principally the life-cycle rites; and (3) *dharma* (life rule seen as law), which lays down the life rules and conduct for any given branch.

The accompanying table of Vedic texts illustrates the relations of the various schools, branches, and texts to each other.

Post-Vedic Sanskrit texts. The post-Vedic sacred literature in Sanskrit is extremely voluminous. It does not have the authority of the Veda, or *Śruti;* it presents itself as secondary to the Vedic literature of which it is an amplification. In fact, most of the texts are original. The nontechnical texts can best be summarized under the following headings: (1) epic, (2) *Purāṇa*, (3) *Tantra*, (4) other.

The epics. As the name indicates, the epics are not primarily sacred texts; nevertheless they are important sources of knowledge of Hinduism. Apart from the extremely important sacred text the *Bhagavadgītā* (see below), which it contains, the *Mahābhārata* ("Great Epic of the Bhārata Dynasty") is a rich source of materials for Hinduism. Amid the epic events are inserted many chapters dealing with religion, especially in the 3rd, 12th, and 13th books. The *Rāmāyaṇa* ("Romance of Rāma") is both the principal text for Rāma worship and a receptacle for many legends.

Purāṇa. *Purāṇa* (literally "ancient story") is a general name of a long series (traditionally 18) of often voluminous texts that treat in encyclopaedic manner myths and legends, as well as genealogies, of gods, heroes, and saints. They can loosely be divided into three groups: those exalting the god Brahmā, the *Brāhma-, Brāhmāṇḍa-, Brahmavaivarta-, Mārkaṇḍeya-, Bhaviṣya-*, and *Vāmana-Purāṇas;* those devoted to the god Viṣṇu (Vishnu), the *Viṣṇu-, Bhāgavata-, Nāradīya-, Gāruḍa-, Pādma-*, and *Vārāha-Purāṇas;* and those devoted to *Śiva* (Shiva), the *Śiva-, Liṅga-, Skanda-, Agni-* (or *Vāyu-*), *Mātsya-*, and *Kūrma-Purāṇas.* The division, however, is an artificia one. Many deal with the same or similar materials.

The Vedic Texts

	Śākhā (branch)	Saṃhitā (compilation)	Brāhmaṇa	Āraṇyaka	Upaniṣads	Śrautasūtra
Ṛgveda	Śakala Bāskala	Śakala Bāskala	Aitareya Kauṣītakī	Aitareya Kauṣītakī	Aitareya Kauṣītakī	Āśvalāyana Śānkhāyana
Yajurveda						
Black	Kāṭhaka Maitrāyaṇa	Kāṭhaka Maitrāyaṇa	Kaṭha	Kaṭha	Kaṭha Maitrāyaṇa	Kāṭhaka Mānava Vārāha
	Taittirīya	Taittirīya	Taittirīya	Taittirīya	Taittirīya Śvetāśvatara	Baudhāyana Vādhūla Bhāradvāja Āpastamba Hiraṇyakeśin Vaikhānasa
White	Mādhyaṃdina Kāṇva	Vājasaneyi Kāṇva	Bṛhad	Śatapatha	Bṛhadāraṇyaka Īsā	Kātyāyana
Sāmaveda	Kauthuma-Rāṇāyanīya	Kauthuma-Rāṇāyanīya	Pañcaviṃśa	Āraṇyakasaṃhitā	Chāndogya	Lātyāyana Drāhyāyana Ārseyakalpa
	Jaiminīya	Jaiminīya	Jaiminīya Śātyāyana	Jaiminīya-Upaniṣad-Brāhmaṇa	Kena	Jaiminīya
Atharvaveda	Śaunaka	Śaunaka	Gopatha		Muṇḍaka Māṇḍūkya Praśna	Vaitāna
	Paippilāda	Paippilāda	Gopatha		Muṇḍaka Māṇḍūkya Praśna	Vaitāna

The most important *Purāṇas* are the *Viṣṇu-*, *Liṅga-*, *Bhāgavata-*, and *Skanda-Purāṇas*. This purāṇic literature continues with *Upapurāṇas* ("sub-Purāṇas") and *Māhātmyas* (glorifications) of temples and sacred places.

Tantra. *Tantras* (literally "looms") are texts exemplifying religion on a more popular level, dealing largely with spells (*mantras*), ritual (particularly temple ritual), and drawn symbols, (*maṇḍalas*). There are Vaiṣṇava, Śaiva, and Śākta *Tantras*, devoted respectively to the worship of Viṣṇu, Śiva, and the goddess Śakti (Shakti). Their metaphysical contents are meagre, but they contain a wealth of practical information. In some, particularly the Śākta *Tantras*, evidence of so-called left-handed practices (esoteric, magical, and sometimes sexual) is found, but the incidence of this material is often exaggerated.

Other. As Sanskrit literature is permeated with religion, many texts, even in the belles lettres, are at times sources for knowledge of the religion. Among them the most important one is Jayadeva's poem the *Gītagovinda* ("Poem Which Sings of the Herd God").

Dravidian texts. *Tamil.* Of the Dravidian languages, Tamil especially offers a variety of important texts. The classical Tamil poetry of the Saṅgam ("academies"; from *c.* AD 200) do not speak of religion. With the exception of the *Tirumurukārruppaṭai* (a "Guide to the Worship of the God Muruganˉ") the earliest Tamil literature is unconcerned with religion. From the 7th century AD on, a rich literature of devotion, both to Śiva and Viṣṇu, emerges.

Nayanārs and Āḻvārs

The hymns of the Śaiva saints, the Nāyaṇārs, are collected in the *Tevāram* ("A Garland of Honey [for Śiva]"), which since AD 1000 has formed the basis of south Indian Śaivism. The collection comprises the songs of Appar, Sambandar, and Sundarar. The *Tevāram* itself heads a series of 11 sacred books, among which the eighth, the *Tiruvācakam* ("Message of the Loved One"; *i.e.*, Śiva) of Māṇikkavācakar, is the most famous. In the 12th century the canon and hagiography is established with the *Periyapurāṇam* (lives of 63 Śaiva saints) of Cēkkiḻār, in which the biographies of the saints are worked out. Of importance is the *Kandapurāṇam* of Kacciyappar, which is parallel to the *Skanda-Purāṇa* in Sanskrit.

Like Śaivism, Vaiṣṇavism also developed a devotional (*bhakti*) movement with its own saints, the Āḻvārs, whose hymns were collected by Nāthamuni in the *Nālāyiram* (4,000 poetical compositions). Among the Vaiṣṇavas the most famous saint is Nammāḻvār (*c.* 800). The later period produces important translations of the epics, especially the *Rāmāyaṇa* version by Kambaṇ. Also Tamil is used as a vehicle for philosophy and theology, notably that of the Śaivasiddhānta. From the 13th to 18th centuries there is a line of Cittars (from Sanskrit *siddha*, "saint"), whose verses continue to be extremely popular among the Tamil-speaking people.

Kannada. Much of the Kannada sacred literature is associated with the Liṅgāyats, a Śaivite sect, founded by Bāsava (*c.* 1150) whose lyric *vācanams* (talks) are important.

Telugu. Much of the religious literature in Telugu centres on translations of the *Mahābhārata* and *Rāmāyaṇa*.

Texts in modern Indo-Aryan languages. *Bengali.* The oldest texts in Bengali are the hymns of the *Caryācaryāviniścaya*, ("decision concerning good and ill conduct") the *Dohākośa* ("Collection of Stanzas") of Saraha, the *Dohākośa* of Kaṇha, and the Tantric *Dākārṇava* (700–1000). The doctrines, those of the Sahajiyā sect (a school in Orissa and Bengal since the 10th century which seeks an equilibrium of self and universe through Yogic practices), are in certain ways a Buddhist hybrid. Later, Kṛṣṇaism (worship of the god Kṛṣṇa, or Krishna) predominates, notably in Caṇḍīdās' *Śri-Kṛṣṇa-kīrtan* ("Glorifications of Lord Kṛṣṇa") and *Padāvali* ("verses") (*c.* 1375) and his contemporary Vidyāpati's poetry (*c.* 1400). In that same period, Kṛttibās Ojhā made the most celebrated Bengali version of the *Rāmāyaṇa*. Nothing of the religious reformer Caitanya (1485–1533) is extant, but his doctrines and life brought about a whole literature. Noteworthy sources for Viṣṇuism are the Brajbuli poets, who glorify the child and cowherd Kṛṣṇa. Religious inspirations have also been ascribed to the works of Rabindranath Tagore, the 19th- and 20th-century Bengali poet, whose poems and songs some have elevated to the level of hymns.

Hindi. Hindi religious literature begins with an early ascetic Gorakhnāth, the authenticity of some of whose works remains doubtful, as does his date. The first great name is that of Rāmānanda (*c.* 1450) whose influence has been profound. He adored Viṣṇu in his incarnation as Rāma. Among his pupils were Kabīr (1440–1518), a social reformer who sought to reconcile Islām and Hinduism. Mīrābāī, a princess and a poetess (1503–73), composed the praises of Kṛṣṇa that are particularly famous. Similarly a devotee of Kṛṣṇa was Sūrdās (1483–1563), whose *Sūrsāgar* ("Ocean of the poems of Sūr") celebrates the young Kṛṣṇa. The predominant figure in Hindi devotional poetry is Tulsīdās (died 1623), whose *Rāmacaritmānas*, a *Rāmāyaṇa* version, is still popular today and annually enacted at the Rām Līlā celebrations of the Daśahrā. Dayananda Sarasvati, the founder of the Arya Samaj, used Hindi in sermons and organizational work.

Marathi. The first important author is Jñānadeva, also known as Jñāneśvar (*c.* 1280), who wrote a famous versified commentary on the *Bhagavadgītā.* Nāmdev (*c.* 1450?) wrote devotional poetry to Viṣṇu as the god Viṭhobā of Pandharpur. The tradition is continued by Tukārām (1607–49) whose 1,300 hymns are the culmination of *bhakti* (devotional) sentiment. In the modern period Marathi literature is distinguished by the *Gītarahasya* of B.G. Tilak (1856–1920) in which the action-philosophy of the *Bhagavadgītā* was stressed in the Indian struggle for independence.

III. Major religious texts: their contents and significance

VEDIC TEXTS

Ṛgveda-Saṃhitā. The Ṛgveda is essentially an anthology of poems addressed to gods, spirits, or deified ritual objects, such as the plant extract *soma.* It contains 1,028 hymns with a total of 10,562 lines, which are distributed over ten books, of which the first and the last are the most recent. A hymn usually goes through three phases: it begins with an exhortation that is followed in the main part by praise of the deity, prayers, and imploration, with frequent references to the deity's mythology, and finishes with a specific request. The most frequently addressed gods are Indra and Agni (the sacrificial fire), to whom are devoted 250 and 200 hymns respectively. While many hymns were eventually used in the liturgy, they were not primarily composed for this purpose but were rather inspired by the ritual. Thus a great number of hymns glorify the *soma* (both as plant and the extracted inebriating juice), the fire, and other ritual objects. Though no doubt the personality of Indra dominates, there is a general tendency of the poets to extol whatever god is being addressed above all other deities. Epithets or even individual feats of a particular god can also be applied to others.

The tenth book contains more varied materials, with some hymns referring to domestic rites, such as a marriage hymn and a set of funeral hymns. This book also contains a number of cosmogonic and cosmological hymns of great interest. In 10:90 the creation of the world is described as the sacrifice and dismemberment of a primeval "man" (*puruṣa*); out of parts of his body the components of the universe are formed. This hymn represents the social model of the four classes (*varṇa*), with the Brahmin emanating from his mouth, the Kṣatriya (kingly and warrior class) from his arms, the Vaiśya (merchants) from his thighs, the Śūdra (low castes) from his feet. This text remains the basic authority for the four classes.

Mythical origin of castes

Elsewhere, creation is described as the emergence of the Existent (*sat*) from the Nonexistent (*asat*), a notion refined in 10:129 in which the world is said to emerge from a chaos beyond either Existence or Nonexistence. Many of these notions are also found in the *Upaniṣads.*

Eventually the significance of the Ṛgveda is found wholly in ritual. The lines of the Ṛgveda are important only in so far as they are recited as part of the liturgy. Their transmission was confined to priestly circles, less interested in their content than in their ritual use. One hears little of the study of them for their own sake, although there are later commentaries, the best known of which is that by Sāyaṇa (*c.* 1400).

Yajurveda-Saṃhitā and Sāmaveda-Saṃhitā. The Yajurveda-Saṃhitā and Sāmaveda-Saṃhitā are completely subservient to the liturgy. The Yajurveda-Saṃhitā contains the lines, usually in prose and brief, with which the executive priest (*adhvaryu*) accompanies his ritual manipulations, addressing the implements he handles and the offering he pours, and admonishing other priests to do their invocations.

The Sāmaveda is a collection of lines from the Ṛgveda (and a very few new ones) that were chanted with certain fixed melodies. Both, unlike the Ṛgveda, were designed for the ritual.

Atharvaveda-Saṃhitā. The Atharvaveda stands apart. It contains 731 hymns—there are also some prose passages—of about 6,000 stanzas divided over 20 books. Books 1–7 contain magical prayers for precise purposes: spells for a long life, cures, imprecations, love charms, prayers for prosperity, charms for kingship and Brahminhood, and expiations for evil committed. They reflect the magical-religious concerns of immediate life interests, on a different level than the Ṛgveda, which glorifies the great gods and their liturgy. Books 8–12 contain similar texts, but also cosmological hymns that continue those of the Ṛgveda and provide a transition to the more advanced speculations of the *Upaniṣads.* Books 13–20 celebrate the cosmic principle (book 13) and present marriage prayers (book 14), funeral formulas (book 18), and other magical and ritual formulas. For knowledge of practical religion and magic, the text is of extreme importance, particularly where it complements the one-sided picture of the Ṛgveda. Many of the formulas had their immediate use when administered by the right priest. Many such rites are laid down in the *Kauśika-sūtra* (manual of the Kauśika family of priests) of the Atharvaveda.

Passage rite ritual formulas

Brāhmaṇas and Araṇyakas. *The Brāhmaṇas* detail and describe the liturgy, invoking, and sometimes inventing, myths to explain the ritual facts. The Brāhmaṇas are of inestimable value because they give account of the actual conduct of the ritual in a detail not to be found elsewhere. Since they are earlier than the *śrauta-sūtras* (ritual manuals), they are the oldest sources for the history of Indian ritual. Little interest is given to the Vedic gods, who are largely treated as ritual accessories. The principal religious concern is with the sacrifice, among which the *soma* sacrifice and its variations hold the first place. The sacrifice is looked upon as the generator of the power that keeps the universe in motion. Elaborate homologies (structural parallels) are devised between the sacrifice, its elements and phases, and the universe and its parts. The universalization of the dynamics of the ritual into the dynamics of the cosmos brings out a universal God-sacrifice, a deity very different from the older gods, who is called the Lord of Creatures (Prajāpati). He is not so much worshipped by means of the sacrifice as generated and regenerated by it.

These cosmic-sacrificial speculations continued in the *Āraṇyakas* (forest books), which contain materials of two kinds: *Brāhmaṇa*-like discussions of rites not believed to be suitable for the village and hence discussed in the "forest," and continuing visions of the relationship between sacrifice, universe, and man. The word Brahman —the creative power of the ritual utterances, which is used to denote the creativeness of the sacrifice as such and which underlies ritual and therefore cosmic order— is prominent in these texts.

Upaniṣads. The *Upaniṣads* continued lines of thought, similar to the *Āraṇyakas,* elaborating on the structure of the human personality and that of the macrocosm, while interest in the ritual receded. Of the *Upaniṣads* the oldest are the *Bṛhadāraṇyaka* and *Chāndogya.* They were loosely composed and not systematic in intention, presenting various accounts of doctrines; a number of them are presented in the form of debates between different sages.

Among early accounts of creation are those that look upon creation as the self-fulfillment of the creator, but the prevailing tone is one of dissatisfaction with a transitory world and the effort is to find irreducible and eternal principles. Macrocosmically, these were found in Brahman or in the Existent (*sat*), microcosmically in breath (*prāṇa*), consciousness, but particularly in the self (*ātman*). In order to shed all transiency man should seek to return to the irreducible first principle from which he derives his being. Macrocosm and microcosm were conceived to be so parallel in structure that in certain places the underlying principles were held to be the same, while elsewhere a measure of distinctness was maintained; but there was no developed systematization of ideas, and the sages were at any time willing to reconsider.

Concept of self and Brahman

Transmigration does not figure prominently in the older texts, but there is an interesting account in *Chāndogya Upaniṣad* 5. After death the migrating self may go two routes, the Road of the Gods, which brings him to Brah-

man; and the Road of the Ancestors, which returns him to earth; his station in a new life is dictated by his previous deeds. Thus the search for the first cause and irreducible principle of being becomes also a quest for release from transmigration. The method was not expressly discussed but was thought of as a reversal of creation. A famous text in *Chāndogya Upaniṣad* 6 holds that the original Existent emanated three elements that by their various combinations produce the world; the return to the Existent is following the process of creation back to its original cause.

There is little theism in the older *Upaniṣads*, although the highest principle may be described as an effulgent God; the main concern was cosmological and protosoteriological rather than popularly religious.

MAHABHARATA AND RAMAYANA

The two epics are quite distinct from the earlier literature. Interest in ritual niceties is minimal, and concerns are eminently practical. Nevertheless, the *Mahābhārata* and, less so, the *Rāmāyaṇa* are important sources; in fact, they are almost the sole documents of Hindu religion from *c.* 400 BC to AD 200, a period that, with the rise of Buddhism, brought incisive changes.

Mahābhārata ("Great Epic of the Bhārata Dynasty"). Distinction should be made between the epic itself and the pseudo-epic or the large blocks of more recent and extraneous texts that have been assimilated into the main story. In the epic the main concern of the heroes is with the *dharma*, the complex of supernaturally sanctioned moral, social, and ritual laws, which are incumbent on an individual according to his station and period of life. While the heroes are still in the process of discovering the *dharma*, in the pseudo-epic there are large portions in which *dharma* is systematically treated, and which become precursors of the later *Dharma-śāstras* (see below *Dharma-śāstras*). Religious practice takes the form of Vedic ritual on official occasions, but principally of pilgrimage, and, to some extent, adoration of gods. Apart from the Bhagavadgītā (see below *Bhagavadgītā*) much of the religious material is found in the Book of the Forest (book 3), in which sages teach the exiled heroes; and in the Book of Peace (book 12), in which the wise Bhīṣma holds forth on religious and moral matters.

Religious contents of the *Mahābhārata*

The Vedic gods have lost importance and survive as figures of folklore. The Brahman of the *Upaniṣads* is popularly personified as the god Brahmā, who presides over the gods and dispenses boons. Of far greater importance is Kṛṣṇa. In the epic he is cast in the mold of a hero, a leader of his people, and an active helper of his friends. His biography as it is known later is not worked out; still the text is the source of early Kṛṣṇaism. Not everywhere, and certainly not by everyone, is Kṛṣṇa considered a god, and even as god his stature is superhuman rather than divine. He is occasionally identified with Viṣṇu, but not significantly. Far more remote than the instantly accessible Kṛṣṇa is Śiva, who is also hailed as the supreme god in several myths recounted of him, notably the Story of the Five Indras, Arjuna's battle with him, and others. The epic is rich in information on sacred places: it is clear that the making of pilgrimages and bathing in sacred rivers constituted an important part of religious life. On occasion, these sacred places are associated with sanctuaries of gods. More frequent are accounts of mythical events concerning the particular place and enriching its sanctity. Numerous descriptions of pilgrimages (*tīrthayātrā*) give the writers opportunities to detail local myths and legends. Besides them, there are countless edifying stories that shed light on the religious and moral concerns of the age. Almost divine are the towering ascetics capable of fantastic feats, whose benevolence is sought and whose curses are feared.

Apart from its influence as a written text, the *Mahābhārata* has made its impact on the ages both in South and Southeast Asia through the continual retelling of its stories in vernacular translations and oral literature. Today the story and its tales are part of the early education of the Hindu, and their influence on Indian and Southeast Asian art has been profound.

Rāmāyaṇa. Vālmīki's *Rāmāyaṇa* ("Romance of Rāma"), the simple story of how Prince Rāma was exiled, the abduction of his wife Sītā by the demon Rāvaṇa, the latter's defeat at the hands of Rāma and his monkey allies led by Hanumān, and Rāma's restoration has had a tremendous impact on the Hindu mind. Rāma is the perfect king, Sītā the perfect wife, Lakṣmaṇa the perfect brother; and Rāma's reign is the epitome of social harmony and prosperity. While not in its oldest form, the *Rāmāyaṇa* identifies Rāma with Viṣṇu as another incarnation (*avatāra*) and remains the principal source for Rāmaism (worship of Rāma). Though not as voluminous as the *Mahābhārata*, the text contains a great deal of comparable religious materials in the form of myths, stories of great sages, and examples of human behaviour.

Its continuing influence as a religious and moral romance is dramatically illustrated by its further history. It was translated into Tamil by Kambaṉ in a poem that is one of the glories of Tamil literature; into Bengālī by Kṛttibās Ojha; and into Hindi in the *Rāmacaritmānas* by Tulsīdās. Even today a continuous reading of the *Rāmāyaṇa* is an act of great merit. A popular enactment of one version is an annual event all over northern India.

THE BHAGAVADGITA

The *Bhagavadgītā* ("The Lord's Song") is the most influential Indian religious text, although it is not strictly *Śruti* or revelation. It is a brief text, 7,000 verses divided into 18 chapters, in quasi-dialogue form. When the opposing parties in the *Mahābhārata* war stand ready to begin battle, Arjuna, the hero of the favoured party, despairs at the thought of having to kill his kinsmen and lays down his arms. Kṛṣṇa, his charioteer, friend, and adviser, thereupon argues against Arjuna's failure to do his duty as a noble. The argument soon becomes elevated into a general discourse on religious and philosophical matters. The text is typical of Hinduism in that it is able to reconcile different viewpoints, however incompatible they seem to be, and yet emerge with an undeniable character of its own. Three different ways of releasing the self from transmigration are set forth. There is the discipline of the task (*karma-yoga*): against the views held by Buddhists and Jains, as well as the Sāṃkhya philosophy, which hold that all acts bind, and that therefore abstention from action is a precondition of release, Kṛṣṇa argues that it is not the acts that bind but the selfish intentions with which they are performed. He argues for a self-discipline in which one does one's duties according to the dictates of one's prescribed tasks (*dharma*), but without any self-interest in the personal consequences of the acts. On the other hand, he does not deny the relevance for a limited group of the discipline of insight (*jñānā-yoga*), in which one seeks release in a yogic (ascetic) course of withdrawal and concentration. Then the tone changes and becomes intensely religious: Kṛṣṇa reveals himself as the Supreme God and grants Arjuna a vision of himself. Still a third, and perhaps superior, way of release is through a discipline of devotion to God (*bhakti-yoga*) in which the self humbly worships the loving God and hopes in release not so much for personal liberation from transmigration but for an eternal vision of God. In response to his votary's devotion God will extend his grace to him, through which he will overcome the bonds of this world.

Theologies of the *Bhagavadgītā*

The *Bhagavadgītā* is not a systematic theological treatise and many different elements, drawn from Sāṃkhya and Vedānta philosophy, are mixed together. Religiously, its important contribution was the new emphasis placed on devotion, which has since remained the governing force in Hinduism. Also the popular theism evidenced elsewhere in the *Mahābhārata* and the transcendentalism of the *Upaniṣads* converge; and a God of personal characteristics is identified with the Brahman of the Vedic tradition. In its three disciplines the *Bhagavadgītā* gives a typology of the three dominant trends of Indian religion: *dharma*-based brahminism, enlightenment-based asceticism, and devotion-based theism.

The influence of the *Bhagavadgītā* has been profound. It was a popular text open to all who would listen and

fundamental for all later Hinduism. Vedānta philosophy recognizes it, with the *Upaniṣads* and the *Brahmasūtras* (brief doctrinal rules concerning Brahman), as the third authoritative text, so that all school founders had to comment on it. Until today, as is evident from the lives of such diverse personalities as the Indian freedom fighters B.G. Tilak (1856–1920) and Mahatma Gandhi (1869–1948) it has continued to shape the attitudes of the Hindu.

PURANAS

It is a characteristic development of Sanskrit religious literature that certain basic texts of religious or moral importance attracted to themselves other texts with similar or related themes. The *Mahābhārata* epic attracted a great number of other texts that grew into a colossal pseudo-epic. A similar development took place with the genre of *Purāṇa* (literally "ancient story"). While it is possible that there was a single original *Purāṇa*, from which the 18 major ones somehow derive, it is more likely that there was a floating Purāṇic repertory of bards and professional reciters. Traditionally this repertory had five themes: creation (*sarga*), periodic recreation (*pratisarga*), genealogy of gods and sages (*vaṃśa*), description of eras (*manvantara*), and the *gestae* ("feats") of the dynasties (*vaṃśānucarita*). Around this core, malleable as it was, adhered a huge number of texts, known as the 18 *Purāṇas*. The *Purāṇas* deal with religious and moral matters in the widest sense of the word for the broadest possible audience. While there are a number of *Purāṇas* without any particular sectarian affiliation, many are devoted specifically to one of the *avatāras* (incarnations) of Viṣṇu, to Śiva, or to Śakti. They document religious concerns in Hinduism from about the close of the *Mahābhārata* (*c.* AD 400) to *c.* 1000. In unsystematic encyclopaedic fashion they describe sacrifices, festivals, rites of expiation, vows, donations, funerary ritual, portents, divination, the construction and erection of icons and temples, places of pilgrimage, etc., all under the general concern for *dharma*.

Bhāgavata-Purāṇa

While all *Purāṇas* have exerted influence on Hinduism —and are in turn reflections of trends in Hinduism—no one can compare in popularity with the *Bhāgavata-Purāṇa* ("The *Purāṇa* of the Devotees of the Blessed Lord Kṛṣṇa"), the *Purāṇa* of the god Kṛṣṇa par excellence. This *Purāṇa* of the Bhāgavatas (those who worship Viṣṇu as the Blessed Lord) is fairly late (*c.* 10th century) and distinguishes itself from the other *Purāṇas* in that it is planned as a unit and that far greater care is taken with both metre and style. It runs to about 18,000 stanzas, divided over 12 books. The most popular part of the *Purāṇa* is the description of the life of Kṛṣṇa, for which it has since remained the principal authority. In this work far greater emphasis than in other texts is placed on the youth of Kṛṣṇa: the threats against his life by the tyrant Kaṃsa, his flight and life among the cowherds at Gokula, and especially his adventures and pranks with the cowherd girls. This treatment has remained classic. The popularity of the text has led to the survival of many manuscripts, some beautifully illustrated. Much of medieval Indian painting and an enormous amount of vernacular literature draw upon the *Bhāgavata-Purāṇa* for their themes.

DHARMA-SASTRA

Among the texts of Vedic inspiration are found the *Dharma-sūtras*, or manuals on *dharma*, which give rules of conduct, rites, and expiations as they were practiced in a number of branches of the Vedic schools (*śākhās*). Their principal contents relate to the duties of men in the life stages of studenthood, householdership, retirement, and asceticism; dietary regulations; offenses and expiations; and the rights and duties of kings. They also address themselves to religious matters, such as purification rites, funerary ceremonies, forms of hospitality, and daily oblations. Finally, they touch on juridical matters. Among these texts the more important ones are the *sūtras* of Gautama, Baudhāyana, and Āpastamba. Although the direct relationship is not clear, this particular concern was continued in the more systematic *Dharma-śāstras*, which in turn have become the basis of Hindu law. They include some of the most important texts of classical India.

Hindu law

First among them stands the *Dharma-śāstra* of Manu (*c.* AD ?200), in 2,694 stanzas distributed over 12 chapters. It deals with various matters such as cosmogony, definition of *dharma*, the sacraments, initiation and Vedic study, the eight forms of marriage, hospitality and funerary rites, dietary laws, pollution and purification, rules for women and wives, royal law, which then leads into more strictly juridical matters, categorized under 18 headings, and then returns to more religious matters, such as donations, rites of reparation, the doctrine of *karman*, the soul, and punishment in hell. Law in the juridical sense is thus completely imbedded in religious law and practice. The framework is provided by the model of the four-class society. The influence of *Dharma-śāstra* has been enormous, as it provided Hindu society with its practical morality. For large parts of the Indian subcontinent Manu's text through its commentaries, notably that of Medhātithi (9th century), has been the lawgiver.

Second only to Manu is the *Dharma-śāstra* of Yājñavalkya, in 1,000 (1,013) stanzas, distributed under the three headings of good conduct, law, and expiation. It has been very influential through its commentary *Mitākṣarā* of Vijñāneśvara (11th century) (see CASTE SYSTEMS).

TANTRAS

Tantra (literally "loom") is the general title of a long series of texts, divisible by sectarian lines, which reflect the beliefs, cults, and practices of medieval India. In many respects they are similar to the *Purāṇas*, but the practical aspect predominates over the philosophical one. Theoretically a *Tantra* deals with (1) "knowledge," or philosophy; (2) *yoga*, or concentration techniques; (3) ritual, which includes the formation of icons and the building of temples; and (4) conduct in religious worship and social practice. In general the last two subjects preponderate, while *yoga* tends to centre on the mystique of certain sound-symbols (*mantras*) that sum up esoteric doctrines. The least developed is the "philosophy," which tends to be a syncretistic mixture of Sāṃkhya and Vedānta philosophic thought, with special and at times exclusive emphasis on the god's power or *Śakti*. The *Tantra* texts divide into three classes: (1) Śaiva *Āgamas* (traditions of the followers of Śiva), (2) Vaiṣṇava *Saṃhitās* ("collections of the Vaiṣṇavas": name borrowed from the Vedic Saṃhitās), (3) Śākta *Tantras* ("the looms of the followers of the goddess Śakti").

Śaiva Agamas. Like much other Hindu sacred literature, this literature is hardly catalogued, let alone thoroughly studied. It is only possible here to summarize classes of texts within the various traditions.

There are four sects of Āgamic Śaivas (Śiva worshippers who follow their own Āgama—"traditional"—texts): the Sanskritic Śaivasiddhāntins; *i.e.*, those who accept the philosophic premises and conclusions of Śaivism, the Tamil Śaivas (in South India), the Kashmir Śaivas (in the extreme north), and the Vīraśaivas or Liṅgāyats (from *vīra*, literally "hero"—signifies their puritanism; *liṅga* is the Śiva emblem they worship in lieu of images). The Śaivasiddhānta traditionally has 28 *Āgamas*, and another 150 sub-*Āgamas*. The principal texts are hard to date; most probably they do not antedate the 8th century. Their doctrine is that Śiva is the conscious principle of the universe, while matter is unconscious. Śiva's power, or Śakti, personified as a goddess, causes bondage and release. She is also the magic Word, and thus her nature can be sought out and meditated upon in *mantras* (sound symbols).

The basic texts of the Tamil Śaivas are the hymns of the Nāyanārs, which are addressed to local manifestations of Śiva and are intensely personal and devout. They date from *c.* 800. Kashmir Śaivism begins with the *Śivasūtra* or "lines of doctrine concerning Śiva" (*c.* 850) as a new revelation of Śiva. The system embraces the *Śivadṛṣṭi* ("A Vision of Śiva") of Somānanda (950) in which emphasis is placed on the continuous recognition of

Śiva; the world is a manifestation of Śiva brought about by his Śakti. The system is called *Trika* (triad), because it recognizes the three principles of Śiva, Śakti, and the individual soul. Vīraśaiva texts begin at about 1150 with the *Vācanams* ("sayings") of Bāsava. The sect is puritanical, worships Śiva exclusively, rejects the caste system in favour of its own social organization, and is highly structured with monasteries and gurus.

Vaiṣṇava Saṃhitās. There are two groups: *Vaikhānasa Saṃhitās* and *Pāñcarātra Saṃhitās*. The latter is the prevailing one; over 200 titles are known, though the official number is 108. *Vaikhānasa Saṃhitās* (collections of the Vaiṣṇava school of *Vaikhānasas*—originally, ascetics) seem to have embodied the original temple manuals for the Bhāgavatas (devotees) but by the 11th or 12th century became supplanted by the *Pāñcarātra Saṃhitās* (collections of the Vaiṣṇava school of *Pāñcarātra*—"the system of the five nights," exact meaning in dispute). The philosophy of the latter is largely a matter of cosmogony, greatly inspired by both the Sāṃkhya and Yoga philosophies.

Notion of *Śakti* in Vaiṣṇavism

Apart from their theology, in which for the first time the notion of *Śakti* is introduced into Vaiṣṇavism, they are important in so far as they give an exposition of Vaiṣṇava temple and cult practices. On the philosophical side it is maintained that the supreme god Kṛṣṇa Vāsudeva manifests himself in four co-equal "divisions" (*vyūhas*), representing levels in creation. These gods emanate as supramundane patrons before the primary creation is started by their Śakti (power). In the primary creation Śakti manifests herself as a female creative force inspired by the Indian Sāṃkhya philosophy's cosmogony. Practically, stress is laid on a new type of "incarnation"—"iconic incarnation"—in which the god is actually present with a portion of himself in a stone or statue, which thus becomes an icon; therefore the icon can be worshipped as God himself.

Śākta Tantras. Śāktism in one form or the other is evidenced since Bāṇa (*c.* 650) wrote his *Hundred Couplets to Caṇḍi* and Bhavabhūti his play *Mālatī-Mādhava* (725), in which Tantric practices are referred to. There is no traditional list of texts that is authoritative; many texts are extant.

Śāktism is an amalgam of Śaivism and folk mother-goddess cults. The Śaivite notion that not Śiva himself but his Śakti (sexual, creative power) is active is taken to its last consequence, that, without Śakti, Śiva is a corpse, and simultaneously Śakti is the creator as well as creation. In the *yoga* part great importance is ascribed to *mantras* (sound symbols), which conjure up the realities with which they are identified. Another important notion (partly derived from Yoga philosophy) is that through the body run subtle canals that carry esoteric powers connected with the spinal cord, at the bottom of which the goddess is coiled around the *liṅga* as Kuṇḍalinī; she is to be made to rise through the body to the top, whereupon release takes place. Important among the Śākta *Tantras* are the *Kulārṇava Tantra* (ocean of Tantrism), which gives details on the "left-handed" cult forms of ritual copulation; the *Kulacūḍāmaṇi* (crown jewel of Tantrism), which embroiders on ritual; and the *Saradātilaka* ("Beauty Mark of the Goddess Śaradā") of Lakṣmaṇadeśika (11th century), which treats of magic.

OTHER TEXTS

Stotras

Since so much of Sanskrit and vernacular literature is suffused by religion, it is often difficult to differentiate between "profane" and sacred literature. Extremely popular was the genre of *stotras*, "hymns of praise," to one god or the other. From 650 date the *Hundred Couplets to Caṇḍī* of Bāṇa and the *Hundred Couplets to the Sun* by Mayūra. Numerous *stotras* are apocryphally attributed to the philosopher Śaṅkara. The most popular collection of all is the *Kṛṣṇakarṇāmṛta* ("Balm to the Ears Concerning Kṛṣṇa") by Līlāśuka Bilvamaṅgala, which contributed, though it originated in the south, a great deal to the development of *bhakti* devotionalism in Bengal.

Most celebrated of all literary-religious works is the *Gītagovinda* of Jayadeva (12th century) on the subject of the loves of Kṛṣṇa and Rādhā. Its structure is an artful combination of the Mahākāvya style of the epyllion (literary short epic) in 12 chapters, interrupted by 24 series of eight-line songs. The style and language is extremely musical, with a great preference for rhyme, alliterations, and assonance.

SACRED LITERATURE IN REGIONAL LANGUAGES

The Sanskrit sources, copious as they are, are further augmented by the vernacular literatures, which in part continue, in part complement them. The vernacular sources, however, are generally less accessible than the Sanskrit ones, because editions are lacking or are defective and harder to survey. Present knowledge of them, therefore, is very imperfect. Generally, except where there was an independent literary tradition, as in Tamil, the texts are either epigonic in character, translating and restating the Sanskrit tradition, or original creations, the latter particularly to be found in the theistic literature of *bhakti*. The original creations tend to adopt a new vernacular medium to put themselves in contrast to the Sanskritic ideals. Thus frequent criticism is made of the caste system, brahmin superiority, the emptiness of *dharma* formalism; and the love of God is exalted as the sole effective means for salvation beyond learning, Veda, and ritual. Also the use of vernaculars encourages a localism not usually found in the Sanskrit sources: the god is seen as the god of a particular locality, temple, or icon. Finally it leads to a greater sense of self-identity; one is not as in the Sanskrit sources recreating what was supposedly always given from the beginning but creating anew. That these original creations themselves become canonic texts for a particular sect, with their own burden of scholastic commentaries that seek to make a system out of them, does not detract from their great value as newer contributions to the sacred literature.

Epigonic texts. "Epigonic" is used for want of a better word, because from the point of view of the earlier literature these texts are indeed epigonic in the sense that they are derived from or inspired by the epics, *Purāṇas*, etc. But as works of literature they may, and often are, not epigonic at all, but new works that frequently stand at the beginning of a new literary and stylistic tradition, though often continuing to borrow some internal features from the earlier literature. From the latter point of view many are more appropriately dealt with under the study of Indian literature. The works of which vernacular versions were most commonly made were the *Rāmāyaṇa*, the *Mahābhārata*, and the *Bhāgavata-Purāṇa;* and the *Mahābhārata*—not in its entirety but its parts. These versions are not literal translations, but free versions in which the authors place their own emphases, different from the original, and from one another. The oldest of the vernacular versions of the *Rāmāyaṇa* is the Tamil one of Kamban (*c.* 12th century), a work suffused with devotion (*bhakti*) and simultaneously a poem of high literary distinction. In Bengali several translations were made, with some interesting and probably authentic variations from the "official" Rāma story by Vālmīki, the best known one by Kṛttibās Ojhā (1450). Equally, if not more, famous is the Hindi version by Tulsīdās (*c.* 1550), entitled *Rāmacaritmānas* ("The Holy Lake of Rāma's Feats"), while a famous translation in Tamil exists from the 18th century from the hand of Villiputturar.

Regional language versions of the *Rāmāyana* and *Mahābhārata*

A Telugu rendering was made by Ranganātha *c.* 1300. The *Mahābhārata* was translated into Bengali (*c.* 1600), into Telugu by Nanayya and Tikkana (13th century). The *Bhāgavata-Purāṇa* was extremely popular both as a text, which was translated frequently (*e.g.*, into Bengali by Maladhar Vasu, 1480), and because it gave the canonical account of Kṛṣṇa's life and especially his boyhood, which is the perennial inspiration of the *bhakti* poets.

In Marathi the teacher Jñānadeva (also known as Jñāneśvar; *c.* 1300) composed a commentary on the *Bhagavadgītā*, which remains a classic in that literature. His work was continued by Eknāth (*c.* 1600), who also composed *bhakti* poetry.

The bhakti lyricists. Although all have their individual genius, the *bhakti* (devotional) lyricists share a number of

common features whatever their language. The premium of a Sanskrit education exacted for authors of Sanskrit texts, which limited them largely to the Brahmin class and thus put a definite stamp on them, was not demanded. Consequently a *bhakti* poet could arise from any class. He brought to his poetry a familiarity with folk religion unknown to or ignored in the Sanskrit texts. The use of the spoken language, even though formalized, made possible immediate expression of an unmediated vision that needed no further context; thus the lyrics are short, intensely personal, and precise. These works illustrate the localistic and reformist tendency evidenced all over India in the vernacular literatures.

Oldest among the *bhakti* lyricists are those of Tamil: they evoke and invoke in particular Śiva and Viṣṇu in numerous local identifications. The Śaiva lyricists (Nāyaṇārs) are Appar, Sambandar, and Cuntarar whose hymns are collected in the *Tevāram* (*c.* 11 century). More or less contemporary were their Vaiṣṇava counterparts, the Āḻvārs, Poykai, Pūtaṇ, Pēyār and Tirumalicaialar Āḻvār, and in the 8th century the poetess Āṇḍaḷ, Periyāḻvār, Kulacēkarar, Tiruppāṇāḻvār, and notably Nammalvar, who is held to be the greatest. The devotion of which they sing exemplifies the new *bhakti* movement that seeks a more direct contact between man and God, carried by a passionate love for the deity who reciprocates by extending his grace to man. These saints also became the inspiration of theistic systematic religion: the Śaivas for the Śaivāsiddhānta, the Vaiṣṇavas for Viśiṣṭādvaita. In Kannada the same movement was exemplified by Basava, whose *vācanams* ("sayings, talks") achieved a very great popularity. His religion, that of Vīraśaivism, was perhaps the most "protestant" of the *bhakti* religions.

Bengali and Hindi devotional texts

A particularly rich tradition developed in Bengal. The inspiration there was largely Kṛṣṇaite, centring on the love of Rādhā, who symbolizes the human soul, for Kṛṣṇa, the supreme God. In this tradition are Caṇḍīdās and the Maithili poet Vidyāpati (*c.* 1400). The greatest single influence was Caitanya (born 1485), who renewed Kṛṣṇaism. He has left no writings but inspired a number of hagiographies, among the more important of which is the *Caitanya-caritāmṛta* ("Nectar of Caitanya's Life") by Kṛṣṇa Dās (born 1517). The religious lyric continues in the so-called *padas* (verses); one of the greatest poets in this *bhakti* genre in which divine love is symbolized by human love is Govinda Dās (1537–1612). The songs of Rāmprasād Sen (1718–75) similarly honour the Śakti as mother of the universe and are still in wide devotional use.

Hindi literature produced its own great religious lyricists beginning with the disciples of Rāmānanda (*c.* 1450), who was a follower of the philosopher Rāmānuja. Among them the most famous is Kabīr (died 1518), whose *bhakti* was nonsectarian. The princess Mīrā Bāī (1503–73) composed elegiacs and praises in honour of Kṛṣṇa, whereas Tulsīdās, apart from his *Rāmacaritmānas*, composed Rāmaite lyrics. Sūrdās (1483–1563), a follower of the Vallabha school of Vedānta, is famous for his Sūrsāgar ("Ocean of the Poems of Sūr"), a collection of poems based on the childhood of Kṛṣṇa, following the account of the *Bhāgavata-Purāṇa*. In the Marāthi tradition Nāmdev (*c.* 1300) celebrated Viṣṇu, particularly in his manifestation as Viṭhobā at the Pandharpur temple; and Tukārām (1607–49), the greatest poet of this literature, sang of the god of love in numerous hymns.

Other regional texts. No exhaustive inventory can be given of all of the major religious texts in the modern vernaculars. An exception may be made for the Bengali *maṅgal-kāvya* literature, which centres on folk deities. *Maṅgals* are distinguished from other forms of religious literature by their contents: a eulogy to the deity to which they are dedicated; often, an account of the conversion or other experiences of the writer; a lengthy, semi-epical account of the *deus ex machina* through which worship of the deity on earth has come about; and an account of other significant activities of the deity on behalf of his devotees. Numerous *maṅgals* have been composed in honour of many local deities, including Manasā (goddess of snakes), Śītalā (goddess of smallpox), Dharma Ṭhakur (or Dharma Ray), Ṣaṣṭhī (goddess of childbearing and infant welfare), and several forms of the Goddess, such as Kālī and Caṇḍī. Most widely known within Bengal is the folk goddess Manasā, a snake deity of surely hoary antiquity, who in the course of time was assimilated to the Śakti of Śāktism. The *Manasā-maṅgal* gives a Purāṇic account of the greatness and powers of the goddess (see HINDUISM, HISTORY OF; SOUTH ASIAN PEOPLES, ARTS OF).

IV. Religious uses of the sacred texts

The various uses to which the texts have been and are put have been indicated, for the most part, under their descriptions. The range of these uses is enormous. The Vedic texts found their use in the huge edifice of ritual that was erected from *c.* 1000 to 600 BC. Their later parts, especially the *Upaniṣads*, became the foundation of the principal Hindu philosophical school, the Vedānta. The epics, on the other hand, were used directly for the instruction and edification of a much larger audience than was intended for the Vedas; and their later role in Sanskrit and vernacular literatures shows sufficiently their effectiveness. Similar instruction was provided by the *Purāṇas*, but here a greater degree of localization sets in. Although they remain Pan-Indian in intention, they may contain glorifications of specific spots; the *Bhāgavata-Purāṇa* is generally believed to be a southern creation evoking Kṛṣṇa traditions from there.

Of immediate impact on family life are those parts of the Veda that deal with the domestic rites, the life-cycle sacraments, the offerings to the dead, the code of behaviour for the four classes of society, and the four stages of life. But also important are the numerous stories of epic and *Purāṇa* in which signal instances of generosity, of paternal and filial love, and of fraternal relations are celebrated. In personal religious life, texts of many kinds find their uses. The caste Hindu is brought up on the ancient stories that become part of the framework in which he experiences life. Special emergencies may necessitate vows in which the recitation of a text is an important part. He may make a particular *mantra* his own as an aid to his meditations.

On public occasions other texts may come into use. The great festivals at the temples call for the use both of the texts on fundamental procedures proper to that temple, and the special recitations for which the festival calls. On such festive occasions there will be professional reciters who may retell in Sanskrit or vernacular languages the old stories of gods, saints, and heroes. At the great pilgrimage places to which millions may flock, sages have their own booths in which they patiently set forth, day after day, the teachings of the *Bhagavadgītā*.

Rām Līlā festival

Perhaps the most dramatic use of a sacred text is that of Tulsīdās's Rāma epic, at the Rām Līlā festival, which is particularly popular in northern India. While the text is continuously recited, actors take the role of the different characters of the *Rāmāyaṇa*, culminating in a pyrotechnic killing and exploding of the giant demon Rāvaṇa, to both the delight and the edification of the public. An entire city can participate in this event. Similar uses of the texts can be found all over the South Asian subcontinent.

Hindu sacred literature thus embraces an enormous repertory of texts of startling antiquity and variety, and touching upon practically every aspect of individual, family, and social life. The texts have provided religion with its substance, philosophy with its principal assumptions, art with its themes, literature with its topics, music and dance with their soul.

BIBLIOGRAPHY. The fullest general description, with an extensive bibliography has been given by J.N. FARQUHAR, *An Outline of the Religious Literature of India* (1922, reprinted 1967); a briefer but more up-to-date treatment may be found in LOUIS RENOU and JEAN FILLIOZAT, *L'Inde classique*, vol. 1 (1947). For the Vedic ritual literature, see A. HILLEBRANDT, *Altindische Ritual-Literatur* (1897).

(J.A.B.v.B.)

Hipparchus

Hipparchus, who flourished between the years 146 and 127 BC, was the greatest astronomical observer of antiq-

uity and an able mathematician. He was born at Nicaea in Bithynia (part of modern Turkey), on the eastern shore of the Propontis (the modern Sea of Mármara). The dates of his birth and death and the place of his death are not known. He carried out his observations in Bithynia, at Rhodes, where he spent much time, and also, it seems, at Alexandria. The year 127 BC is usually cited as the last date known for his actual work, and a French astronomer, Jean-Baptiste-Joseph Delambre (1749–1822), clearly demonstrated that some observations of Hipparchus on the star Eta Canis Majoris could well have been carried out in that year.

Most of contemporary knowledge of Hipparchus is contained in the writings of Strabo of Amaseia (flourished *c.* AD 21) and in the great astronomical compendium, *Almagest* by Ptolemy (flourished AD 127–151). Ptolemy often quotes Hipparchus, and it is obvious that he thought highly of him; indeed, as a result of the slow progress of early science, he speaks of him with the respect due a distinguished contemporary, although almost three centuries separated the work of the two men. It is difficult always to determine to which of them credit is due.

Influence of Hipparchus on Ptolemy

It is certain, however, that in all his work Hipparchus showed a clear mind and a dislike for unnecessarily complex hypotheses. He rejected not only all astrological teaching but also the heliocentric views of the universe that seem to have been proposed, according to Archimedes (*c.* 287–212 BC), by Aristarchus of Samos (flourished *c.* 270 BC) and that were resuscitated by Seleucus the Babylonian, a contemporary of Hipparchus. In this connection, it is necessary to recall that strong arguments had been advanced against the idea of the motion of the Earth, and the general climate of opinion had never been favourable to following up the lead given by Aristarchus. Moreover, the system of movable eccentrics, and that of epicycles and deferents, accounted well for most of the irregularities observed in the motions of the Sun, the Moon, and the planets. These two systems were based on the erroneous belief that all celestial movement is regular and circular, or at least that it is best described in terms of a system of regular motion in circles. In the system of movable eccentrics, the centres of the supposed orbits of bodies around the Earth were themselves revolving around the Earth. In the other, epicycles were small circles theoretically imposed on the great circular orbital paths, which were called deferents. The epicycle–deferent mechanism was found with that of the movable eccentric in Ptolemy's late form of the geocentric system of cosmology. It was, of course, this Ptolemaic geocentric system that was handed down to western European science, but it must be remembered that the views of Hipparchus had a profound influence on Ptolemy, as he himself acknowledged. It was not until the 15th century that regular observations over very long periods showed the geocentric hypothesis to be too complex to be acceptable and Copernicus proposed that the Sun is the centre of the universe.

Few details are known of the instruments that Hipparchus used. It seems likely that he observed with the usual devices current in his day, although Ptolemy credits him with the invention of an improved type of theodolite with which to measure angles.

Precession of the equinoxes. Hipparchus is best known for his discovery of the precessional movement of the equinoxes; *i.e.*, the alterations of the measured positions of the stars resulting from the movement of the points of intersection of the ecliptic (the plane of the Earth's orbit) and of the celestial equator (the great circle formed in the sky by the projection outward of the Earth's equator). It appears that he wrote a work bearing "precession of the equinoxes" in the title. The term is still in current use, although the phenomenon is more usually referred to merely as "precession." This notable discovery was the result of painstaking observations worked upon by an acute mind. Hipparchus observed the positions of the stars and then compared his results with those of Timocharis of Alexandra about 150 years earlier, and with even earlier observations made in Babylonia.

He discovered that the celestial longitudes were different, and that this difference was of a magnitude exceeding that attributable to errors of observation. He therefore proposed precession to account for the size of the difference and he gave a value of 45″ or 46″ (seconds of arc) for the annual changes.

Measured differences in star positions

This is very close to the figure of 50.26″ accepted today and is a value much superior to the 36″ that Ptolemy obtained.

Tropical and sidereal years. The discovery of precession enabled Hipparchus to obtain more nearly correct values for the tropical year (the period of the Sun's apparent revolution from an equinox to the same equinox again), and also for the sidereal year (the period of the Sun's apparent revolution from a fixed star to the same fixed star). Again he was extremely accurate, so that his value for the tropical year was too great by only 6½ minutes.

Star catalog. Observations of star positions measured in terms of celestial latitude and longitude, as was customary in antiquity, were carried out by Hipparchus and entered in a catalog—the first star catalog ever to be completed. Hipparchus measured the stellar positions with greater accuracy than any observer before him, and his observations were of use to Ptolemy and even later to Edmond Halley. To catalog the stars was thought by some of Hipparchus' contemporaries to be an impiety, but he persevered. Hipparchus had been stimulated in 134 BC by observing a "new star." Concluding that such a phenomenon indicated a lack of permanency in the number of "fixed" stars, he determined to catalog them, and no criticism was able to deflect him from his original purpose.

Hipparchus' catalog, completed in 129 BC, listed about 850 stars (not 1,080 as is often stated), the apparent brightnesses of which were specified by a system of six magnitudes similar to that used today. For its time, the catalog was a monumental achievement.

Lunar and solar work. In his work on the Sun and Moon Hipparchus used the observations of others as well as his own. He showed that the system based on movable eccentrics, and that based on epicycles and deferents, were equivalent in the motions they gave for the Sun and Moon and, indeed, for the planets. Both methods gave the position of the Sun correct to within 1′, and Hipparchus rejected the peculiar notion, prevalent in his day, that the Sun moved in an orbit inclined to the ecliptic. Hipparchus also redetermined the inclination of the ecliptic and obtained a value correct to within 5′ of the modern figure.

The motion of the Moon is more complex than that of the Sun, owing to the perturbations that the Moon suffers from both Earth and Sun; in consequence, there are more irregularities to be taken into consideration. Hipparchus satisfactorily accounted for that inequality of the Moon's motion that is now known to be due to the elliptical form of its orbit; he utilized the system of circular epicycles and deferent but proposed that the deferent was inclined at an angle of 5° to the ecliptic. His theory gave reasonably satisfactory results for the motion at Full and New Moon. Hipparchus was dissatisfied however, for, as he appreciated, the errors at quadrature (when the Moon stands at first and last quarters) were too great. He concluded that there was some further inequality in the Moon's motion, but he was unable to discover any means of solving this problem, and he said candidly that he was leaving the solution of this question to those who were to follow him.

Lunar and solar parallax

Hipparchus also attacked the problem of the relative size of the Sun and Moon and their distance from the Earth. It had long been appreciated, of course, that the apparent diameter of each was the same, and various astronomers had attempted to measure the ratio of size and distance of the two bodies. Eudoxus obtained a value of 9:1, Phidias (father of Archimedes) 12:1, Archimedes himself 30:1, while Aristarchus believed 20:1 to be correct. The present-day value is, approximately, 393:1. Hipparchus followed the method used by Aris-

tarchus, a procedure that depends upon measuring the breadth of the Earth's shadow at the distance of the Moon (the measurement being made by timing the transit of the shadow across the Moon's disk during a lunar eclipse). This method really gives the parallax (the apparent change in the position of a celestial body when observed from two different directions), and thus the distance, of the Moon, the parallax for the Sun being too small to give a significant result; moreover the accuracy obtainable for the distance even of the Moon is poor. Dissatisfied with his results, Hipparchus attempted to find the limits within which the solar parallax must lie for observations and calculations of a solar eclipse to agree; he hoped that differences between solar and lunar parallax might thus also be revealed. He obtained no satisfactory result from his efforts, however, and concluded that the solar parallax was probably negligible. At least he appreciated that the distance of the Sun was very great indeed.

Planetary studies. Hipparchus was unsuccessful in forming a satisfactory planetary theory and was a scientist enough to avoid building hypotheses on insufficient evidence. Theon claims that Hipparchus preferred the mathematical device of the epicycle to that of the movable eccentric, but it seems clear that he looked upon such devices purely as mathematical cŏnveniences and was not concerned with the reality or otherwise of these combinations of circular motion. Ptolemy went further in his description of the planetary work of Hipparchus and wrote

> It was, I believe, for these reasons and especially because he had not received from his predecessors as many accurate observations as he has left to us, that Hipparchus, who loved truth above everything, only investigated the hypotheses of the sun and moon, proving that it was possible to account perfectly for their revolutions by combinations of circular and uniform motion, while for the five planets, at least in the writings which he has left, he has not even commenced the theory, and has contented himself with collecting systematically the observations and showing that they did not agree with the hypotheses of the mathematicians of his time. He explained in fact not only that each planet has two kinds of inequalities but also that the retrogradations of each are variable in extent, while other mathematicians had only demonstrated geometrically a single inequality and a single arc of retrograde motion; and he believed that these phenomena could not be represented by eccentric circles nor by epicycles carried on concentric circles, but that, by Jove, it would be necessary to combine the two hypotheses. (J.L.E. Dreyer, *History of the Planetary Systems From Thales to Kepler,* Cambridge University Press, 1906.)

Those who studied this problem before the time of Hipparchus had but one main aim, to explain the irregularities that occurred when a planet was in opposition; *i.e.*, in a direction opposite to that of the Sun. No notice was taken of those irregularities that occurred when a planet was in other parts of its orbit, doubtless because of the habit of observing only around times of opposition. Hipparchus advocated more frequent observations made over longer periods, and he carried out such work himself with an accuracy unsurpassed in his time. Ptolemy used the older observations that Hipparchus had sifted, those that Hipparchus himself had made, and his own before he formulated the planetary views enshrined in his *Almagest.* In his own planetary work Hipparchus adopted the generally accepted order for the Sun, Moon, and planets. With the Earth as the centre, they were, in order from the Earth, the Moon, Mercury, Venus, the Sun, Mars, Jupiter, and Saturn.

Mathematics. It is to be expected that the astronomical work of Hipparchus should have led him to develop certain departments of mathematics. He made an early formulation of trigonometry and tabulated a table of chords—*i.e.*, the length of the line joining two points on a unit circle corresponding to the given angle at the centre; *e.g.*, chord of $\alpha = 2 \sin(\alpha/2)$: he is known to have had a method of solving spherical triangles. It is also generally agreed that the theorem in plane geometry known as "Ptolemy's theorem" was originally due to Hipparchus and was later copied by Ptolemy. The French statesman and mathematician Lazare Carnot (1753–1823) showed that the whole of plane trigonometry can be deduced from these formulas.

Geographical work. Hipparchus criticized severely the geographical work of Eratosthenes (*c.* 276–*c.* 194 BC) and himself did some work in this field. His main contribution was to apply rigorous mathematical principles to the determination of places on the Earth's surface, and he was the first to do so by specifying their longitude and latitude—the method used today. Hipparchus was, no doubt, led to this method by his work on the trigonometry of the sphere. He tried to measure latitude by utilizing the ratio of the longest to the shortest day at a particular place instead of following the customary method of the Babylonians of measuring the difference in length of day as one travels northward. Hipparchus also divided the then known inhabited world into climatic zones, and suggested that the longitude of places could be determined by observing, from these places, the moments when a solar eclipse began and ended; but this bold scheme, while theoretically satisfactory for a small area of the Earth's surface, was not a practical proposition in his day.

BIBLIOGRAPHY. J.L.E. DREYER, *History of the Planetary Systems from Thales to Kepler* (1906; reprinted as *A History of Astronomy from Thales to Kepler*, 1953), a readable but scholarly book in which the work of Hipparchus is clearly set out; GEORGE SARTON, *A History of Science*, vol. 2, *Hellenistic Science and Culture in the Last Three Centuries B.C.* (1959), a volume containing an excellent well-written résumé of Hipparchus and his achievements; PTOLEMY, *The Almagest*, reprinted in an English translation in "Great Books of the Western World," vol. 16 (1952), the main original source of references to the astronomical work of Hipparchus, and the clarity of Ptolemy's text makes clear his contribution.

(C.A.R.)

Hippocrates

Hippocrates of Cos was a Greek physician of the 5th and 4th centuries BC. Referred to from antiquity as "Hippocrates the Great" to distinguish him from others of that name, he was regarded from classical times as the father of medicine. His name is traditionally associated with the so-called Hippocratic Oath—certainly not written by him —which in modified form is still often required to be taken by medical students on graduating.

The Hippocratic tradition

Life. Trustworthy information about Hippocrates' life is scanty. His younger contemporary Plato referred to him twice. In the *Protagoras* Plato called Hippocrates "the Asclepiad of Cos" who taught students for fees and implied that Hippocrates was as well-known as a physician as Polyclitus and Phidias were as sculptors. It is now widely accepted that an "Asclepiad" was not a temple priest or a member of a physicians' guild but instead was a physician belonging to a family that had produced well-known physicians for generations. Plato's second reference occurs in the *Phaedrus.* Hippocrates is referred to as a famous Asclepiad who had a philosophical approach to medicine. Further, Hippocrates regarded the body as "a whole"—that is, as an organism. His medical practice resulted from his collection of information regarding parts of the body into an embracing concept and, thereafter, the division of the whole into its parts.

Meno, a pupil of Aristotle, specifically stated in his history of medicine the views of Hippocrates on the causation of diseases, namely, that undigested residues were produced by unsuitable diet and that these residues excreted vapours, which passed into the body generally and produced diseases. Aristotle said that Hippocrates was called "the Great Physician" but that he was small in stature (*Politics*).

These are the only extant contemporary, or near-contemporary, references to Hippocrates. Five hundred years later, the Greek physician Soranus wrote a life of Hippocrates, but the contents of this and later lives were largely traditional or imaginative. It seems possible, however, that Hippocrates was the son of a physician and was born about 460 BC on the island of Cos. Throughout

his life he travelled widely in Greece and Asia Minor practicing his art and teaching his pupils, and he frequently taught at the medical school at Cos. There is a strong tradition that he died about 377 BC at Larissa. His birth and death dates are traditional but may well be approximately accurate. Undoubtedly Hippocrates was a historical figure, a great physician who exercised a permanent influence on the development of medicine and on the ideals and ethics of the physician.

By courtesy of the Soprintendenza Alle Antichita Di Ostia, Italy



Hippocrates, Roman bust copied from a Greek original, *c.* 3rd century BC. In the collection of the Antichità Di Ostia, Italy.

Multiple authorship of the Hippocratic Collection

The Hippocratic Collection. From shortly after the Hippocratic period, references were made to named works by "Hippocrates," and this tradition continued. The number of works "by Hippocrates" known in ancient times was about 70, but the number now extant is about 60. They became known as the Hippocratic Collection (*Corpus Hippocraticum*), of which the earliest surviving manuscript dates from the 10th century AD.

Even in antiquity it was realized that not all the works attributed to Hippocrates had actually been written by him—hence the later attempts to designate the "genuine works." This endeavour started at least as early as the 2nd century AD and continues to the present day. The works differ enormously in length and style, in the opinions expressed, and in the types of their intended users. Some are written for professional physicians, some for their assistants and students, some for laymen, and some are philosophical works. From internal and other evidence the approximate dates of some of the treatises are known, and it seems fairly certain that at least a century —and possibly much longer—separates the date of the earliest work from that of the latest. One feature is common: all the works were written in the Ionic dialect, which thus became the language of Greek science.

There has long been general agreement that the collection constituted the library of a medical school, probably that at Cos, and that, during the 3rd or 2nd century BC, it passed to the great library at Alexandria, where the works were edited and made available. The collection deals with the following subjects: anatomy; clinical subjects; diseases of women and children; prognosis; treatment by diet and drugs; surgery; and medical ethics.

Prominent among the works in the Hippocratic Collection were a treatise on *Epidemics*, in seven books and written by at least two authors; *On the Sacred Disease*, a treatise on epilepsy; *Prognostics; Airs, Waters and Places;* and *Aphorisms*, a collection of 412 short counsels regarding diagnosis, prognosis, and treatment (see MEDICINE, HISTORY OF: *Early Greek and Roman Medicine*).

BIBLIOGRAPHY. The only modern edition of the whole of the Greek text of the Hippocratic Collection is EMILE LITTRE, *Oeuvres complètes d'Hippocrate*, 10 vol. (1839–61, reprinted 1961). This work also gives a French translation of the complete collection, and is the only complete translation into any modern language. A selection of 28 of the treatises are given in Greek text and English translation by W.H.S. JONES and E.T. WITHINGTON in *Hippocrates*, 4 vol. ("Loeb Classical Library," 1923–31, reprinted 1957–59). The English translation of selected works by FRANCIS ADAMS, *The Genuine Works of Hippocrates*, 2 vol. (1849), is still of great value. An excellent modern translation of 13 treatises may be found in JOHN CHADWICK and W.N. MANN, *The Medical Works of Hippocrates* (1950).

An excellent discussion of Hippocrates and his influence is in CHARLES SINGER, *Greek Biology and Greek Medicine* (1922). A shorter discussion, from a slightly different aspect, is in CHARLES SINGER and E.A. UNDERWOOD, *A Short History of Medicine*, 2nd ed. (1962). The whole Hippocratic question is very fully discussed, from the medical and philological aspects, in H.E. SIGERIST, *A History of Medicine*, vol. 2 (1961). For the Hippocratic oath, see W.H.S. JONES, *The Doctor's Oath* (1924); and LUDWIG EDELSTEIN, "The Hippocratic Oath," *Bull. Hist. Med.*, suppl. no. 1 (1943), reprinted in Edelstein's *Ancient Medicine* (1967). For a modern discussion of the therapeutic armamenterium of Hippocrates, see J. STANNARD, *Bull. Hist. Med.*, 35:497–518 (1961).

(E.A.U.)

Hiroshige

A noted Japanese print artist, Andō Hiroshige was the last major figure in the development of what has become known as the *Ukiyo-e* ("Pictures of the Floating World") school. In particular, he consolidated the landscape print as an independent genre and depicted the total range of his nation's scenic beauties in a manner that could be, for the first time, readily appreciated by the common man. When the Japanese woodcut was rediscovered in the Europe of the later 19th century, it was Hiroshige who gave such Western artists as Whistler, Cézanne, Toulouse-Lautrec, Gauguin, and van Gogh a new vision of nature.

Hiroshige was born in 1797, the son of Andō Genemon, warden of the Edo (Tokyo) fire brigade, which was entrusted with the protection of Edo Castle. Hiroshige had three sisters, two of them considerably older than himself. Various episodes indicate that the young Hiroshige was fond of sketching and probably had the tutelage of a fireman, Okajima Rinsai, who had studied under a master of the traditional Kanō school of painting.

In the spring of 1809, when Hiroshige was 12, his mother died. Shortly after, his father resigned his post, passing it on to his son. Early the following year, his father died as well.

Despite his age, the young fire warden conducted himself with distinction in times of actual conflagration, but, although holding a position of some prestige, Hiroshige's actual daily duties were minimal, and his wages were small. Undoubtedly, these factors, plus his own natural bent for art, eventually led him to enter, in about 1811, the school of the *Ukiyo-e* master Utagawa Toyohiro.

Schooling under Toyohiro

Hiroshige is said to have first applied to the school of the more popular artist Toyokuni but was turned down due to the latter master's large number of applicants. Toyokuni's less renowned confrere Toyohiro, too, is said to have first refused Hiroshige, acquiescing only after the latter's persistent requests. Had Hiroshige been accepted as pupil by Toyokuni, he might well have ended his days as a second-rate imitator of that artist's specialty of gaudy prints of girls and actors. It was doubtless the more modest and refined taste of Toyohiro that helped form Hiroshige's own style—and led his genius eventually to find full expression in the new genre of the landscape print.

Although receiving a *nom d'artiste* and a school license at the early age of 15, Hiroshige was no child prodigy, and it was not until six years later, in 1818, that his first



"Kambara" by Hiroshige, woodblock colour print from the series, "Fifty-three Stages on the Tōkaidō," 1834. 25 cm × 37 cm.
By courtesy of Peter Morse

published work appeared. In the field of book illustration, it bore the signature Ichiyūsai Hiroshige. No earlier signed works are extant, but it is likely that, during this student period, Hiroshige did odd jobs (*e.g.*, inexpensive fan paintings) for the Toyohiro studio and also studied, on his own, the Chinese-influenced Kanō style and the impressionistic Shijō style—both of which were to strongly influence his later work.

As soon as he was able, Hiroshige transferred to his own son the post of fire warden and devoted himself to his art. As is customary with artists of the plebeian *Ukiyo-e* school, early biographical material regarding Hiroshige is scarce: he and his confreres were considered to be only artisans by the Japanese society of the time, and, although their works were widely enjoyed and sometimes even treasured, there was little interest in the personal details of their careers. Thus, Hiroshige's adult years must be traced largely through his works.

Stages in Hiroshige's artistic life

Hiroshige's artistic life may be characterized in several stages. The first was his student period, from about 1811 to 1830, when he largely followed the work of his elders in the field of figure prints—girls, actors, and samurai, or warriors. The second was his first landscape period, from 1830 to about 1844, when he created his own romantic ideal of landscape design and bird-and-flower prints and brought them to full fruition with his famed "Fifty-three Stages of the Tōkaidō" (a product of his sketching journey in 1832 on the famed Tōkaidō highway between Edo and Kyōto) and with such later series as "Famous Places in Japan," "Views of Kyōto," "Eight Views of Lake Biwa," "Sixty-nine Stations of the Kiso Highway," and various series of views featuring Edo and environs. His last stage was his later period of landscape and figure-with-landscape designs, from 1844 to 1858, during which overpopularity and overproduction tended to diminish the quality of his work.

Hokusai (*q.v.*), Hiroshige's early contemporary, was the innovator of the pure landscape print. Hiroshige, who followed him, was a less striking artistic personality but frequently achieved equivalent masterpieces in his own calm manner. Hokusai's approach is more powerful and realistic. Hiroshige lacked Hokusai's broad and yet intensive training in many styles but created his own genial, poetic view out of the tradition of Japanese popular painting. There was in his work a human touch that no artist of the school had heretofore achieved; these pictures revealed a beauty that seemed somehow tangible and intimate, even to men who had hitherto thought little of nature's wonders.

Unlike Hokusai, Hiroshige's life was relatively uneventful: his life was his work, with neither peaks nor valleys. Hiroshige leaves the impression of a largely self-taught artist who limited himself to the devices and capacity of his own nature.

Hiroshige was fond of travel, loved wine and good food, and in his other tastes was a true citizen of Edo. Dying in the midst of a cholera epidemic in the autumn of 1858, he enjoined his family to refrain from an excessive funeral ceremony by quoting in his final will and testament an old verse that well expresses the hedonism of old Edo:

When I die
don't cremate me don't bury me:
just throw me in the fields
and let me fill the belly of
some starving dog.

Hiroshige's own farewell verse went:

Leaving my brush behind
in Edo
I set forth on a new journey:
let me sightsee all the famous views
in Paradise!

MAJOR WORKS

PRINTS AND BOOK ILLUSTRATIONS: "Famous Views of the Eastern Capital" (*c.* 1831); "Flower and Bird" sets (*c.* 1832–34); "Fifty-three Stages on the Tōkaidō" (1833–34); "Views of Kyōto" (1834); "Eight Views of Lake Biwa" (1835); "Eight Views in the Environs of Edo" (*c.* 1835); "Tōto meisho" (*c.* 1833–34); "Edo meisho" (*c.* 1833–43); "Kōto meisho" (*c.* 1833–34); "Honchō meisho" (*c.* 1833–43); "Kiso kaidō" (*c.* 1839); "Kanazawa hakkei" (*c.* 1839); "Set of Fishes" (*c.* 1840); "Reisho Tōkaidō" (*c.* 1848–54); "Wakan rōei shū" (*c.* 1853); "Sixty-odd Provinces" (*c.* 1853); "Thirty-six Views of Mt. Fuji" (1854–58); "One Hundred Views of Edo" (1856–58); "Triptychs of Naruto Rapids, Moonlight on Kanagawa, Mountains and Streams of Kisō" (1857).

BIBLIOGRAPHY. EDWARD F. STRANGE, *The Colour-Prints of Hiroshige* (1925), the principal reference work in English, though often based upon outdated sources; YONE NOGUCHI, *Hiroshige*, 2 vol. (1934–40), an impressionistic monograph, well illustrated; RICHARD LANE, *Masters of the Japanese Print* (1962), includes a critical survey of Hiroshige's work, based on original sources; MINORU UCHIDA, *Hiroshige*, text in Japanese, with English index (1930), a comprehensive source work; JUZO SUZUKI, *Hiroshige*, text in Japanese, with English catalog of plates (1970), scholarly study of the basic Hiroshige materials, well illustrated.

(Ri.L.)

Histamine and Antihistamines

Histamine is a compound found in nearly all tissues of the mammalian body and in some plants and foodstuffs. It belongs to a group of substances called autocoids that are released from tissues during conditions of stress, inflammation and allergy. Other autocoids are serotonin, polypeptides, adenyl compounds, and prostaglandins. The term H-substance was once used to describe a material (found in extracts of mammalian skin) that was released under conditions of oxygen deficiency and injury and was thought to be responsible for local expansion of blood vessels (vasodilatation). It has now been established that H-substance is histamine.

Antihistamines are synthetically prepared compounds

that selectively counteract the pharmacological effects of histamine, following its release from certain large cells (mast cells). Antihistamines replace histamine at its active sites, thereby preventing histamine-triggered reactions.

General properties of histamine. *Chemistry.* Formed by the action of the enzyme histidine decarboxylase on the amino acid histidine, histamine is stored chiefly in tissue mast cells and blood mast cells (*i.e.*, basophils). It can also be prepared synthetically. Histamine readily combines with heparin (the anticoagulant); this complex is stable in water but not in the presence of negative ions. In most tissues, histamine is bound with heparin, in large granules of mast cells.

Physiology. The lung and gastrointestinal tract of most mammals contain relatively large amounts of histamine. There is, however, no simple relationship between the histamine content of a tissue and the ability of that tissue to form histamine, to destroy it, or to respond to it. The physiological importance of histamine remains unsolved. It is a potent activator of acid gastric secretion, and this may be its major role in mammals. It has also been identified as forming part of (1) a mechanism that affords protection, by localizing damage and assisting repair of tissues; (2) a process related to growth, increasing the availability of food and oxygen to developing tissues; and (3) a process involving pregnancy, contributing to the control of uterine movement and producing dilatation of fetal and maternal blood vessels. The presence of histamine in the hairs of stinging nettles of plants may also be protective since it is partly responsible for the swelling and itching produced by contact with *Urtica* leaves. It is also present in many insect venoms.

Possible roles in mammals

The release of histamines from lung mast cells causes local contraction of the smooth muscle cells and can result in asphyxiation. The smooth muscles of the uterus and intestine of most vertebrate species (the rat and mouse are exceptions) are very sensitive to histamine, which causes spasms of these muscles. Histamine, however, produces relaxation of vascular smooth muscle and a profound fall in blood pressure as permeability of the blood vessels increases and fluid passes from them into the tissues. The combination of vasodilatation and loss of circulating volume result in shock and collapse. These effects are often observed after injuries. Intradermal injections of histamine produce the "triple response"; *i.e.*, dilatation of skin vessels, increased permeability leading to edema, and dilatation of the surrounding arterioles. Histamine stimulates acid gastric secretion, probably by acting directly on the glandular secretory cells of the gastric mucosa; this is the only action of histamine that is not antagonized by antihistamines.

Histamine liberators

Certain chemicals are capable of releasing histamine from tissue stores, either by displacing it from the binding site or by damaging the mast cell membrane. Compounds that release histamine include snake and wasp venoms, certain enzymes, detergents, and a large group of chemicals (*e.g.*, amines, amidines, and guanidines). Histamine release may also follow the ingestion of foodstuffs (*e.g.*, fresh egg white, strawberries, and shellfish).

Following the injection of a foreign protein (termed an antigen) into an animal, compounds called antibodies are formed from blood globulins (see IMMUNITY). These circulate in the bloodstream and are absorbed by the tissues. If the same antigen is injected intravenously into such a sensitized animal three or more weeks later, the antigen combines with the antibodies, tissue mast cells are disrupted and their constituents (including histamine) are released. These events can lead to a condition termed anaphylactic shock. The effects of the second antigen injection are similar in many respects to those produced by an injection of histamine. An anaphylactic shock response may occur in a sensitized man after repeated ingestion of certain foods or treatment with drugs such as penicillin. Sensitization for certain antigens as manifested by allergy and asthma is related to the phenomenon of anaphylactic shock (see ALLERGY AND ANAPHYLACTIC SHOCK).

In the body histamine is changed to 1,4-methylhistamine, which is either excreted directly or deaminated (nitrogen is removed) by action of the enzyme histaminase before excretion. The product imidazolylacetic acid is excreted in the urine, usually combined with the sugar ribose.

General properties of antihistamines. Antihistamines oppose selectively all the pharmacological effects of histamine except those on gastric secretion. Development of antihistamines dates from about 1937, when French workers discovered compounds that protected animals against both the lethal effects of histamine and those of anaphylactic shock. The first antihistamines were derivatives of ethylamine; aniline-type compounds, tested later and found to be more potent, were too toxic for clinical use. However, in 1942, the forerunner of most modern antihistamines (an aniline derivative called Antergan) was discovered; then more potent, more specific, and less toxic compounds were prepared. More than 100 antihistaminic compounds soon became available for treating patients.

Since histamine is involved in the production of some symptoms of allergy and anaphylaxis, antihistamines can control certain allergic conditions, among them being hay fever and vasomotor rhinitis; the nasal irritation and watery discharge are most readily relieved. Persons with urticaria, edema, itching, and certain sensitivity reactions respond well. Antihistamines are not usually beneficial in treating the common cold and asthma. Antihistamines with powerful antiemetic properties are used in the treatment of motion sickness and vomiting. Used in sufficiently large doses, nearly all antihistamines produce undesirable side effects; the incidence and severity of the side effects depend both on the patient and on the properties of the specific drug. The most common side effect in adults is drowsiness; cerebral stimulation may occur in infants. Other side effects include gastrointestinal irritation, headache, blurred vision, and dryness of the mouth. If a patient does not improve after three days of treatment with antihistamines, it is unlikely that he will benefit from them. Antihistamines are readily absorbed from the alimentary tract, and most are inactivated by monoamine oxidase enzymes in the liver.

BIBLIOGRAPHY. M. ROCHA E SILVA (ed.), *Histamine and Anti-Histaminics* (1966), an encyclopaedia on histamine and antihistamines, including chemistry, metabolism, physiological and pharmacological actions; G. KAHLSON, "A Place for Histamine in Normal Physiology," *Lancet*, 1:67–71 (1960), a review of the evidence that histamine may be concerned in metabolic machinery common to the synthetic processes involved in development, growth, and repair; W.D.M. PATON, "Histamine Release by Compounds of Simple Chemical Structure," *Pharmacl. Rev.*, 9:269–328 (1957), a review of compounds that release histamine from animal tissues, and the mechanisms involved; J.F. RILEY and G.B. WEST, "The Presence of Histamine in Tissue Mast Cells," *J. Physiol., Lond.*, 120:528–537 (1953), evidence that histamine content of tissues parallels mast cell population; B. UVNAS, "Mechanism of Histamine Release in Mast Cells," *Ann. N.Y. Acad. Sci.*, 103:278–284 (1963), study of biochemical mechanisms of histamine release.

(G.B.W.)

Historiography and Historical Methodology

Modern historians aim mainly at reconstructing an accurate record of human activities and at achieving a more profound understanding of them. This conception of their task is quite recent, dating only from the development in the late 18th and early 19th centuries of scientific history, cultivated largely by professional historians. It springs from an outlook that is very new in human experience: the assumption that the study of history is a natural, inevitable kind of human activity. Before the late 18th century, historiography (the writing of history) did not stand at the centre of any civilization. History was almost never an important part of regular education, and it never claimed to provide an interpretation of human life as a whole. This was more appropriately the function of religion, of philosophy, even perhaps of poetry and other forms of imaginative literature.

This article is divided into the following sections:

I. Historiography in the West
Ancient historiography
Historiography in the Middle Ages
Byzantine historiography
Historiography in the Renaissance
Early modern historiography
Historiography in the age of the Enlightenment
Historiography in the 19th and 20th centuries
II. Non-Western historiographical traditions
Muslim historiography
East Asian historiography

I. Historiography in the West

ANCIENT HISTORIOGRAPHY

Classical historiography. The older, pre-18th-century outlook has been particularly well studied in the historiography of the ancient Greeks and Romans. But, although two of the most important ancient historians, Herodotus and Thucydides, wrote as early as the 5th century BC, when recorded Greek historiography was only just beginning, they had few successors of comparable quality. It is a symptom of the relative lack of importance attached in antiquity to this type of activity.

Ancient history was a branch of literature. The most appreciated historians were the writers who, like Thucydides, were able to touch on universal human problems or who, like the Roman author Tacitus (died *c.* AD 120), wrote in a dramatic way about important events or who, at least, attracted readers by their excellent style and skill in composition. Many of the works that lacked some of these literary qualities failed to survive.

About 1,000 ancient Greeks wrote in antiquity on historical subjects, but most of these writers are mere names. Many of the losses appear to have occurred in antiquity itself. Even historians of first rank have fared badly. Only in a few cases have complete texts of all their writings survived. Of the voluminous history of Polybius (covering originally the period 220–144 BC) only about one-third survives. Nearly half of Livy's Roman history (originally covering the period 753–9 BC) is lost. The text that remains is reasonably good only through the efforts of a group of Roman aristocrats who, in about AD 500, were trying to salvage the chief glories of Roman literature. A considerable part of Tacitus is missing, and the surviving portions of his *Annals* and *Histories* (originally AD 14–96) derive from two unique manuscripts.

Herodotus, whom the Roman statesman Cicero called "the father of history," came from the western coast of Asia Minor. The writers who preceded him were mainly Ionians from the Greek settlements in the same area. The origin of Greek historiography lies in the Ionian thought of the 6th century. The Ionian philosophers were doing something unprecedented: they were assuming that the universe is an intelligible whole and that through rational inquiries men might discover the general principles that govern it. Hecateus of Miletus, the most important Ionian predecessor of Herodotus, was applying the same critical spirit to the largely mythical Greek traditions when he wrote, early in the 5th century, "the stories of the Greeks are numerous and in my opinion ridiculous." Herodotus was more of a traditionalist, but he introduced his work as an "inquiry" (*historia*).

Egyptian and Babylonian historiography

A glance at the older historiography of the Egyptians, the Babylonians, and the other peoples of the ancient Near East will heighten one's appreciation of the novelty of the task undertaken by Herodotus. The kings of Egypt, of Babylonia and Assyria, and of the Hittites and the Persians all sought to preserve their glorious deeds for posterity in monumental inscriptions. The more important rulers also accumulated large archives, including both ordinary administrative documents and records specially commemorating their achievements. Some 20,000 clay tablets remain from the collections written for Ashurbanipal of Assyria (668–627 BC). Both in Egypt and in Babylonia lists of kings were kept in the temples, and these were sometimes supplemented by brief annals recording the principal events, though the hatred felt by certain rulers for their predecessors led to periodic destructions of older material. The exceptional meagreness of the narrative sources for Babylonian history before 747 BC seems due to the obliteration of the older annals by Nabonassar of Babylonia (ruled 747–734). Apart from changes in literary style, there was surprisingly little development over a period of more than 1,000 years in all these types of commemorative records. The inscriptions and temple records were normally intended to perpetuate the glory of the gods in whose service these rulers had accomplished great deeds. The names and dates of dynasties and of particular rulers can be reconstructed fairly adequately with the aid of these sources, but it would be absurd to expect much accurate or even true information about particular events. Nor, with very rare exceptions, were the men who had access to this material interested in using it to write continuous histories.

Herodotus and his immediate Ionian predecessors shared a very novel outlook. Its distinctive features were a lively curiosity and a capacity to treat sources in a critical spirit. Boundless curiosity about people and their diverse customs is one of the most endearing traits of Herodotus. Like other Greeks from western Asia Minor, he was particularly stimulated by contacts with the great Persian Empire, which offered opportunities for reasonably secure travel. The resultant immense widening of historical perspective is illustrated by a story told by Herodotus about Hecateus. When the latter assured the Egyptian priests at Thebes that he could trace his descent through 16 generations, the Egyptians showed him evidence of the descent of their high priests through 345 generations. Herodotus was the first to link his geographic inquiries with true history. His descriptions of the barbarian world that confronted the Greeks provided an introduction to the epic of the successful Greek resistance to the Persians.

The types of history written by the ancient Greeks and Romans influenced profoundly all subsequent historiography down to the 18th century. In order to interpret sympathetically this classical historiography, it is necessary to bear in mind the literary conventions that governed this branch of literature. The ancient Greeks distinguished between history and biography. The origin of both forms can be traced back to at least the 5th century BC, and the differences between them were observed throughout antiquity. The writer of history was supposed to aim at giving a true story, but the purpose of biography was to praise and to edify. The biographer was entitled to treat historical personages in a manner that resembled legend rather than sober history. There existed, of course, some exceptions. The lives of the early Roman emperors written by Suetonius in the 2nd century AD, while conforming to the traditional, topical arrangement of biographies, constitute an unusually valuable historical source, especially for Augustus, whose correspondence is repeatedly quoted. Yet another distinction was drawn between history and the study of "antiquities," to use a term employed by Varro (116–27 BC), perhaps the greatest of all the ancient Roman scholars. This distinction was already implicit in Aristotle's contemptuous dismissal of history (in his *Poetics*) as a branch of literature dealing with the particular rather than with things of general significance. The histories he condemned provided chronological narratives of wars and political events. Aristotle and his disciples were engaged in several enterprises that they regarded as something quite different from history. For example, they embarked on the study of the constitutions of all the Greek states. Such work was to be based on systematic inquiries. The student of the "antiquities" tried to use a wider range of evidence than the sources normally consulted by the ancient historians, and he arranged his results systematically by topics.

Ancient biography

In antiquity a writer of history was usually preoccupied at least as much with style as with content. A generation before Aristotle, the rules of rhetoric, as they might be applied to history, were fully elaborated by Isocrates, a teacher of rhetoric at Athens. Cicero tried (especially in his *De oratore*, 55 BC) to familiarize the Romans with

these Isocratean precepts. History was to be written in a clear but solemn style, akin to fine oratory. The historian was to introduce all manner of literary embellishments but was also to stress the moral lessons of his story. At its worst this type of historiography could lead to serious misrepresentations of the past. Among the Roman historians, Livy (died AD 17) was an important practitioner of this kind of writing, which was particularly well suited to the patriotic myths that he was trying to immortalize, of a Rome that owed its magnificent destiny to the unique virtues of its citizens and the perfection of its antique institutions. Some outstanding historians, such as Polybius (2nd century BC) and Caesar (died 44 BC), eschewed these rhetorical precepts, but in all the ancient writers an important element of literary artifice was always present. This is one of the reasons why they offend modern standards, which demand absolute accuracy in the presentation of evidence. One of the most striking contrasts is the reluctance of the ancient historians to quote documents. Tacitus might rely heavily on the archives of the Roman Senate, but he never mentions his documentary sources. An inscription discovered at Lyons, France, preserves a speech delivered by the emperor Claudius to the Senate in AD 48, and it is clear that Tacitus utilized another version of the same text. His skill in using it is matched by the freedom with which he adapts it to suit his purpose.

The greatest and the most original achievement of the best Greek historians lay in their clear grasp of the need to distinguish truth from fiction and their conscious preoccupation with the methods of achieving this. This is admirably conveyed in a famous passage of Thucydides.

Methods of Thucydides

> And with reference to the narrative of events, far from permitting myself to derive it from the first source that came to hand, I did not even trust my own impressions, but it rests partly on what I saw myself, partly on what others saw for me, the accuracy of the report being always tried by the most severe and detailed tests possible. My conclusions have cost me some labour from the want of coincidence between accounts of the same occurrences by different eye-witnesses, arising sometimes from deficient memory, sometimes from deficient impartiality.

His practice did not fully live up to this ideal, however. The greatest of his Greek successors, Polybius, is reasonably impartial, except in his treatment of some of the events in Greece. Among the Romans, the writing of history was chiefly the preserve of members of the senatorial class, who almost invariably had some personal axes to grind. But the correctness of the rules formulated by Thucydides was accepted, in principle, by most ancient historians.

Thucydides had deliberately restricted himself to the history of his own time, and many of the subsequent ancient historians did likewise. They could depend on their own experience or could question well-informed contemporaries. The surviving fragments of Livy relating to his own lifetime (64/59 BC–AD 17) are much more vivid and convincing than the earlier books of his history (surviving today only down to 167 BC). The tendency to prefer contemporary history was strengthened by the practical bent of many of these writers. Several ancient historians were men of action familiar with warfare and politics. Interested in history as a source of instruction for statesmen, they could write with authority only about wars and political transactions of their own time. Polybius, the exiled Achaean general and a great traveller, derides unpractical, sedentary historians such as Timaeus, who had been writing about the peoples of the western Mediterranean without stirring for 50 years from Athens.

The historians of antiquity were much less skillful in dealing with noncontemporary history, for which they relied on older historians. Where none was to be found, they felt lost, as Livy complains in the early portions of his Roman history. The modern recourse to non-narrative sources was alien to the habits of most ancient historians. They were usually incapable of doing this successfully, just as they were ill equipped to discuss critically the sources used by the older writers.

Herodotus chose for his theme the successful resistance of the Greeks against the Persians at the beginning of the 5th century BC. Thucydides wrote about the Peloponnesian War, in which virtually all the Greek states became involved in the last decades of that century. These were limited subjects of obvious importance for which it was possible to find ample evidence. The strength of the ancient historians lay precisely in imposing an interesting pattern on the events of a selected period, usually contemporary or fairly recent, for which they had manageable sources. The best of them could thereby achieve a sense of dramatic unity and produce literary masterpieces. The speeches that Thucydides invented for some of the main protagonists in his story are artistically the most satisfying parts of his work, and at times they even seem to recapture the spirit of what might have been said on these occasions. In a superb writer like Tacitus, whose political career had included long periods of frustration and insecurity, one does not look for impartiality or for scrupulous truthfulness but, rather, for fascinating insights into what the development of Roman imperial power from Augustus to Domitian (the period AD 14–96) meant to the proud, sophisticated Roman aristocracy for whom he was writing.

Classical study of "antiquities"

The study of "antiquities," as opposed to narrative history, did not normally produce works of literary merit, and this is probably the main reason why most of them disappeared. One important group of such writings originated with Aristotle and his collaborators, writing in the third quarter of the 4th century BC. They were interested in both literary "antiquities" and in the systematic study of the constitutions of Greek states. They had described 158 different constitutions, though only their account of Athens now survives. A comparison of its two main parts illustrates the contrast between the deficiencies of ancient historiography and the impressive achievements of the antiquarian researchers. In the introductory, historical section, Aristotle was baffled by the problem of dealing with the fairly remote past. For each particular period he tried to follow some contemporary sources. The resultant juxtaposition of several writers differing widely in their political outlook produced an account full of contradictions. The second part, however, containing a systematic description of the Athenian constitution, is a masterpiece of shrewd analysis, as are the empirical portions of Aristotle's *Politics* (Books IV–VI), which are based on a wealth of concrete examples derived from the different Greek states.

Aristotle inspired in the 3rd and 2nd centuries BC a great mass of philological and antiquarian research. The most important scholars were to be found in the new Hellenistic states, especially at Alexandria in Egypt and at Pergamum in Asia Minor. Among the surviving Hellenistic fragments, there are commentaries on Herodotus and Thucydides. The Hellenistic scholars were interested in many subjects connected with history and did pioneering work in chronology, geography, and topography. They were accustomed to using every kind of source and to quoting documents extensively. Their greatest Roman disciple was Varro, who tried to recover all the vestiges of the old Roman society and to make a systematic survey of Roman life based on the evidence provided by language, literature, religion, and ancient customs. Most of his writings have been lost, but he supplied the conjectural (though incorrect) date of 753 BC for the foundation of Rome and knowledge of the probable boundaries between some of the groups whose union produced the city of Rome. Unfortunately, antiquarian researches of such penetrating nature were almost never applied in antiquity to the writing of narrative histories.

Early Christian historiography. The triumph of Christianity in the Roman Empire during the 4th century assured the predominance of a type of historiography radically different from the works of the pagan Greek and Roman historians. Its origins were Jewish. The Jews were the only people of antiquity who had the supreme religious duty of remembering the past because their traditional histories commemorated the working out of God's plan for his chosen people. By contrast, no Greek ever heard his gods ordering him to remember. It was the

Purpose of Jewish histories

duty of every Jew to be familiar with the Jewish sacred writings, which were ultimately gathered into what became the Old Testament. The writers of these biblical books only gave an authoritative version of what everybody was supposed to know, and they were only concerned with the selection of such facts as seemed relevant in interpreting God's purpose. In addition, the Jews also cherished unwritten traditions. To quote Josephus, a Jewish historian of the 1st century AD, "what had not been written down, was yet entrusted to the collective memory of the people of Israel and especially of its priests."

The Christians took over the Old Testament and added to it an additional body of sacred history. The writers of the four Gospels included in the New Testament were bearing witness to assured truths that the faithful ought to know, and no convincing reconstruction of historical facts is possible from these books of the New Testament. The only avowedly historical book in it is the Acts of the Apostles. The New Testament as a whole represents merely a selection from the early Christian writings. It includes only what conformed to the doctrine of the church when, later on, that doctrine became fixed in one form. Between the Acts of the Apostles, dating probably from the late 1st century, and the writings of Eusebius of Caesarea (died *c.* 340) and his contemporaries in the first quarter of the 4th century, there is an almost complete gap in Christian historiography.

For the Christian writers the story of Jesus, as recorded in the Gospels, represented the fulfillment of the prophecies that could be found in various parts of the Old Testament. The Jewish part of the Bible also assured for Christianity the authority of a long antiquity. The history contained in the two parts of the Bible, now indissolubly linked together, became the only authentic record of God's revelation for mankind, dwarfing into insignificance all the records of other peoples and religious groups. The concept of a universal history had not been wholly unknown to the pagan world, but the Christians were the first to apply it effectively. Christian history had to be a universal history, though of a very peculiar sort, where only one sequence of privileged events, Jewish and Christian, deserved detailed record. The Christian claims must have seemed more extravagant to the pagans than even the Jewish ones. Thus Eusebius stated that the Christians were, in fact, born with the world, anticipating St. Augustine's vision of the city of God existing since the beginning of time.

Universal chronologies

In defending their religion against hostile critics, the early Christians were forced to fit some pagan history into their universal scheme. This was achieved by means of universal chronologies from the creation of the world to each writer's own time. The events of Jewish and Christian history were thus synchronized with the main dates of the pagan myth and history. Sextus Julius Africanus, who wrote in the early 3rd century, is the first Christian writer known to have attempted this feat. He allotted 6,000 years to the whole span of human history and placed the birth of Christ in the year 5500 from the creation of the world. This work provided the model for the more elaborate *Chronographia* (*Chronicle*) of Eusebius. It became the foundation for a long succession of Greek chronographies produced by Byzantine writers. A Latin adaptation by St. Jerome (died 419/420) was immensely influential in western Europe for more than 1,000 years. A modern scholar is filled with mingled admiration and despair at the ingenuity of Eusebius and of his more eminent successors and at the absurdity of many of their conclusions. But they did originate and impose on the world a unified scheme of universal chronology. The dating from the birth of Christ was introduced by Dionysius Exiguus, who wrote at Rome in the early 6th century, and it was successfully popularized in the 8th century by the English historian Bede.

The writing of history of their own time was not an essential task for the Christians of the 4th and 5th centuries. When they did so, they wrote primarily in defense of their religion against the pagan world or against rival Christian groups branded as heretical. All these histories belong to religious apologetics. They suffer from inevitable distortions in the choice of what should be mentioned and what must be suppressed, and they are often excessively unfair to outsiders and opponents. These faults were not uncommon among the classical historians, though the Christians were somewhat unusual in their extreme conviction that they alone must be right. A comparison between the Christian historians and an outstanding pagan writer, such as Ammianus Marcellinus (second half of the 4th century), who was very ready to admire those Christians who merited it, brings out the intolerance and narrowness of outlook of his Christian contemporaries.

Eusebius was the earliest and the most important of the Christian historians of the 4th century. He is quite frank about the practical and apologetic aims of his *Historia ecclesiastica* (written 312–324; *Ecclesiastical History*) designed to show how, through a long series of acts of Divine Providence, a Christian empire was finally brought into existence by Constantine. He admits that "we shall introduce into this history in general only those events which may be useful first to ourselves and afterward, to posterity." This work, like his other historical writings, is a mixture of devout fiction and invaluable detail. But there is plenty of the latter in *Ecclesiastical History*. Contrary to the usual practice of the ancient historians, Eusebius tries to specify his sources, and he quotes from them extensively in order to document as fully as possible the developments that resulted in the triumph of Christianity. He provided in this respect a valuable model for his medieval successors. The most astonishing thing about Eusebius was his capacity to handle his sources critically, in matters where it seemed permissible to do so. In one passage of his *Chronicle* he sets aside the authority of St. Paul in favour of a piece of evidence contained in the Book of Judges. In later patristic literature nothing similar is found.

Biography, as it was habitually written in antiquity, could be readily adapted to Christian purposes. St. Jerome modelled himself on Suetonius in compiling the lives of 135 Christian writers (written in 392) as a way of demonstrating the high level of culture attained by his co-religionists. The ancient biographers had freely mingled fact with fiction for the edification of their readers and could be readily imitated by the writers of the lives of Christian saints. The life of St. Anthony of Egypt by St. Athanasius (mid-4th century) set the pattern for this most popular type of medieval literature.

St. Augustine

St. Augustine, the greatest of the Latin Church Fathers of the 4th and 5th centuries, was certainly not concerned with writing of history in any ordinary sense of the term. In his *De civitate Dei* (*City of God*) he might invoke historical evidence to demonstrate the utter degradation of all the non-Christian societies, and he encouraged his pupil Orosius to develop this theme more fully in the latter's *Historiarum libri VII adversus paganos* (*Seven Books of History Against the Pagans,* to 417). Nearly 200 manuscripts of Orosius have survived, testifying to the immense popularity of his work in the Middle Ages. Augustine's greatest influence on historiography lay in his main message. His vision of the divine and the earthly cities confronting each other dominated the outlook of all the medieval Christian thinkers and profoundly affected their treatment of history. Within that divine plan for the world, purely secular history seemed an insignificant thing.

HISTORIOGRAPHY IN THE MIDDLE AGES

The 5th to the 11th century. The period stretching from the 5th to the 11th century was a time of very profound cultural decline in regions that had once constituted the western half of the Roman Empire. Almost all the inhabitants of these provinces again became illiterate. There are long periods for which there are virtually no narrative sources, and the bulk of surviving historical writings consists merely of meagre factual annals. Virtually all the writers were ecclesiastics, in marked contrast to the Byzantine lands, where a strong tradition of lay historiography persisted throughout the Middle Ages.

The annalists and chroniclers of the West were predominantly monks, and their lack of experience of the secular world outside their cloisters made them into blinkered and unpractical historians. This was true even of Bede, an Anglo-Saxon monk, who was by far the greatest historian of the early Middle Ages.

All the historians of this period were seriously affected by the cultural decline around them. They were having to write in part for a more uncultured audience. Sulpicius Severus, probably the best Western historian of the early 5th century, still intended his *Chronica* (to 403) for educated Roman Christians, but his life of St. Martin of Tours is a piece of medieval hagiography. This model could inspire lives full of folklore and miracle, from which the real human personalities of the saints were almost wholly absent. The same duality of purpose is a notable feature of Bede's voluminous writings. He explicitly recognized that he must adapt himself to his audience when he explained that he was writing in a simple Latin style so that he might be more easily understood by his Anglo-Saxon readers. There is a marked contrast of tone between his theological and his historical writings. As a theologian, Bede follows Eusebius and the earlier Church Fathers in not exaggerating the frequency of miracles and in believing that they were most common in the earliest days of Christianity. But Bede's lives of the English saints and his *Historia ecclesiastica gentis Anglorum* (*Ecclesiastical History of the English People*), covering chiefly the years 597–731, are full of miracles and visions. There is one or other on almost every page. It is possible that some of these incidents were included by Bede because he thought that his readers expected mentions of these familiar, traditional stories.

Bede's *Ecclesiastical History*

In preparing his historical works, Bede not only took great care to assemble the widest possible collection of sources but also tells the reader what he is using. In dedicating his *Ecclesiastical History* to King Ceolwulf of Northumbria, he requests that

> in order to remove all occasions of doubt about those things I have written, either in your mind or in the minds of any others who listen to or read this history, I will make it my business to state briefly from what sources I have gained my information.

An impressive list follows, including mentions of documents copied for him by friends at Rome, Canterbury, and other places. Like Eusebius, on whom Bede modelled himself, he quotes some of the documents integrally. Bede's methods of securing and recording information are so similar to the practices of modern historians and the judicious tone of his writing is so impressive that the reader is almost taken in into treating him as if he were a modern scholar. But Bede's *Ecclesiastical History* was written as a work of edification in order to strengthen the faith of his readers in Divine Providence, through which, as he saw it, his Anglo-Saxon countrymen had been converted to Christianity. All matters not connected with his main theme are ignored. Bede's handling of evidence on subjects that he regarded as embarrassing inspires mistrust. But these are small matters in comparison with the enormous mass of information that he alone has preserved and the encouragement that Bede continued to give for many centuries to the writing of history.

The influence of Bede and other Anglo-Saxon scholars was greatly felt during the later 8th and the 9th centuries in the Frankish kingdom, where under Charlemagne and his successor, Louis the Pious, there was a modest revival of historical writing. Besides the annals kept at various monasteries, which tended to convey information in a manner that suited the Frankish rulers, there were a few more ambitious ventures. The important *Historia Langobardorum* (*History of the Lombards*), written *c.* 774–785 by Paulus Diaconus, or Paul the Deacon, was the work of one of the best educated men of the time. Nithard, a grandson of Charlemagne, left an invaluable narrative of the disintegration of the Carolingian state during his lifetime. The work that exerted the greatest influence on the medieval writers of biographies was Einhard's *Vita Karoli imperatoris* (written 817–830; *Life of Charlemagne*). The author was a leading official and a close companion of Charles, and his work was naturally intended as a eulogy of the great king. Einhard says that Charlemagne retreated safely from Spain, returning with his army safe and sound, except that on a ridge of the Pyrenees, on the way home, he happened to experience some small effects of Gascon perfidy. Nobody would gather from this that the Franks had narrowly escaped a major disaster. Einhard was merely echoing the story told in the semi-official contemporary annals. Another source of distortion was Einhard's use of a classical model, the *Lives of the Caesars* by Suetonius. The subject headings under which he described Charles and even the very words used were partly borrowed from the lives of Roman emperors, but his Charlemagne is probably in essentials an authentic and credible portrait.

If bulk alone is to be taken as a criterion, annals were the main product of medieval historiography. The annalist merely sets down the most important events of the current year. In the case of the earliest medieval annals, the events were often noted down in Easter tables, in the blank spaces between the dates calculated for the forthcoming Easters. Such paschal annals would be extremely brief. When, as often happened, annals came to be written down in separate manuscripts, distinct from the Easter tables, there was room for the expansion of individual entries. In either case, the resultant annals cannot be regarded as history since the events are necessarily recorded in isolation. But they preserve in a right order the essential facts, which could be rearranged into a continuous narrative. Such a narrative, if it still followed the chronological arrangement of its various annalistic sources, should properly be termed a chronicle.

Medieval historians show little awareness of the process of historical change. They were unable to imagine that any earlier age was substantially different from their own. The unawareness of the meaning of anachronism helps to explain the strange wanderings of medieval annals and chronicles. If a religious community wanted to acquire a historical narrative, it copied some work that happened to be most readily accessible. A continuation might then be added at the manuscript's new abode, and, later on, this composite version might be copied and further altered by a succession of other writers. Hence there are at least six main versions of the annals known as the *Anglo-Saxon Chronicle*. They all derive from the annals kept down to 892 at Winchester, the West Saxon capital. Thereafter, copies were acquired by religious centres in the most diverse parts of England, and one manuscript was being kept up to date at the abbey of Peterborough as late as 1154. An extreme case of wanderings is represented by the annals of the cathedral church of Cracow, the medieval Polish capital. The first section is based on Orosius, the next comprises annals beginning with the death of Bede and containing notices of Frankish and German events, while the Polish section starts with the conversion of Poland to Christianity (965–966) and ends in the 13th century.

The *Anglo-Saxon Chronicle*

The 12th, 13th, and 14th centuries. Historians are accustomed to regarding the late 11th and 12th centuries as an age of intensified progress in culture and learning; this development, however, did not greatly affect historiography. There was a modest revival of interest in some of the ancient Latin writers, but would-be historians were unsure which ancient models they ought to imitate. A whole series of attempts was made to apply to other races the theme in Virgil's *Aeneid* of a noble group of people guided by the gods toward a splendid destiny. The first essential step was to establish the descent of one's nation from the ancient Trojans and then to trace subsequent history through a series of heroic conquests. The most ambitious of these writings was the *Historia regum Britanniae* (*History of the Kings of Britain*), by Geoffrey of Monmouth (died 1155), which attempted to establish for the Celts a historical destiny greater than any other. Although some, even contemporary, readers were not deceived by the work, and William of Newburgh, one of the best English historians of the 12th century, denounced it as a tissue of absurdities, many seriously accepted it as history.

With a few exceptions, the ablest minds of the 12th century were attracted into enterprises that ignored history; they were more concerned with systematization of thought and with philosophical speculations. One of the exceptions was Otto, bishop of Freising, in Bavaria. He was a grandson of the Holy Roman emperor Henry IV. He received the best education that his age could give, but he was also briefly a Cistercian monk during the most austere period of that order's history. Otto was torn between conflicting impulses to seek the city of God as the only reality and yet to hope for the progress of the German empire. Out of this conflict came his first work, *Chronica* (*The Two Cities*), a chronicle of world history to 1146, perhaps the most profound medieval attempt at a Christian philosophy of history. As Otto himself confessed, it was composed "in bitterness of spirit . . . in the manner of tragedy." The election in 1152 of his nephew and friend Frederick Barbarossa, as emperor, filled Otto with a new elation. The excellence of his second work, *Gesta Friderici I imperatoris* (*The Deeds of Frederick Barbarossa*), derives in a considerable measure from a quality rare in medieval historians, a sense of optimistic belief in the value of writing history because it might become a record of human progress. *The Deeds of Frederick Barbarossa* contains a penetrating analysis of the problems encountered by the German rulers in trying to rule the precociously urbanized Italian society.

The works of Otto of Freising

As in antiquity, the best medieval works were accounts of contemporary history by men who had participated in the events that they were describing. It is, however, very significant that some of the writers that are prized most highly today survive in only very few manuscripts and were presumably not appreciated by most of their contemporaries. One such work was the *Historia pontificalis* ("Pontifical History") covering the period 1148–52, of John of Salisbury, one of the most accomplished scholars of his age, who was writing about the period when he was in the papal service. Another instance of undeserved neglect is furnished by the *Liber de regno Siciliae* ("Book of the Kingdom of Sicily") covering the period 1154–69, written by an anonymous member of the Sicilian court.

Unlike the ancient historians, the medieval writers of contemporary history had no inhibitions about extensively quoting official documents. In England, a succession of writers preserved a large quantity of such texts. Roger of Hoveden was, in the last quarter of the 12th century, treated by the English kings as a kind of court historian. He preserved valuable legal and administrative records with which he was familiar through his activities as a royal official and justice. Matthew Paris, the most important English monastic historian of the 13th century, was highly regarded by King Henry III and had excellent sources of information. He left behind a collection of transcripts of royal and ecclesiastical documents that today fills a large printed volume. Some writers made their chronicles into an anthology of official records, thinly connected by the author's brief comments. Such is the chronicle of Robert of Avesbury, consisting mainly of the military dispatches of King Edward III and other interesting documents to 1356. Another variant of the same method was for a wholly mediocre chronicle to incorporate exciting pieces of eyewitness narratives by other writers. A dull English monastic product of the late 14th century, the *Anonimalle Chronicle*, includes a narrative of the Peasants' Revolt of 1381, which is one of the most dramatic and interesting eyewitness accounts to be found in medieval historiography.

The most popular histories of the 13th and 14th centuries were encyclopaedic compilations giving all the important facts neatly arranged under the dates of popes, emperors, and other rulers. There were even more ambitious ventures aiming at summarizing all the important facts from all the different branches of human activity. The Dominican Order, created at the beginning of the 13th century, was especially concerned with producing such aids for the dissemination of useful knowledge. The best known of these Dominican works is the immense *Speculum historiale* ("Mirror of History"), by Vincent of Beauvais, written under the patronage of King Louis IX of France. It is a compilation made up of excerpts from many authors.

The 13th and 14th centuries were not a period of any fundamental innovations in the techniques and nature of historiography, but there was a growing diversity of types of historical writing. Very detailed, chatty narratives multiplied, often badly organized and inaccurate, but conveying the authentic atmosphere of the times and vividly portraying leading personalities. Such were the St. Albans chronicles of Matthew Paris (to 1259), the reminiscences of Joinville about St. Louis during the Seventh Crusade (1248–54), the Lombard chronicle of Fra Salimbene (to 1287), or the vast history of the first part of the Hundred Years' War written in the second half of the 14th century by Froissart. Memoirs and histories written in vernacular languages, such as those of Joinville and Froissart, came to be quite common. Laymen began to write histories. Some were great men, like Geoffroi de Villehardouin, one of the leaders of the Fourth Crusade (which captured Constantinople 1202–04), of which he wrote an account. Important urban chronicles began to appear, such as the Florentine chronicle of Giovanni Villani, with its invaluable statistics of Florentine population and activities around 1338. The extraordinary personality of St. Francis, who died in 1226, inspired lives of him more convincingly human than any previous medieval biographies of saints.

Vernacular histories

The Humanist historians of the 15th century tried to make a deliberate break with the tradition of medieval historiography. By their insistence on a more coherent arrangement of subject matter, by their superior critical outlook, and, above all, by their much more accurate awareness of the process of historical change, they had introduced innovations of fundamental importance. In part they owed their grasp of these new possibilities to the influence of Byzantine scholars. In historiography, as in other matters, the new humanistic scholarship was a joint product of Western and Byzantine traditions.

BYZANTINE HISTORIOGRAPHY

During the millennium that elapsed between the collapse of the Roman Empire in the West in the 5th century AD and the Italian Renaissance of the 15th century, in no part of Europe did the writers of history consistently maintain as high a standard of achievement as in the Byzantine Empire. Parts of the 7th and the 8th centuries form lengthy gaps in the record of Byzantine historiography, but this seems mainly to be the result of subsequent losses of manuscripts. When, in the middle of the 9th century, Photius, future patriarch of Constantinople, compiled a record of some 280 books that he had read, he mentioned works of 33 Greek historians, dating mostly from the late Roman Empire and the Byzantine period, 20 of which are now lost. But, among the Byzantines of the 7th and 8th centuries, there was certainly no parallel to the Dark Ages in western Europe.

Sources of Byzantine historiography

The Byzantine historians were heirs to the combined traditions of classical Greek writing, of the subsequent Hellenistic historiography, and of the Christian historical writing of the 4th century. Few ancient Latin historians were ever translated into Greek, and their influence on the Byzantines was, therefore, very slight. The older classical Greek historians provided the Byzantines with their cherished models of language and style. Like all educated Byzantines, the historians continued for a millennium to write in a literary language that soon became unintelligible to the vast majority of their compatriots. Hence, from the 6th century onward, there appeared, side by side with the learned historiography, a succession of popular chronicles written in the ordinary language. Most of these popular writings form—in their prejudice, ignorance, and crudity—a startling contrast to the works of the more eminent classicizing historians, but they do provide valuable glimpses of the sort of hagiographical history, more religious myth than sober fact, that ordinary Byzantines apparently wanted to read.

Herodotus and Thucydides were frequently invoked by

Byzantine historians as models of fine prose. The influence of these two writers on the substance of what was written usually remained slight and superficial, however. The only Byzantine writers who seriously modelled themselves on these two oldest Greek historians wrote during the 15th century. The earlier Byzantine historians owed most to Polybius and to the Greek biographer Plutarch (died *c.* AD 119), the two Hellenistic writers who had the greatest influence on Byzantine notions of how history and historical biography should be written.

Like Polybius, the majority of Byzantine historians, including most of the best ones, preferred to write about their own times; and within these limits they produced some real masterpieces. Unlike the majority of the ancient historians, Polybius had included much autobiographical detail, and his influence reinforced the readiness of the Byzantine historians to talk about themselves, thus providing abundant information about several of these authors. Their histories are likely to be one-sided and full of details about what interested them, while remaining silent about a great mass of other contemporary happenings. They are frequently gossipy and patently prejudiced, inspiring much less confidence than the austere, impartial writings of authors such as Thucydides. This is one of the main reasons why the Byzantine historians have often been excessively underestimated by modern readers. The bulk of the Byzantine contemporary histories were written by statesmen, high officials, and prelates—men with access to important information. They have to be used critically and cautiously but can be immensely valuable.

Priscus of Panium (*c.* 450), a member of a Byzantine embassy to Attila's camp, is the best source of information about that terrible king of the Huns and his followers. A century later, the reconquest of Vandal Africa and of Ostrogothic Italy by the emperor Justinian was the main theme of the *History of the Wars* of Procopius, a leading civilian adviser of Belisarius, the Byzantine commander. Subsequently, Procopius also wrote a *Historia arcana* (*Secret History*), containing a horrible indictment of the activities of Justinian and Belisarius. Many of his details about the corruption at court and the oppressive nature of the government may be substantially correct. In the 11th century Michael Psellus, who wrote a history of his own times, was a leading Byzantine scholar and official, for a time even the chief adviser of emperors. His *Chronographia* is concerned almost entirely with the happenings at the Byzantine court and is one of the most gossipy and amusing narratives ever written on such a subject. His psychological insight and his lively and subtle style delighted the educated Byzantines. Anna Comnena, the daughter and biographer of the emperor Alexius I, greatly admired Psellus. Her own *Alexiad* is a much less fascinating work, but the recovery of the Byzantine power under her father provided her with an important theme.

Procopius' histories

The last, increasingly disastrous, centuries of Byzantine history are recorded by a series of scholarly and interesting historians. Nicetas Acominatus, a high imperial official, provides a surprisingly balanced eyewitness account of the siege and capture of Constantinople by the forces of the Fourth Crusade (1202–04). George Acropolites, a leading adviser of the Greek emperors of Nicaea, carries the story from 1203 to the recapture of Constantinople by the Byzantines in 1261. The later 13th and 14th centuries are covered by a succession of writers deeply immersed in contemporary theological disputations. Perhaps the most readable of all Byzantine histories is the largely autobiographical work of the leading politician and emperor John VI (reigned 1347 to 1354), written after his deposition during his years of enforced retirement in a monastery. George Sphrantzes, a close friend of the last emperor, Constantine XI, included in his history an eyewitness account of the siege and capture of Constantinople by Mehmed II in 1453. Two of Sphrantzes' contemporaries chose to write primarily about the Turks. Their methods place them among Renaissance historians. Laonicos Chalcocondyles wrote (in about 1464) an account of the rise of the Turkish state. He did so in the manner of Herodotus, with long digressions on various neighbouring nations. A little later, Critobulos of Imbros, in his account of the Turkish conquest of Constantinople, made Mehmed II his chief hero and modelled his history on Thucydides.

The study of what might be called "historical antiquities" was not much cultivated by Byzantine scholars. The most notable exception was the emperor Constantine VII, but only some fragments of his voluminous collections have survived (dating from about 940 to 959). They include a very interesting account of the various peoples with whom the Byzantines had to deal. Such ancient Greek literature as still survives, including that of all the historians, was preserved by the Byzantine scholars. When, around the year 1400, the teaching of Greek was introduced into Italian universities by Byzantine scholars, they brought also their superior techniques of literary scholarship, transforming thereby the study of Latin authors as well as introducing into western Europe the treasures of Greek literature. One result was the emergence of the new Renaissance historiography.

HISTORIOGRAPHY IN THE RENAISSANCE

If there is one thing that united the men of the Renaissance, it was the notion of belonging to a new time. Lorenzo Valla, one of the ablest of the early Humanists, in a preliminary draft of his history of King Ferdinand I of Aragon (written in 1445–46), proudly enumerates the modern technical inventions made in recent centuries, and especially near his own day. The sense of the novelty and excellence of their achievements was particularly felt by the men of the Renaissance in connection with their attempts to imitate the works of the ancient Greek and Roman writers and artists. They were not yet claiming that an era of unlimited progress was dawning for mankind—such concepts belong to the 18th century—but the belief in the progressiveness of their own age soon spurred the best Renaissance scholars and artists into achievements that, in some important respects, surpassed their ancient models. This happened in historiography, and especially in the sciences connected with it. The pace of change must not be exaggerated, however. Despite promising beginnings, historiography as a systematic discipline did not emerge during the Renaissance and, in fact, this development did not occur until the 19th century. The reasons for this delay form one of the main problems in any study of historiography between the years 1400 and 1800.

In the early Renaissance one by-product of the newly won sense of modernity was the tendency to regard the millennium between the collapse of the Roman Empire in the West and the 15th century as an era of prolonged decline. The concept of the Middle Ages was thus introduced for this intervening period. Two very important histories written in the first half of the 15th century deliberately concentrate on the medieval centuries. Their authors were leading Italian Humanists. The first to appear was the *Historiae Florentini populi* ("History of Florence") of Leonardo Bruni, the city's chancellor from 1427 to 1444. The second, the *Historiarum ab inclinatione Romanorum imperii decades* ("Decades"; mainly devoted to Italy) was written by Flavio Biondo, an important papal official. It covered the period from the sack of Rome by Alaric in AD 410 to the writer's own time. The "invention" of the Middle Ages as a separate historical period remains one of the most enduring legacies of Renaissance historiography.

Denigration of the Middle Ages

Unlike the medieval historians, the Renaissance Humanists became much more acutely aware of the process of historical change. This was a gradual development. They were trying to understand the ancient writers, whom they were seeking to emulate, and they became increasingly aware of the need to replace these writers in their correct historical setting. When Petrarch (1304–74), the pioneer Italian Humanist, unearthed in 1345 a collection of Cicero's letters, he was shocked to discover that Cicero was not a cloistered scholar of the medieval tradition but a busy politician who wrote his dialogues in moments of banishment from active life. In 1361, in a

letter to the Holy Roman emperor Charles IV, Petrarch was able to use his increased familiarity with classical documents to expose a medieval forgery of the Austrian archduke masquerading as a charter of Julius Caesar.

Between about 1440 and his death in 1457, Valla was one of the most influential Humanists. His *Elegantiae linguae latinae* (1444; "Elegancies of the Latin Language") was a treasury of information about correct Latin usages. For Valla the meaning of words was not natural but conventional and historical, because it was derived from changing custom. Thus a sense of ceaseless historical evolution was planted at the very centre of Humanist preoccupations with the recovery, the correction, and the interpretation of ancient texts.

In 1440 Valla's patron, King Alfonso of Naples, at war with the papacy, asked Valla to write some treatise against Pope Eugenius IV. Valla obliged by decisively disproving, on both linguistic and historical grounds, the genuineness of the "Donation of Constantine." From the middle of the 8th century, when this document was probably concocted, it had been used by the popes as one of the weightiest justifications for their claims to secular authority in Italy. Its authenticity had been sometimes questioned in the past by some of the acutest minds, such as Bishop Otto of Freising in the 12th century and Marsilius of Padua in the first half of the 14th century, but it required Valla's expert techniques to dispose of the "Donation" forever. The validity of Valla's methods of historical criticism was at once recognized by at least one other leading Humanist. Biondo wrote the relevant portions of his "Decades" of papal and Italian history between 1440 and 1443, while remaining in the service of the very same Eugenius IV who had been the chief object of Valla's attack. Yet Biondo tacitly accepted Valla's conclusions, and he never mentions the "Donation of Constantine." Biondo's critical outlook found still another expression in his summary dismissal of the fabulous history of Geoffrey of Monmouth. In his copy of Geoffrey he entered only a single note: "I have never come across anything so stuffed with lies and frivolities."

Historical philology

Valla's work on the texts of the New Testament proved in the long run to be one of the most influential applications of the new science of historical philology. His aim was to recover, so far as possible, the original Greek version through the use of the oldest extant manuscripts. He defended these researches by pointing out that he was not correcting the Holy Scriptures but merely the Latin Vulgate translation of St. Jerome that had been adopted by the Catholic Church. The revolutionary nature of Valla's historical approach comes out most strikingly in his comment that "none of the words of Christ have come to us, for Christ spoke in Hebrew and never wrote down anything." The corrections assembled by Valla became generally known when, in 1505, Erasmus published them as *Annotationes* on the New Testament. They provided a model for Erasmus' edition of a Greek New Testament in 1516, from which stem all the new Protestant versions of the 16th century.

The new historical philology was also soon applied to the study of philosophical and legal texts. In this, the most striking progress was made in the second half of the 15th century by Politian, who lectured at Florence, and by his friend Ermolao Barbaro, who taught at Padua. They were inaugurating the history of ideas and of intellectual movements. In his studies of Aristotelian texts, Barbaro insisted on using only the commentators of antiquity. In his lectures and writings (1489–94), Politian tried to re-establish from internal evidence the correct sequence of Aristotelian treatises, and he traced the gradual liberation of Aristotle's thought from the influence of Plato. The meaning of the terms used by Aristotle was rigorously investigated in the light of the linguistic usage of his Greek contemporaries. Politian's ventures into the field of legal texts proved particularly influential. He had at his disposal a very good 6th-century version of the Digest—that is, the section of Justinian's Corpus (body of civil law) based on the rulings of the Roman jurists. Politian's collation of it with the first printed edition of the Digest (in 1490) formed part of an inquiry into the transmission of the texts of the Roman law during the Middle Ages. Politian's researches stimulated a remarkable school of Humanist jurists, mostly Frenchmen, headed by Guillaume Budé, who published the first historical commentary on the Digest in 1508. In the course of the 16th century, these scholars laid the foundations of a new branch of scholarship, the history of laws and institutions.

The methods of textual criticism used by Politian and his friends were designed to produce definitive editions of classical texts. Politian was aware of the need to establish the correct descent of manuscripts and to disentangle the best textual tradition. In all this he was far ahead of almost all his contemporaries, and he was anticipating the procedures that were systematically adopted for the first time by Karl Lachmann and other German scholars in the 19th century. The historical philology of Politian was a program for the future rather than a dawn of a new era in the editing of classical texts. In contrast to his methods, most of the other Humanist editions of the Latin and Greek classics are very unsatisfactory. This is particularly true of the editions produced between about 1400 and 1550. The reckless emendations of Humanist editors, coupled with the subsequent disappearance of some of the manuscripts used by them, created grave problems for later scholars. Ever since the 17th century the task of the more modern editors has consisted largely in reconstructing, so far as possible, the manuscript versions available before 1400.

Modern historiography was created in the 19th century through a successful combination of the use of narrative sources with every other type of evidence. Some 15th-century Italian Humanists were already aware of these possibilities. The idea of recovering an entire civilization through a systematic collection of all the relics of the past was not alien to them. Biondo used mainly conventional narrative sources for his "Decades" of Italian history, but his description of the city of Rome in antiquity (*Roma instaurata*, 1444–46) was based on a novel combination of the narratives of other historians with a wide range of miscellaneous sources. These included topographical guides, public and private documents, studies of surviving buildings, inscriptions, and coins. But in practice most histories and biographies continued to be written in a conventional way, while the revived study of "antiquities" was cultivated in separation from narrative historiography.

Biondo's description of ancient Rome

Imitation of ancient models is the feature most often stressed in the modern descriptions of Humanist histories. This meant that style mattered at least as much as content and that historical truth might be obscured by literary conventions. On the more positive side, there was the renewed insistence on the choice of definite, clearly delimited subjects and on a more coherent arrangement of material. The abler Humanist historians, however, were also making innovations that bring their practice a little nearer to present notions of writing history.

Several Humanist historians were particularly attracted to the study of the origins of the states about which they were writing. In the 15th century Bruni did this for Florence, and Biondo and Bernardo Giustiniani for Venice, to mention some notable examples. In the 16th and early 17th centuries, French and English scholars inaugurated a critical study of the origins of their national institutions. Humanist historians prided themselves on their critical ability to overthrow the legends in which various countries had concealed their ignorance of their own origins. The incentives to revise the earliest history were often political. Bruni deemed it essential to prove that Florence had not been founded under the tyranny of the Roman emperors but in the time of the free republic. He happened to be right. The Humanist historians were more confident than their ancient predecessors that they could write competent histories of a remote past. In practice they were much less successful in this than they imagined. In dealing with periods before their own time, they usually followed only a restricted number of earlier narratives, though the best of them, such as Bruni and Biondo, displayed in their histories of medieval Italy

a novel ingenuity in combining well-chosen sources. Biondo, for example, made effective use of Dante's correspondence.

There was also some modest progress through the better use of documentary sources. This is often far from obvious, because Humanist historians, like their ancient predecessors, do not usually refer to their sources, even when they quote texts verbatim. Hence came Leopold von Ranke's utter misjudgment of the historical value of the *Storia d'Italia* ("History of Italy") of Francesco Guicciardini. Before Ranke's time it was universally accepted as the most authoritative contemporary history of Italy in the years 1494 to 1534. Ranke, who became one of the pioneers of "scientific" history in Germany, first established his reputation in 1824 by his attack on the reliability of Guicciardini. Ranke argued that the statements of that great Florentine statesman were contradicted by documentary evidence and that his history must have been based on unreliable secondary authorities. The discovery in the 20th century of Guicciardini's private archive proved that his history was scrupulously based on original documents of the highest value.

Guicciardini's history of Italy

Guicciardini, in a work that forms the nearest Renaissance parallel to the history of Thucydides, tries to comprehend the succession of tragedies that befell Italy from the start of the French invasions in 1494. This desire to recapture the rational causes of events is one of the most mature features of the best Renaissance historiography.

EARLY MODERN HISTORIOGRAPHY

Italian Humanist historians provided models that could be imitated easily in other countries. Almost everywhere in western and central Europe, local writers were encouraged to produce descriptions and histories of their own lands, intent with patriotic pride. In such countries as Spain and Poland, which had only recently achieved their unity, this was a way of commemorating their newly won cohesion. In the 15th century it was the object of a pioneer work on the earliest antiquities of Spain, the *Paralipomena Hispaniae*, by the Catalan Humanist bishop Joan Margarit i Pau, and of the invaluable *Annales seu cronicae incliti regni Poloniae* ("History of Poland"), by Jan Długosz, which included an exceptionally precise geographic description of his country. In Germany a sense of national identity could be vindicated by Humanist historians striving to minimize the importance of the continued political division of their land. The *Germania* of Tacitus was printed in Germany as early as 1473 and started the fashion of using this collective name for that country. Tacitus called the Germans "the indigenous inhabitants." This was used by a leading patriotic Humanist, Conradus Celtis, as a proof that Germany should be free from all foreign domination. Celtis and his other Humanist contemporaries deliberately hunted for manuscripts of medieval German writers to prove that their country, despite its disunity, could have a national history. Some important masterpieces were recovered, including the histories of Otto of Freising. Celtis' pet project of a description of Germany modelled on Biondo's *Italia illustrata* was carried out in 1530 by Sebastian Münster, and Münster's fuller *Cosmographia* (1544; "Cosmography"), though purporting to describe the known world, devoted one-half of its 818 pages to "the German nation." There was also a spate of histories of Germany, mostly very laborious and unreflective but incorporating the newly rediscovered medieval narratives and even some documentary sources. Greater originality came only in the wake of the Reformation. The same thing happened in France and in England. In both countries patriotic preoccupations were a leading feature of works written by Humanist historians, and the appearance of Protestantism reinforced in a peculiar way the existing nationalist tendencies.

The influence of the Reformation on historiography must first be discussed at a more universal level. As the philosopher Francis Bacon shrewdly observed, Martin Luther had been obliged "to awake all antiquity and to call former times to his succours . . . so that the ancient authors . . . which had a long time slept in libraries, began generally to be read." This was not because Luther would have regarded himself as a historian. But as early as 1519, in his disputation with Johann Eck, he encountered the assertion that the primacy of the pope was of divine origin. In order to disprove this and to demonstrate that they alone represented the true church, the Protestants had to retell in a new way the entire history of Christianity. In a preface to the *Vitae Romanorum pontificum* ("Lives of the Pontiffs"), published by Robert Barnes in 1535, Luther himself confessed that, although he himself had not originally attacked the papacy with historical arguments,

> now it is a wonderful delight to me to find that others are doing the same thing . . . from history—and it gives me the greatest joy . . . to see . . . that history and Scripture entirely coincide in this respect.

Protestant history

The starting point for the Protestant rewriting of Christian history could best be found in St. Augustine's teachings. The true church, the city of God, had always existed, even though at times it seemed to be overshadowed by the enemies of the divine order. Those enemies were not only the pagans and the heretics, as St. Augustine had believed. In more recent times they had included also the upholders of the papal authority and the persecutors of such medieval true Christians as John Wycliffe (died 1384) and John Hus (died 1415). The writings of Eusebius provided the model for chronicling the sufferings of the faithful until the dawn of freedom for the true church in the 16th century. These views about the correct history of Christianity were presented with exceptional cogency in John Calvin's *Christianiae religionis institutio* (fullest edition 1559; *Institutes of the Christian Religion*) and were shared by most Protestant scholars. The only obvious disagreements arose when Protestants tried to pinpoint the moment at which the church took the fatal turn away from God's true purpose. While the radical sectarians considered that the papacy had always been corrupt, less extremist Protestants were prepared to accept the earlier popes and to argue that the rot set in at some date between the time of Eusebius (died *c.* 340) and the 7th century. The choice of precise date might depend on the national traditions of each country. Thus, Bishop Richard Davies, in his preface to the New Testament in Welsh (1567), treats Pope Gregory the Great (died 604) as a special enemy because Gregory's effort to convert the Anglo-Saxons led ultimately to the subjugation of the autonomous British church.

Historians writing in this spirit were incapable of impartiality. But the historical controversies between the Catholics and the Protestants produced from both sides huge compilations. Their authors were determined to prove their respective cases by a stupendous marshalling of authorities and documentary sources. The habit of giving copious references and long, exact quotations, missing from the Humanist historiography, was reintroduced by the religious controversialists. On the Protestant side, the largest work is the *Ecclesiastica historia*, or the so-called *Centuriae Magdeburgenses* (13 volumes, 1559–74; "Magdeburg Centuries"), retelling the history of the church down to 1200. The Catholic reply, equally huge and graceless, was produced in 12 volumes by Cardinal Baronius. The chief Protestant critic of this work, the great Greek scholar Isaac Casaubon, was astonished by the Cardinal's ignorance of Greek and Hebrew, his gross mistakes, and his boundless credulity.

The narratives of contemporary events written in the 16th and early 17th centuries by the participants in the religious struggles, though equally partisan, include some works of great historical value and high literary merit. The earliest and best German Protestant narrative, that by Johannes Sleidanus, received a grudging tribute from his great opponent, the Holy Roman emperor Charles V, who remarked that "the rogue has certainly known much . . . ; he has either been in our privy council or our Councilors have been traitors." John Foxe's *Book of Martyrs* (1563) contains a great mass of exact information about the persecution of reformed religion in England and Wales during the reign of Mary Tudor, and it has influenced many generations of British Protestants.

Foxe's *Book of Martyrs*

The achievements of Queen Elizabeth I and the Anglican Church's settlement of her reign found an outstanding defender in William Camden, who was encouraged to write by Elizabeth's leading ministers. In his *Annales Rerum Anglicarum, et Hibernicarum Regnante Elizabetha* ("Annals of Elizabeth's Reign") Camden made excellent use of a mass of official records at his disposal, though his treatment of confidential matters had to be discreet.

Out of a conflict between Venice and the papacy in the first years of the 17th century was born the *Istoria del consilio tridentino* (1619; *History of the Council of Trent*) of Fra Paolo Sarpi. A Catholic friar, but a passionate defender of Venetian autonomy, Sarpi drew a dark picture of worldly papal policies and the unscrupulous machinations of the Jesuits. It is a bitter, prejudiced, but splendidly written and well-informed work, which profoundly influenced the anticlerical historians of the 18th century. All these contemporary narratives, however, have one serious limitation. They deal almost exclusively with political events and with changes in ecclesiastical organization. The Protestant schism is treated as merely a revolt against the abuses of the old church, and the deeper reasons for the alienation of the Protestants from the Catholic faith are never explained. Furthermore, these historians, by attributing the origins of the schism almost exclusively to Luther's sudden conflict with the papacy, obscured the existence in the early 16th century of numerous Catholic reformers, whose sole aim was to transform the Catholic Church from within. This one-sided approach to the history of the Reformation was destined to persist for a long time. Two influential histories published in the years 1683–88, one by a great Catholic prelate, Bishop Jacques-Bénigne Bossuet, and the other by Pierre Jurieu, a leading Protestant, still agreed on the same superficial account of the causes of the Reformation.

The rewriting by the Protestants of universal church history naturally involved a drastic revision of the history of the national churches. In Germany, particularly, the history of the church had become inextricably intermixed with the destinies of the German empire. Their hatred of the papacy made the Lutherans visualize the course of German history with unusual clarity. Nobody before them had attempted to impose on that history a single intelligible pattern of any sort. Theirs was bound to be a prejudiced pattern, a story of gradual national disintegration as the result of the successive defeats of the German emperors by the papacy. Johannes Stumpf's tragic chronicle of the Holy Roman emperor Henry IV (published in 1556) treated his struggles with Pope Gregory VII as the beginning of the empire's tribulations. The whole course of German history was retraced in this fashion under the influence of Luther's chief Humanist collaborator, Philipp Melanchthon, in the so-called *Chronicle* of Carion, written in its final versions (1572–73) by Melanchthon's son-in-law, Caspar Peucer.

One of the most novel features of the English Protestant historiography was the reawakening of scholarly interest in the period before the Norman Conquest of England in the 11th century. Matthew Parker, Queen Elizabeth's first archbishop of Canterbury, thought he could discern in the pre-Conquest church elements of true Christianity that were destroyed thereafter and had only been reintroduced by the Protestants. The Anglican Church could be represented as a return to the traditional practices and beliefs of the early English Christians. Thus the replacement of Latin by English in the Protestant church services could be justified by citing the presence in Anglo-Saxon England of Bibles, liturgies, and devotional literature in the Old English language. Parker and his friend Lord Burghley, Elizabeth's most trusted minister, gathered around them a circle of enthusiastic scholars, whose work preserved most of the important Anglo-Saxon texts as well as of some leading post-Conquest chronicles. Parker's own method of editing texts horrifies modern scholars, but some of the antiquarian works published by members of this group were of high quality. Camden's *Britannia* (first edition 1586, later much enlarged) was a pioneer work on the topography of Roman and early medieval Britain. The edition by Sir Henry Spelman of the records of the pre-Conquest church councils was the first serious attempt to apply to an important type of early sources the methods of best continental scholarship.

Camden's *Britannia*

The growth of a historical outlook can be traced in the 16th century in many diverse fields of learning. For the first time men were realizing that there was a historical side to every branch of knowledge concerned with human affairs. "I have become aware that law books are the products of history," wrote the French legal historian François Baudouin in 1561. In each branch of study there developed a special historical technique particularly appropriate to it. The most sophisticated scholarship was to be found in the field of classical studies. A group of scholars active in the second half of the 16th century were achieving results much superior to the work of the earlier Renaissance classicists. They combined philological expertise with a determination to reach a really adequate understanding of the ancient Greek and Roman civilizations. A few were Italians, such as Carlo Sigonio, but most of the important works were written in France and in the Protestant centres of Switzerland and Holland. As textual critics these scholars were reacting sharply to the earlier, more haphazard, methods of emending and editing classical authors. They were trying to bring the text of one writer after another to a state of near perfection. Some leading ancient historians, such as Tacitus, benefitted greatly from this treatment (edition of Lipsius in 1575). Though their methods do not quite reach the standards of modern scholarship, they anticipate intelligently many of the procedures more systematically adopted in the 19th century. Isaac Casaubon was the first to point out in his edition of Suetonius (1595) that Einhard's 9th-century life of Charlemagne was modelled on the work of that Roman historian. Casaubon's friend Joseph Scaliger renewed the science of classical chronology (1583) and was the first to reconstruct the original Greek *Chronicle* of Eusebius lying behind St. Jerome's Latin translation. Sigonio's pioneer work on the rights and duties of Roman citizens (1560) was later much used by Theodor Mommsen, one of the founders in the 19th century of the modern study of Roman history.

In the course of the 16th century, non-narrative historical work of the highest originality and complexity was being carried on in the legal faculties of French universities. One important stimulus was provided by the existence in France of different legal systems—the uncodified provincial customs in the north and the written law in the south. The latter ultimately derived from the Roman law, and, in the southern French universities, there arose an eager demand for the introduction of the new Italian methods of interpreting the Roman legal texts. Andrea Alciato, a pioneer in the historical treatment of the Roman law, taught at Bourges from 1529 to 1533, and his pupils founded the "Romanist" school of French legal historians.

Legal histories

Important advances were made in the study of both the Roman law and of the origins of the French legal customs, laying virtually the foundations of a new branch of scholarship, the history of law and institutions. François Baudouin published in 1545 the first historical survey of the development of the Roman legal science. The treatise on the custom of Paris by Charles Dumoulin (published 1539–58) resulted from his advocacy of the codification of the northern French legal customs. It was the first scholarly exposition of a body of customary French law derived from feudal practices, and it amounted to a first comprehensive history of European feudalism. It prompted a series of controversial works by a succession of scholars. The Roman, the Germanic, and the Celtic roots of feudalism all found advocates, and the respective claims of Lombard and Frankish texts to provide the best clues were vigorously canvassed. The complexity of the problems presented by the unravelling of the origins of feudalism dawned on scholars for the first time. The most valuable of these attempts to rediscover the "ancient French constitution" were the re-

searches on "the antiquities of France" of Étienne Pasquier (published 1560–1607), which form a basis for all later study of medieval French institutions.

One of the novel features of European civilization in the later 16th and 17th centuries was a secularization of mental interests. Secular learning could now produce ideas more fascinating to intelligent men than theology. History was one of the most popular types of literature sought by a growing reading public. Several treatises on the proper way of writing history appeared in the third quarter of the 16th century. An anthology consisting of 12 such works, including the famous *Methodus* of the French political philosopher Jean Bodin, was published at Basel in 1576. Nearly 100 years later a "Catalogue of the Most Vendible Books in England" (1657) showed that history books constituted a large proportion of the total works published. It has been estimated that between 1460 and 1700 at least 2,500,000 copies of 17 leading ancient historians were published in Europe.

The late 16th century and the 17th witnessed the publication of several great collections of historical materials. The men who undertook these gigantic tasks often were antiquarians accumulating miscellaneous records rather than historians, but they were supplying materials for generations of future historians. Some of the most important publications of sources appeared in France and the Netherlands. Pierre Pithou was a pioneer in editing materials for the history of the Frankish period. The collections of André Duchesne are a vast storehouse of chronicles and other sources for the study of medieval French history. Le Nain de Tillemont edited 20 volumes of records devoted to Roman and church history during the first six centuries of the Christian Era, which a century later furnished one of the principal sources for Edward Gibbon's work *The History of the Decline and Fall of the Roman Empire*. In 1629 a Belgian Jesuit, Jean Bolland, embarked systematically on the editing of records connected with all the saints whose feasts had at any time been celebrated by the church, and this series of publications has been continued to the present day. In the second half of the 17th century, the French Benedictine congregation of Saint-Maur started an immense series of publications commemorating the history of the Benedictines and of other monastic orders. The greatest Maurist scholar, Jean Mabillon, was accepted throughout Europe as the most erudite historian of his time.

In spite of its popularity among an expanding reading public and of the large number of learned editions of materials that it inspired, history was not, for most of the 17th century, one of the sciences that made men proud of living in a modern age. Immense progress was taking place in mathematics, astronomy, and physics. History not only did not seem capable of much further development, but scientifically minded men were beginning to dismiss it as a branch of knowledge that would never be worthy of serious respect. Mabillon's *De Re Diplomatica* (1681) helped to challenge this pessimistic view, but a further century elapsed before history began to be accepted as an authoritative discipline.

Hostility of rulers toward historiography

One major obstacle to the progress of historiography was the hostility of rulers to publications that did not favour their governments. The growth of an influential reading public made rulers increasingly suspicious of historical writings; for example, the censorship exercised by Cosimo I de' Medici, ruler of Florence from 1537 to 1574, precipitated the decline of Florentine historiography. Comparisons with the past also could be invidious. In 1599 Elizabeth I of England censured an author for describing the deposition of one of her predecessors, Richard II, 200 years earlier. Fear of possible trouble made highly intelligent scholars into one-sided historians. The great jurist Hugo Grotius avoided in his history of the wars of the Dutch against Spain discussions of the religious aspects. Samuel Pufendorf, the historian of the Swedish conquests, carefully left out the internal developments in 17th-century Sweden.

The scholars who in that century were responsible for the great advances in the mathematical sciences were convinced that their achievements would ultimately give mankind a novel mastery over its natural environment. This is particularly true of Francis Bacon and of René Descartes. Their optimism was laying the foundations for a belief in a possibility of continuous progress without which the purposeful and assured historiography of the 19th century would be inconceivable. But the attitude toward history of most of the leading thinkers and scientists of the 17th century was not helpful to its immediate development. Bacon, who wrote a readable and rationally argued biography of King Henry VII of England, attached no importance to accuracy; for example, he antedated Henry's death by a whole year and could not be bothered to undertake any detailed research. Gottfried Wilhelm Leibniz was a great mathematician, but his attempts to apply science to historiography led to mechanistic constructions from which real human beings were largely missing. Numerous influential thinkers were decidedly hostile to history. Descartes, the most eminent of the anti-historical scientists, was not simply disgusted by the unsystematic and imprecise methods of the historians of his time but also doubted whether, strictly speaking, history could be regarded as a branch of knowledge at all. But it is important to remember that much of the 17th-century criticism of history was an attitude of men who simply had other priorities and were concerned to attack doctrines that, for one reason or another, historians seemed to support. In the late 17th century the most successful defenders of history were the members of certain particularly scholarly Catholic orders. Catholicism rested its authority on tradition to a much greater extent than did its Protestant opponents. For Catholic scholars such as Mabillon, the defense of history became really a defense of their religion. They were trying to show that historians were capable of discovering scientifically demonstrable truths. The decisive publication was Mabillon's *De Re Diplomatica* of 1681. A member of a rival order, the Jesuit Daniel Papebroch, had challenged (in 1675) the authenticity of the oldest charters of two French Benedictine monasteries, Saint-Denis and Corbie. Mabillon applied his powerful critical intelligence not only to vindicating these documents but also to formulating the general rules that must be used to prove the authenticity of medieval records. He illustrated his rules by admirable examples and stated his conclusions with a candor and a common sense that convinced most readers. Mabillon's survey of the tests that must be applied by scholars covered the writing materials, the scripts (thus founding the science of medieval Latin paleography), the seals and other devices of authentication, the official formulas, and the vocabulary used at different periods. Above all, he stressed that the authenticity of a document usually rested not just on isolated details but on consistent correctness of all its features.

The rules of diplomatics

Mabillon was not just a "historical scientist." He had a passionate interest in the past and a vivid historical imagination. He displayed these qualities abundantly in his last and most important work, the *Annales Ordinis s. Benedicti* ("Annals of the Benedictine Order," to 1066). In the *Traité des études monastiques* (1691; "Manual of Monastic Studies"), he defended the importance of scholarly work as the principal activity of an elite of Benedictine monks. But it would be an anachronism to regard Mabillon and his chief associates as fully comparable to modern historians. They were constrained by the limitations of their time and of their special position as monks. For example, Bernard de Montfaucon, Mabillon's most important successor, is the creator of the science of medieval Greek paleography. But he shares with most of his contemporaries a complete inability to treat the Old Testament as a historical source.

Historical and antiquarian studies developed in 17th-century England in several very distinctive ways. The political struggles and religious controversies of that period made some issues of older English history into matters of immediate practical importance. The other distinctive feature was the delay in the absorption of European continental learning, so that the great progress made in the study of feudal origins in the 16th century began to affect the thinking of English scholars only by

about 1625. But there persisted also elements of continuity growing out of earlier Tudor scholarship. The interest in the Anglo-Saxon church and civilization continued to stimulate important editions of records throughout the 17th and early 18th centuries, including, especially, Sir Henry Spelman's edition of the records of church councils and Sir William Dugdale's *Monasticon Anglicanum* (1655–73), which is still valuable today. Another element of continuity with the Tudor period was the perennial interest of the English notables in heraldry, genealogy, and the antiquities of their native regions. Dugdale's *Antiquities of Warwickshire* (1656) set a pattern and a standard for county histories.

Students of English law and institutions, lacking the stimulus provided for French lawyers by the diversity of legal systems and by the notable progress in the study of Roman law in that country, continued to ascribe immemorial origins to the common law of England and to approach the development of English institutions in a completely unhistorical spirit. Among the parliamentary opposition to the Stuarts, these attitudes were part of a belief in the "ancient constitution," which these sovereigns were supposed to be defying. Spelman, a devout Anglican and a royalist, though a moderate one, was perhaps the first major scholar to break away from this myth. Under the influence of continental publications and correspondents, he accepted that feudal tenure had been introduced into England after the Norman Conquest and that all the English institutions after 1066 must be redefined in feudal terms. But his discoveries were hidden in a dictionary of antiquarian words (*Archaeologus*, vol. 1, 1626; 2 vol. 1664) and made very little impact until some 50 years had elapsed. Spelman had an acute sense of historical development, and he sadly castigated his countrymen for their lack of it in their attitude to parliamentary origins:

Achievement of Sir Henry Spelman

> when States are departed from their original Constitution and that original by tract of time worn out of memory; the succeeding Ages viewing what is past by the present, conceive the former to have been like to that they live in. (*Of Parliaments*, written in about 1640, published 1698.)

His greatest contribution to English history was to grasp that parliaments had developed out of feudal assemblies convoked by the Norman kings and that the Commons were introduced into parliaments subsequently, as a result of the growing prosperity of the lesser landholders. These views first became generally accessible in the 1664 edition of Spelman's dictionary. They were adopted by Robert Brady (in 1681) and by other partisans of the Stuarts and expanded into a Royalist statement of the English past. Violently polemical though this view was, it did at least lay to rest the myth of the immemorial "ancient constitution." The Whig triumph at the Glorious Revolution of 1688, which established a doctrine that the king ruled by parliamentary consent, led to the neglect of these discoveries for much of the 18th century. This was the common fate of much of the research of 17th-century antiquarians, who were very much ahead of their time and were writing for a limited audience. John Aubrey's pioneer description in the 1670s of the prehistoric sites of Avebury and Stonehenge had to wait two centuries for full publication. Even the best of these antiquarians, such as Spelman and Dugdale, were less critical in their handling of the original sources than Mabillon was. Higher standards were reached by a few of their successors in the early 18th century, especially by Thomas Madox, whose *Formulare Anglicanum* (1702) imitated Mabillon by attempting a systematic introduction to English medieval documents. But this did not save Madox from prolonged oblivion. After about 1730 this English tradition of antiquarian scholarship largely ended and remained unfashionable for most of the 18th century.

HISTORIOGRAPHY IN THE AGE OF THE ENLIGHTENMENT

The impulse given to historiography by the Italian Humanists and the religious controversialists had largely spent itself by about 1715. Men knew again how to write rationally satisfying contemporary histories, though often it needed courage to do so. Much less progress had been achieved in reconstructing the more distant past. Impressive collections of historical materials were being accumulated, but most scholars still lacked the capacity to rethink the thoughts of past generations and thus really to understand them. Mabillon could write with insight about early Benedictine history, as he possessed both sympathy with the subject and adequate technical expertise, but he was exceptional. Spelman had grasped that a particular society would be molded in a peculiar way by its institutions. He could not reconstruct and explain the gradual changes from one set of institutions to a later one, but he was aware of the problem.

Judged by the quality of its historical output, the 18th century was not, on the whole, an age of successful historians, but some of the defects of earlier historiography were beginning to be overcome. There were also losses, however, for some of the achievements of the preceding period were in danger of being forgotten. In the leading countries of western Europe, religious controversies were becoming less important, and a massive secularization of interests took place, which affected even ecclesiastical scholars. The French Maurists continued until 1790 to publish imposing historical collections, but their choice of subjects was determined much less than in the time of Mabillon by religious priorities. The greatest Italian ecclesiastical disciple of Mabillon was Ludovico Antonio Muratori, a social reformer. In a divided country like Italy, the best way of expressing his patriotism lay in reminding Italians of the former greatness of their country. Muratori spent much of his long life on his editions of Italian medieval sources.

The nationalist motivation shown by Muratori was peculiar to Italy and also to parts of Germany, another divided country. Elsewhere in Europe there was a danger that, as men lost interest in constitutional or religious disputes that might be settled by appeals to the past, they might turn away altogether from history or at least neglect long stretches of it. This did happen to some extent in the 18th century. Some of the radical French reformers, such as Jean Le Rond d'Alembert, one of the main inspirers in the 1750s of the French *Encyclopédie*, wanted to jettison completely much of the past. The Marquis de Condorcet (*q.v.*), an early prophet of the doctrine of endless progress of mankind and a pioneer historian of European civilization, was a prominent member of a French parliamentary commission that in 1792–93 deliberately destroyed some of the royal records as comprising relics of past servitude.

During much of the 18th century it was safer and easier to publish controversial works of history than it had been in the past. The point is important, as without this greater freedom, the peculiarly radical "philosophical" historiography, so typical of that century, would have been inconceivable. In Italy such writing was still dangerous. Pietro Giannone, the author of an anticlerical history of Naples (1723), was tracked down by the Inquisition and spent 12 years in prison, where he died in 1748. Even the great Muratori, who tried to help Giannone, came into danger of having some of his works banned and had to be rescued by the personal intervention of Pope Benedict XIV. In France, Louis XIV in 1714 imprisoned Nicolas Fréret in the Bastille for alleging (correctly) that the Franks were originally a confederacy of German tribes and not descendants of more illustrious ancestors. Under the successors of Louis, nothing quite so absurd happened again, but critics of the government or the church were often in trouble. Great Britain, Holland, Switzerland, and parts of Germany, on the other hand, provided safe oases where most things could be published. It was no accident that the most independent and historically minded group of German professors should have congregated at the University of Göttingen, founded in 1734, in the Hanoverian territory of the kings of Great Britain.

"Philosophical" historiography

A real renewal of historiography in the 18th century could only come if fresh reasons were discovered for making it again worthwhile. Nationalism could supply one such motive; but this only became decisively influential in the 19th century. An alternative was a histo-

riography inspired by the progress in the natural sciences and based on formulating the general rules governing the development of human societies. The chief features of this "new" historiography were a sense of the unity of all human history, including an interest in the continents outside Europe; a capacity for bold generalizations about the salient features of particular periods or societies; and a preference for topics connected with the progress of human civilization. Condorcet's historical sketch of the progress of the human mind, written in 1794, subdivided all known history into nine periods, each starting with some great invention or with geographical discoveries.

The shortcomings of this "rationalistic" historiography have been rehearsed often enough. For many of its writers it was primarily a weapon of propaganda against their enemies in church and state. Their redeeming virtue was the fearlessly critical attitude to all existing authorities, however august or sacred. The vast scale of their generalizations often precluded any detailed research. This was particularly true of the attempts to write histories of civilization, as the existing collections of printed materials did not cater for such interests, while systematic research in archives was seldom possible in the 18th century. In preparing his pioneer essay on the history of civilization, covering the millennium from the Carolingians to Louis XIV (*Essai sur les moeurs et l'esprit des nations*, 1745–53), the French author Voltaire had to collect bits and pieces from most diverse sources.

One of the most valuable achievements of the thinkers of the 18th century was their capacity to study particular societies as coherent units and to formulate the theory that the various aspects of each society's life were closely interrelated. This was not an entirely novel idea, but it first became commonly accepted during this period. Nor were all its adherents anticlericals. Giambattista Vico (*q.v.*), a Neapolitan Catholic, was ahead of his contemporaries in his particularly subtle sense of the complex influences by which one phase of society gives place to another. In his reconstruction of these transitions during the early stages of Roman history, he makes no clear lines between periods. His countryman Giannone explains in his autobiography that he had studied Roman law not for its own sake but in order to understand the changes in the society of the Roman Empire. The French philosopher Montesquieu, who owed much to Giannone, was not really a historian, but he displays an acute sense of historical realities. His *De l'esprit des lois* (1748; *The Spirit of the Laws*), more than any other book, accustomed his contemporaries to ponder the complex factors that shaped each society. It inspired Gibbon's definition of the kind of history he wanted to write. It was to be a "history related to and explained by the social institutions in which it is contained."

Gibbon's *Decline and Fall*

This ideal was realized in Gibbon's *History of the Decline and Fall of the Roman Empire* (1776–88), one of the masterpieces of "philosophical" historiography. Gibbon was preoccupied above all with the problem of human progress. The belief that continuous progress was possible for mankind had been publicly formulated in the mid-18th century by Anne-Robert-Jacques Turgot in France and by Adam Smith in Scotland, independently, it seems, of each other. Gibbon had read works and known scholars influenced by both these thinkers. A belief in continuous progress would confer a new purposefulness on the study of the entire course of human history and could justify a lengthy account of what otherwise might have seemed very obscure stretches of the past. Such a justification was to inspire most of the historiography of the 19th century. But the problem of progress had a special urgency for Gibbon's generation, which worried at the thought that their own enlightened civilization might also subsequently collapse. By unravelling the causes of the decline of the Roman Empire, Gibbon was determined to show that the Europe of his own day had attained a much superior degree of development and was immune from the fate of the ancient world.

In the 18th century, historiography was still only very rarely connected with the universities; and thus, except in such isolated places as Göttingen in Germany, no continuous schools of history could develop. Some of the most important achievements of the 18th-century historians meant much less to their contemporaries than to their successors in the 19th century. Gibbon was a pioneer in utilizing in a "rationalist" history the vast materials accumulated by generations of erudite antiquarians, but he had no immediate followers. The German archaeologist Johann Joachim Winckelmann tried to revive the true understanding of Greek sculpture and to make the history of art into something more than just the biographies of artists, but his work bore little fruit until the next century. The saddest fate was that of Vico's work. He was hardly ever read before the 19th century, when he at last influenced Barthold Georg Niebuhr and the rest of the German historical school, while Jules Michelet's rediscovery of Vico in 1824 started a new era in French writing on the Middle Ages.

HISTORIOGRAPHY IN THE 19TH AND 20TH CENTURIES

From the early 19th century, historiography began to develop in a radically different way. The decisive changes occurred among the German historians, largely through a reaction to the French Revolution and to a temporary subjugation of their country by Napoleon. Organized teaching of history in schools and universities became a matter of national importance, first in Prussia and then in other parts of Germany. As universal education spread to most European countries in the course of the 19th century, history was accepted everywhere as a necessary subject in schools. For the first time the bulk of historical writing came to be done by professional historians, for whom it became a condition of securing academic appointments or of consolidating their standings as university teachers. Historiography eventually became a continuously cooperative venture, where the achievements of past historians could be used systematically by their successors. But the growth of specialization and the bewildering number of types of works that came to be published constituted a new danger. In the past, important discoveries were frequently lost through lack of interest. But, by the second half of the 20th century, discoveries were in danger of being simply overlooked amid the flood of publications.

Growth of specialization

Another great change lay in the growth of intellectual freedom. Free expression of independent or unorthodox ideas had become dangerous during the French Revolution and under Napoleon, both in the territories controlled by the French and, by way of frightened reaction, in the lands of their unconquered opponents. After 1815 conditions for freer historiography improved gradually in much of Europe. Charles Darwin's *Origin of Species* (1859), which put forth a theory of evolution at first unacceptable to church authorities, probably could not have been published with the same impunity any earlier.

One feature of the growing tolerance of governments toward historiography was the gradual creation of public archives, such as the British Public Record Office in London, created in 1838, and the freer opening of the collections already in existence. Even the papacy accepted these changes, and Pope Leo XIII opened up the papal archive in 1883 as part of a deliberate new policy of encouraging historical study of Catholicism. For the first time historiography came to be based largely on unpublished records, and scholars were tempted into excessive reliance on original documents while unduly neglecting the older types of narrative sources.

In the 20th century some grievous threats to the persistence of free scholarship recurred, and historiography suffered with other branches of humane studies. The establishment of a Communist regime in Russia led, at first, to the rejection of most pre-1917 history as a fit subject for schools and universities. This decision was reversed in the 1930s, and from 1945 Communist countries were encouraging a form of historiography especially concerned with economic history and the class struggles of the past. There was also an enthusiastic interest in the material remains of past ages, leading to an impressive development of archaeology, particularly in Poland. The rise of dictatorships in Italy and Germany

had disastrous effects on historiography in those countries, and recovery after World War II was only gradual.

Judged merely by the number of "practicing" historians and of their publications, historiography seemed in a very flourishing state in the 1970s. Its European traditions had spread to all the other continents and were largely accepted in all non-Communist countries.

The *Introduction aux études historiques* (*Introduction to the Study of History*) of Charles V. Langlois and Charles Seignobos (1898), supplemented by critical comments of another outstanding French historian, Ferdinand Lot (in *Le Moyen Age*, 1898), provides an excellent starting point for the discussion of modern historical methods. History is an autonomous branch of learning, and some of its methods may be unique. Historians should not try to formulate general laws; their branch of learning merely "aims at explaining reality." Langlois and Seignobos particularly stress that history is not a science of observation but a science of reasoning how to extract from imperfect documentary or narrative records some glimpses of what actually happened.

The historian's task

A historian has to subject his sources to a whole series of preliminary investigations. First comes "external criticism," aimed at determining whether the sources are appropriate and adequate for the particular task in hand. The provenance, date, and authenticity of each source must be established by using the techniques of diplomatic, the detailed study and assessment of documents, and of paleography, the study of ancient handwriting, and of other auxiliary sciences that were elaborated after the 17th century. In France a special institution for teaching some of these techniques, the École des Chartes, was created in 1821. The first specialized seminar for instruction in these subjects was established in 1854 at Vienna by Theodor von Sickel, one of the greatest medievalists of the 19th century, and it was gradually imitated by leading German universities. One of the most important critical refinements introduced in the course of the 19th century was the improved handling of narrative sources brought about by seeking to discover the literary sources that lay behind them. Leopold von Ranke, one of the foremost German historians, who began his career as a teacher of classics, was gradually attracted to history through a desire to understand better the sources of the Greek and Latin authors whom he was expounding. In the later decades of the 19th century, such a quest became a normal feature of historical scholarship.

Once a historian has decided, through the application of "external criticism," on the sources that are relevant to his purpose, he must next, by "internal criticism," make sure that he fully understands what he has selected. German classical philologists were the first to bring these latter investigations to a high degree of perfection. Karl Lachmann, an editor of the Latin poets is justly regarded as the creator of modern textual criticism in its most rigorous forms, and historians gradually adopted similar methods. The language of the sources must be understood, corruptions in the text must be eliminated, and the historian must, as accurately as possible, penetrate the minds of the authors with whom he is dealing.

All these critical operations on the sources are merely preliminaries, and the work of the historian proper only starts when he attempts a synthesis of his materials. F. Lot stresses that in this qualities other than the erudite skills come into play. There must be sympathy with the subjects under study, for without it there can be no imaginative insight into the past. Ideally, a historian must display capacities akin to those of a poet or an artist.

Such a quality was, by and large, lacking in the work of the historians of the Enlightenment, who had been unable to achieve imaginative insight into civilizations very different from their own. The greatest shortcoming of Gibbon was his temperamental inability to appreciate religion. The new historiography of the 19th century was created chiefly by Germans, who, through a reaction to the ungodly and cosmopolitan Enlightenment, were endowed to excess with a passion for extolling the unique nature of their fatherland and for tracing the roots of this uniqueness through the whole course of German history. These developments in German historiography can be traced back to some strands of German thought in the 18th century, especially to some features of the writings of Johann Gottfried von Herder. He denied that the purpose of history was to provide a bird's-eye view of the progress of the human mind. It was, rather, to reconstruct history as it had been, which means that all countries and periods are equally deserving of study. This view anticipated Ranke's oft-quoted aim to describe what has actually happened and his conviction that the description of all human history displays the workings of God's providence. The disasters inflicted upon Germany by Napoleon brought forth a patriotic school of historians whose urgent task it became to propagate these views as a means of restoring German independence. The centre of this movement was in Prussia, at the newly founded University of Berlin (1809). Wilhelm von Humboldt, its effective founder, believed that the task of the historian lay in discovering the ideas behind the facts. The concepts that had special validity for him were ideas of religion and of a national state. The German historical school prided itself on the scientific precision of its methods, on its determination to get all the details right, and on the scrupulous quotation of sources. This display of exact scholarship represented a great gain for historical sciences, but its chief purpose was to convince the reader. Yet these German historians were fundamentally inspired by a prejudiced, arbitrary set of assumptions. It is particularly difficult to detect Ranke's hidden bias, as he made a parade of refusing to pass judgments on the past. His preference for the study of foreign relations between states and his treatment of states as natural entities with a right to fulfill their individual destinies justified the successes of Prussia. The defeat in 1848 of the German aspirations to national unity inspired his pupil Wilhelm von Giesebrecht to write the history of the medieval German empire to remind his countrymen of their past glories. When German unification was achieved in 1871, Giesebrecht doubted whether there was any need to bring out any further volumes of his great work. But many German historians, having contributed mightily to the unification of Germany, continued to describe complacently the triumphs of the Bismarckian state. This was one of the purposes of the school of historical economists led by Gustav von Schmoller. There were some dissenting voices. Theodor Mommsen (*q.v.*), the greatest historian of antiquity produced by the 19th century, deplored the tendency of his countrymen to worship state power. Friedrich Meinecke, a leading German historian of political ideas, who until 1914 accepted the ordinary nationalistic assumptions of his countrymen, gradually entirely changed his views and, after the defeat of Germany in two world wars, pleaded in his *Deutsche Katastrophe* (1946; *The German Catastrophe*) for a historiography concerned with the higher values of general civilization. Among the German historians, particularly striking progress was achieved in medieval studies. Meanwhile, attempts at imaginative reconstructions of the past were being made in other countries of western Europe. Jules Michelet wrote in 1833–43 the first history of medieval France based on the French national archives, of which he was at that time keeper. Macaulay's *History of England* (1848–61), covering chiefly the years 1685–1702, represented again a remarkable though prejudiced attempt to relive the past.

German patriotic historians

German scholarly techniques and the methods of German historical teaching spread to other countries in the course of the later 19th century, though it is important to note that until 1914 a significant proportion of leading historians from states outside Germany spent some time in that country. This is particularly true of some of the greatest Russian scholars, such as M.I. Rostovtzeff, one of the most important modern historians of antiquity. In England, William Stubbs, though self-taught, applied the results of German scholarship to the reconstruction of English medieval history. Gabriel Monod, who had studied in Germany, was prominent in introducing more scientific techniques into medieval French historiography, and he founded in 1876 the *Revue Historique*

as the main organ of French historical scholarship. A succession of American students went to Germany, and some, on their return home, reorganized historical studies. Measured by the sheer bulk of publications, the amount of American history written since the 18th century is probably greater than that of any other modern nation. But apart from editions of sources, very few works on American history published before around 1900 are of much practical use today. The most influential pioneer in organizing scientific historiography was Herbert Baxter Adams, who between 1876 and his death in 1901 made the Johns Hopkins University at Baltimore into the foremost American centre of historical studies. He was also one of the founders of the American Historical Association in 1884 and played a large part in successfully launching the *American Historical Review* in 1895 as the main organ of historical scholarship. Some of Adams' pupils became great scholars in various fields of general history. Charles Homer Haskins' works on Norman institutions and on science and culture in the 12th and 13th centuries made him one of the foremost medievalists of the 20th century. But a movement for creating a purely American history was launched in 1893 by another of Adams' pupils, Frederick Jackson Turner, who inaugurated a "progressive" school of historians through his conviction that the fundamental fact of American history down to 1890 was the settlement of a continent. In Turner's eyes the main theme of American history in the 19th century was the conflict between the patrician and capitalist groups of the Eastern Seaboard and the needs of the new settlers in the Middle West. Charles A. Beard inaugurated by his *Economic Interpretation of the American Constitution* (1913) an attempt to rewrite the entire history of the U.S. in terms of conflicts between different groups of economic interests. The weakness of this type of historiography was that it encouraged an excessive parochialism. After 1945 the "progressive" historians came under fire both from more conservative scholars who preferred to stress elements of common tradition and purpose in American development and from the historians of the "new left." In the 1960s and 1970s the close connection between writings on American history and the active political life was infusing great variety and vitality into its historiography, though making it perhaps too susceptible to rapidly changing exernal pressures.

The work of Henry Baxter Adams

Non-Western historiographical traditions

MUSLIM HISTORIOGRAPHY

A study of Muslim historiography in Asia and North Africa and of the historical writings of the peoples of South and East Asia confirms one main feature of the parallel European story. Except when the needs of religion necessitated it, as happened to some extent during the early centuries of Muslim expansion, the writing of history has seldom formed a vital component of any civilization until very recent times. Muslim historiography appears to have originally developed independently of European influences. Until the 19th century Muslim writers only very seldom consulted Christian sources and almost never noted events in Christian countries. Fortunately, they displayed at times more curiosity about the non-Muslim peoples of Asia. The first and best history of the Mongol conquests in the first half of the 13th century was the work of a Persian, Joveynī. On a visit to Mongolia in 1252–53, he was able to consult the recently compiled, earliest Mongol narrative (*Secret History of the Mongols*). The older Chinese historiography developed for over two millennia in isolation, and this led to a series of striking contrasts with Muslim and European tradition of historical writing.

Origins of Arabic historiography

The origins of Arabic historiography still remain obscure because of the gap between the legendary traditions of pre-Islāmic Arabia before the start of the Muslim era (AD 622) and the sophisticated and fairly exact chronicles that began to appear in the later 8th and 9th centuries. But while the detailed stages of this development still await reconstruction, the main influences shaping the early Muslim historiography are clear enough. As in the case of the ancient Jews, it was created and perpetuated by religion. Muḥammad (died 632) regarded himself as a successor to a long series of Jewish and Christian prophets, and he made Islām a religion with a strong sense of history. The Qur'ān, Islām's holy book, is full of warnings derived from the lessons of history.

Teachings of Muḥammad not included in the Qur'ān came to be regarded after his death as authoritative tradition left behind by him. All his sayings and actions were therefore carefully treasured and ultimately came to form, in combination with the Qur'ān, the foundation for the body of Muslim law (Sharīʿah), common to all Islāmic communities. These traditions (Ḥadīth) were transmitted orally for several generations, until they were written down in the 8th and 9th centuries. The resultant collections were only partly historical, as myths and inventions crept into them. The scholars who were engaged in preserving and verifying these traditions were chiefly preoccupied with organizing them into legal and theological systems, and they were frequently hostile to the historians. The earliest authoritative life of Muḥammad, written by Ibn Isḥāq (died 768), was attacked by a leading exponent of the legal "traditionist" learning. This confirms the independence of the historical scholars from the theological and legal interests. But both groups shared some common materials, and the strict rules evolved by the legal "traditionists" for recording their sources and tracing a continuous chain of authoritative transmitters of the traditions encouraged similar exact habits in the Muslim historians. The resultant histories were often pedantic, full of unrelated facts, and deficient in reflective comment, though there are some astonishing exceptions, such as the writings of Ibn Khaldūn (1332–1406). But the better Muslim historians scrupulously quoted their authorities and tried to be truthful. This was particularly true of the "classical" school of historians, who were writing at the centre of the ʿAbbāsid caliphate in Iraq in the 9th and 10th centuries. Aṭ-Ṭabarī (died 923), the most authoritative of them all, wrote his "History of Prophets and Kings" as a supplement to his earlier commentary on the Qur'ān, and subsequent Muslim historians were content to follow his reconstruction of the early Islāmic history. The Syrian and Iraqi historiography of the 12th and early 13th centuries is at least as valuable as the Western historical writing of this period, and sometimes it is clearly better.

To orthodox Muslims, the development of the Islāmic community represents a continuous manifestation of God's purpose. Consequently, the recording of the religious progress of the Islāmic society continued to be sacred duty. One of the original features of Muslim historiography is the large amount of attention devoted to the lives of devout men and of scholars. To many Muslim historians, these spiritual and intellectual activities were of much greater importance than the doings of princes and warriors. One of the peculiarities of Muslim historiography was the liking for encyclopaedic dictionaries of famous men. The earliest of these were devoted to the Companions of Muḥammad and to the early transmitters of the Muslim traditions. For a thousand years extremely diverse types of biographical collections have continued to appear in the Muslim world. Those devoted to religious scholars attained a particularly wide diffusion. Saladin (Ṣalāḥ ad-Dīn), who took Jerusalem from the crusaders in 1187 and later opposed the Third Crusade, offered to the Muslim writers the particularly congenial subject of a ruler dominated by a sense of religious duty. A particularly fine example of medieval Muslim historiography is the biography of Saladin by Bahā' ad-Dīn (died 1234), which gives an exceptional insight into Saladin's motives for many of his critical decisions.

The works of Ibn Khaldūn

But the greatest Arab historian and one of the most penetrating thinkers about historiography in any time or place was undoubtedly Ibn Khaldūn. The introduction (*al-Muqaddimah*) to his *Kitāb al-ʿibar*, a universal history (begun in 1375), is, in A.J. Toynbee's judgment (1934), "the greatest work of its kind that has ever yet been created by any mind." Ibn Khaldūn had absorbed all the learning accessible to a Muslim of his time. He

was a master of religious learning, an outstanding judge, a writer on logic. He turned a subtle and most disciplined mind to historiography in order to explain his personal tragedy. He had served a succession of rulers in Islāmic Spain and the Maghrib (Northwest Africa) as a general, a politician, and even once as a chief minister, and his activities had always ended in disaster. In order to explain what had gone wrong, he sought to achieve a correct understanding of the forces that governed the societies known to him. He concluded that political stability had become impossible in his native Maghrib, because over centuries economic prosperity had declined excessively and the forces of lawlessness had become too strong.

As a detailed chronicler of events Ibn Khaldūn is not always exact, but, like contemporary historians, he knew how to reconstruct correctly the main trends over several centuries. His ability to formulate general laws that govern the fate of societies and to establish rules for the criticism of sources provided him with an intelligent framework for the correct reconstruction of past history.

Ibn Khaldūn's *Muqaddimah* has survived in at least a score of manuscripts, but he has had no effective influence on Muslim historiography until recently; after his time, as before, the writing of history continued to be a normal feature of Muslim civilization in the more advanced Islāmic societies. In several countries, notably in parts of India, the first works that deserve the name of history appeared only after the Muslim conquest or the conversion to Islām. After the 12th century Arabic ceased to be the main language of Muslim historiography. Distinguished histories were written in Persian in the 13th century, and subsequently Turkish and other vernaculars came to be used by historians in different parts of the Islāmic world. But, in its isolation from non-Muslim influences and its traditional interests, Islāmic historiography underwent no intrinsic change until the 19th century, when it began to be affected by the impact of modern Western civilization.

EAST ASIAN HISTORIOGRAPHY

The preservation of some records of historical events can be traced in China to at least the early part of the 1st millennium BC. Confucius (551–479 BC) was credited, rightly or wrongly, in the later Chinese tradition with editing the annals of his native state of Lu. But the appearance of the first works fully deserving the name of histories resulted from the unification of China under a single ruler in 221 BC. The first such work to survive, the *Shih chi* ("Historical Records"), dates from *c.* 85 BC. Its author, Ssu-ma Ch'ien, is quite justifiably called the father of Chinese historiography. His history exhibits many of the main features of the later Chinese official histories as they continued to be written down to the deposition of the last Chinese imperial dynasty in 1911. Within this fairly unified tradition, China produced a mass of historical writings unequalled by any other country before modern times. Until the late 19th century, Japanese historiography formed an offshoot of this tradition.

The "Historical Records" of ancient China

Chinese scholars showed an interest in the history of China from the earliest times. According to the Chinese conception, history makes sense only if it can furnish practical directives for action or supply correct information upon which action can wisely be based. All the schools of Chinese thought quoted the lessons of history. Confucius, with his stress on the moral content of these lessons, formed part of this universal belief in the value of history. One of the duties inculcated by him was the scrupulous transmission of authentic records. When, some centuries after his death, the unified Imperial state began to recruit its bureaucracy among the Confucian scholars, the recording of all the necessary information and the careful preservation of records became one of the main functions of the Chinese government, both centrally and locally. A long series of official histories and of records connected with them has survived from the time of the T'ang dynasty (618–907) onward. From then on, the great bulk of Chinese history was written by bureaucrats for bureaucrats. From a practical point of view this immense body of historical writings fulfilled a very useful purpose. Such histories were bound to be highly stereotyped and restricted in content to what interested the higher officialdom. It is easy to condemn it by modern Western standards for its excessive preoccupation with concrete details and inability to produce works of wider synthesis. But this Chinese tradition did gradually evolve in the direction of greater rationality and subtlety. Its scope widened as the sphere of government expanded. Furthermore, within this tradition there appeared from time to time writers of genius, men of bold critical spirit, genuine historical insight, and overriding integrity. One of the greatest was Liu Chih-chi (661–721), the writer of the *Shih t'ung*, the first thorough treatise in Chinese, or any other language, on historical method, which also constituted in effect a history of Chinese historiography. He had a successor in Ssu-ma Kuang (1019–86), the author of the first fairly comprehensive general history of China (covering the years 403 BC–AD 959). In the 17th century a remarkable group of historical scholars virtually founded a school of critical Chinese philology. None of these writers succeeded in radically transforming Chinese historiography, but they created an increasingly sophisticated and critical tradition. Their successors in the 20th century assimilated some valuable features of modern Western historiography.

BIBLIOGRAPHY

General Works: C.V. LANGLOIS and C. SEIGNOBOS, *Introduction aux études historiques* (1898; Eng. trans., *Introduction to the Study of History*, 1898); HAROLD TEMPERLEY (ed.), *Selected Essays of J.B. Bury* (1930, reprinted 1964); JAMES T. SHOTWELL, *The History of History* (1939); JAMES W. THOMPSON and BERNARD J. HOLM, *A History of Historical Writing*, 2 vol. (1942); ROBIN G. COLLINGWOOD, *The Idea of History* (1946); HERBERT BUTTERFIELD, *Man on His Past* (1955).

Classical: J.B. BURY, *The Ancient Greek Historians* (1909, reprinted 1958); M.L.W. LAISTNER, *The Greater Roman Historians* (1947, reprinted 1963); MOSES I. FINLEY (ed.), *The Greek Historians* (1959), selected passages in translation with a valuable introduction; MAURICE PLATNAUER (ed.), *Fifty Years of Classical Scholarship* (1954; rev. ed. with appendixes, *Fifty Years (and Twelve) of Classical Scholarship*, 1968), especially chapters 6 by G.T. GRIFFITH and 13 by A.H. MACDONALD; ARNALDO MOMIGLIANO, *Studies in Historiography* (1966), a selection from his vast collection of valuable articles in *Contributo alla storia degli studi classici e del mondo antico*, 5 vol. (1959–69); T.A. DOREY (ed.), *Latin Historians* (1966) and *Latin Biography* (1967).

Byzantine: GYULA MORAVCSIK, *Byzantinoturcica*, 2nd ed., vol. 1 (1958), not always reliable in its judgments. There is no satisfactory study in English. There is much useful information in the appendixes to J.B. BURY's edition of *The History of the Decline and Fall of the Roman Empire*, 7 vol. (1896–1900, reprinted 1909–14).

Medieval: THOMAS F. TOUT, "The Study of Mediaeval Chronicles," *Bulletin of the John Rylands Library*, 6:414–438 (1921), reprinted in *The Collected Papers of Thomas Frederick Tout*, 3 vol. (1932–34); REGINALD L. POOLE, *Chronicles and Annals* (1926); M.L.W. LAISTNER, *Thought and Letters in Western Europe, A.D. 500 to 900*, new ed. (1957); CHARLES H. HASKINS, *The Renaissance of the Twelfth Century* (1927); T.A. DOREY (*op. cit.*); RICHARD W. SOUTHERN, "Aspects of the European Tradition of Historical Writing," *Transactions of the Royal Historical Society*, 5th Series, vol. 20–22 (1970–72).

The Renaissance: WALLACE K. FERGUSON, *The Renaissance in Historical Thought* (1948); DENYS HAY, "Flavio Biondo and the Middle Ages," *Proceedings of the British Academy*, 45: 97–128 (1960); MYRON GILMORE, *Humanists and Jurists* (1963); PAUL O. KRISTELLER, *Eight Philosophers of the Italian Renaissance*, ch. 2 (1964), on Valla; FELIX GILBERT, *Machiavelli and Guicciardini: Politics and History in Sixteenth-Century Florence* (1965); IDA MAIER, *Ange Politien* (1966); L.D. REYNOLDS and N.G. WILSON, *Scribes and Scholars: A Guide to the Transmission of Greek and Latin Literature* (1968); ROBERTO WEISS, *The Renaissance Discovery of Classical Antiquity* (1969); E.J. KENNEY, "The Character of Humanist Philology," in R.R. BOLGAR (ed.), *Classical Influences on European Culture, A.D. 500–1500* (1971).

Early modern (16th–17th centuries to c. 1680): J.G.A. POCOCK, *The Ancient Constitution and the Feudal Law: A Study of English Historical Thought in the Seventeenth Cen-*

tury (1957), also valuable for France; HERBERT BUTTERFIELD, "The History of Historiography and the History of Science," *Mélanges Alexandre Koyré*, vol. 2 (1964); and "Delays and Paradoxes in the Development of Historiography," in KENNETH BOURNE and D.C. WATT (ed.), *Studies in International History: Essays Presented to W. Norton Medlicott* (1967); GLANMOR WILLIAMS, *Reformation Views of Church History* (1970); DONALD R. KELLEY, *Foundations of Modern Historical Scholarship: Language, Law, and History in the French Renaissance* (1970); MAY MCKISACK, *Medieval History in the Tudor Age* (1971); G. STRAUSS, "The Course of German History: The Lutheran Interpretation," in ANTHONY MOLHO and JOHN A. TEDESCHI (eds.), *Renaissance: Studies in Honor of Hans Baron* (1971).

The creation of the science of diplomatic and the Enlightenment (c. 1680–1789): DAVID C. DOUGLAS, *English Scholars, 1660–1730*, 2nd rev. ed. (1951); MARTIN L. CLARKE, *Greek Studies in England, 1700–1830* (1945); DAVID KNOWLES, "Jean Mabillon," *Journal of Ecclesiastical History*, vol. 10, no. 2 (1959), and *Great Historical Enterprises* (1962); CHRISTOPHER DAWSON, "Edward Gibbon," *Proceedings of the British Academy*, 20:159–180 (1934); and EDWARD GIBBON, *The History of the Decline and Fall of the Roman Empire*, 6 vol. (1776–86, best modern edition by J.B. BURY, *op. cit.*); ARNALDO MOMIGLIANO, *Studies in Historiography* (1966) and *Terzo Contributo alla storia degli studi classici e del mondo antico* (1966), studies in English of Vico, Muratori, and other Italian scholars; HUGH TREVOR-ROPER, "The Historical Philosophy of the Enlightenment," in *Studies on Voltaire and the Eighteenth Century*, 27:1667–87 (1963), and "The Idea of the Decline and Fall of the Roman Empire," in *The Age of the Enlightenment* (1967).

Prehistory, archaeology, and other auxiliary sciences: SIR EDWARD M. THOMPSON, *An Introduction to Greek and Latin Palaeography* (1912); GLYN DANIEL, *A Hundred Years of Archaeology* (1950); J.G.D. CLARK, *Prehistoric Europe: The Economic Basis*, rev. ed. (1962); O.G.S. CRAWFORD, *Archaeology in the Field* (1953); J.G.D. CLARK, *World Prehistory* (1960); CHARLES SAMARAN (ed.), *L'Histoire et ses méthodes*, vol. 11 of the *Encyclopédie de la Pléiade* (1961); VIVIAN H. GALBRAITH, *An Introduction to the Study of History* (1964).

Writings on the history of the United States: ARTHUR M. SCHLESINGER (ed.), *Historical Scholarship in America* (1932); WILLIAM T. HUTCHINSON (ed.), *The Marcus W. Jernegan Essays in American Historiography* (1937); HUGH H. BELLOT, *American History and American Historians* (1952); DONALD SHEEHAN and HAROLD C. SYRETT (eds.), *Essays in American Historiography: Papers Presented in Honor of Allan Nevins* (1960); JOHN HIGHAM (ed.), *The Reconstruction of American History* (1962); THOMAS J. PRESSLY, *Americans Interpret Their Civil War*, 2nd ed. (1962); RICHARD HOFSTADTER, *The Progressive Historians, Turner, Beard, Parrington* (1968); MARCUS CUNLIFFE and ROBIN W. WINKS (eds.), *Pastmasters: Some Essays on American Historians* (1969).

The emergence of modern historiography (late-18th and 19th centuries): LORD ACTON, *Historical Essays and Studies* (1907), see especially "German Schools of History"; J.B. BURY, *The Idea of Progress: An Inquiry into Its Origin and Growth* (1920, reprinted 1960); EDMUND WILSON, *To the Finland Station* (1940), especially for French historiography; PIETER GEYL, *Napoleon, voor en tegen in de Franse geschiedschrijving* (1946; Eng. trans., *Napoleon, For and Against*, 1949); *Some Modern Historians of Britain: Essays in Honour of R.L. Schuyler* (1951); G.P. GOOCH, *History and Historians in the Nineteenth Century*, 2nd rev. ed. (1952); FERDINAND SCHEVILL, *Six Historians* (1956), particularly interesting on Ranke; GEORG G. IGGERS, "The Image of Ranke in American and German Historical Thought," in *History and Theory*, 2: 17–123 (1962); and *The German Conception of History* (1968); HENRY E. BELL, *Maitland: A Critical Examination and Assessment* (1965); FREDERICK M. BARNARD, *Herder's Social and Political Thought: From Enlightenment to Nationalism* (1965).

Twentieth century: EUGENE N. ANDERSON, "Meinecke's *Ideengeschichte* and the Crisis in Historical Thinking," in *Medieval and Historiographical Essays in Honor of James Westfall Thompson* (1938); MARC BLOCH, *Apologie pour l'histoire; ou, Métier d'historien* (1949; Eng. trans., *The Historian's Craft*, 1953); *Architects and Craftsmen in History: Festschrift für Abbott Payson Usher* (1956), biographies of leading historians; JACK H. HEXTER, *Reappraisals in History* (1961); RICHARD PARES, *The Historian's Business, and Other Essays* (1961), particularly valuable for the works of ARNOLD TOYNBEE; FRIEDRICH MEINECKE, *Die Idee der Staatsräson in der neueren Geschichte* (1957; Eng. trans., *Machiavellism: The Doctrine of Raison d'État and Its Place in Modern History*, 1957); DONALD C. WATT (ed.), *Contemporary History in Europe* (1968).

Russian historiography: There is no satisfactory general survey in English. The following can be useful for particular periods or historians: ANATOLE G. MAZOUR, *Modern Russian Historiography*, 2nd ed. (1958); ALEXANDER S. VUCINICH, *Science in Russian Culture: A History to 1860* (1963); and RICHARD PIPES (trans.), *Karamzin's Memoir on Ancient and Modern Russia* (1966); JOHN S. CURTISS (ed.), *Essays in Russian and Soviet History, in Honor of Gerold Tanquary Robinson* (1962), especially on Semevsky; ALAN D. FERGUSON and ALFRED LEVIN (eds.), *Essays in Russian History: A Collection Dedicated to George Vernadsky* (1964); JOHN KEEP and LILIANA BRISBY (eds.), *Contemporary History in the Soviet Mirror* (1964).

Muslim historiography: FRANZ ROSENTHAL, *A History of Muslim Historiography*, 2nd rev. ed. (1968).

East Asian historiography: WILLIAM G. BEASLEY and E.G. PULLEYBLANK, *Historians of China and Japan* (1961); CHARLES S. GARDNER, *Chinese Traditional Historiography* (1938, reprinted 1961).

(E.B.Fr.)

History, Philosophy of

The term history may be employed in two quite different senses: it may mean (1) the events and actions that together make up the human past, or (2) the accounts given of that past and the modes of investigation whereby they are arrived at or constructed. When used in the first sense, the word refers to what as a matter of fact happened, while when used in the second sense it refers to the study and description of those happenings. The notion of philosophical reflection upon history and its nature is consequently open to more than one interpretation, and contemporary writers have found it convenient to regard it as covering two main types of undertaking. On the one hand, they have distinguished philosophy of history in the traditional or classical sense; this is conceived to be a first-order enquiry, its subject matter being the historical process as a whole and its aim being, broadly speaking, one of providing an overall elucidation or explanation of the course and direction taken by that process. On the other hand, they have distinguished philosophy of history considered as a second-order enquiry; here attention is not focussed upon the actual sequence of events themselves but, instead, upon the procedures and categories used by practicing historians in approaching and comprehending their material. The former, often alluded to as speculative philosophy of history, has had a long and varied career; the latter, which is generally known as critical or analytical philosophy of history, has only risen to prominence during the 20th century.

SPECULATIVE THEORIES

The idea of an order or design in history. The belief that it is possible to discern in the course of human history some general scheme or design, some all-encompassing purpose or pattern, is very old and has found expression in various forms at different times and places. The reasons for its persistence and vitality are numerous, but two very general considerations may be identified as having exercised a fairly continuous influence. First, it has often been supposed that, if the belief in an overall pattern is abandoned, one is obliged to acquiesce in the view that the historical process consists of no more than an arbitrary succession of occurrences, a mere agglomeration or patchwork of random incidents and episodes. But such a view (it has been contended) cannot be seriously entertained, if only because it conflicts with the basic demand for system and order that underlies and governs all rational enquiry, all meaningful thought about the world. Second, it has frequently been felt that to refuse to allow that history is finally intelligible in the required manner implies a skepticism concerning the value of human life and existence that constitutes an affront to the dignity of human nature. The 18th-century German philosopher Immanuel Kant, for example, spoke of the "repugnance" that is inevitably experienced if the past is viewed

The necessity for a meaningful view of history

> as if the whole web of human history were woven out of folly and childish vanity and the frenzy of destruction, so that one hardly knows in the end what idea to form of our race, for all that it is so proud of its prerogatives.

In more recent times, a comparable attitude is discernible beneath Arnold Toynbee's uncompromising repudiation of the idea that history is "a chaotic, disorderly, fortuitous flux, in which there is no pattern or rhythm of any kind to be discerned." Thus, it has been the object of a long line of theorists, representative of widely divergent outlooks, to demonstrate that such pessimism is unjustified and that the historical process can, when appropriately viewed, be seen to be both rationally and morally acceptable.

Theological origins. Western speculation concerning the meaning of history derived in the first instance chiefly from theological sources. The belief that history conforms to a linear development in which the influence of providential wisdom can be discerned, rather than to a recurrent cyclical movement of the kind implicit in much Greco-Roman thought, was already becoming prevalent early in the Christian Era. Traces of this approach are to be found in the conception of the past set forth by St. Augustine in his *City of God* and elsewhere; it is, for example, compared on one occasion to "the great melody of some ineffable composer," its parts being "the dispensations suitable to each different period." Yet the cautious subtlety of Augustine's suggestions and the crucial distinction he drew between sacred and secular history make it important not to confuse his carefully qualified doctrines with the cruder positions advanced by some of his self-proclaimed successors. This applies, *par excellence*, to the work of the most renowned and thorough of these, Jacques-Bénigne Bossuet. Written 1,000 years after Augustine's death, Bossuet's *Discours sur l'histoire universelle* (1681; "Discourse on Universal History") is imbued throughout with a naïve confidence that the entire course of history owes its pervasive character to the contrivance of a "higher wisdom." In the eyes of Bossuet, to grasp and understand the great procession of empires and religions was "to comprehend in one's mind all that is great in human affairs and have the key to the history of the universe." For the rise and fall of states and creeds depended in the end upon the secret orders of Providence, the latter being the source of that manifest historical justice and retribution to which, on nearly every page, the annals of the past bore clear and unmistakable witness. Bossuet's vast survey was, in fact, the last major contribution to its genre. Though it made a considerable impression when it was first published, it appeared just before the discoveries of Sir Isaac Newton effected a massive transformation of the European outlook, and the book's impact was short-lived. Thus, the development of historical speculation in the 18th century was generally marked by a tendency to reject theological and providential interpretations in favour of an approach more closely aligned, in method and aim, to that adopted by natural scientists in their investigations of the physical world.

Augustine and Bossuet

Secular approaches: the Enlightenment and beyond. For many Enlightenment and post-Enlightenment thinkers, the project of establishing a science of history and society, comprising hypotheses and laws of an explanatory power analogous to that attained by theories in the physical sciences, acquired an almost obsessive importance. The age of religious and metaphysical conjectures concerning the destiny of human affairs had, in their opinion, come to a close. The task that now presented itself was one of constructing, upon the basis of hard observable facts, interpretations that would not only rescue the human studies from ignorance, uncertainty, and primitive superstition but also put into men's hands an instrument for predicting and controlling their fate. Thus, the idea of creating a universally valid social science, capable of accounting for the phenomena of history in terms of causal principles comparable to those employed in the natural sphere, came to be linked with the promotion of reformist and revolutionary ideals. Men such as Condillac and Condorcet in the 18th century and Henri de Saint-Simon, Auguste Comte, John Stuart Mill, and Henry Thomas Buckle in the 19th century all believed that it was feasible to apply scientific procedures to the study of human development. But equally—though in widely different ways—they were men deeply concerned with practical objects and committed to changing existing institutions and ways of life. To these men, theory was complementary to practice; knowledge was power.

Yet even in the 19th century, when speculation of this type was at its height, there were informed skeptics—Joseph de Maistre and Arthur Schopenhauer, for example, and later the great Swiss historian Jacob Burckhardt—who challenged the optimistic and rationalistic presuppositions on which it was founded. It was pointed out that notions such as that of the perfectibility of man or of the existence of some foreseeable goal toward which the course of events was inexorably leading were not empirically established truths but mere articles of faith; in subscribing to them, historical theorists often appeared to be tacitly importing into their allegedly scientific interpretations teleological conceptions of a kind that it had been their declared intention to banish forever from social enquiry. These objections have been repeated and amplified by 20th-century critics such as Karl Popper, who have also maintained that the theorists in question were, in any case, working with an unacceptably crude notion of scientific reasoning and that their high-sounding generalities conspicuously failed to measure up to the requisite standards of conceptual precision and observational testability.

Nineteenth-century skepticism

Although such strictures have considerable force, they should not obscure the significant contribution that had been made toward extending human knowledge and understanding. The tendency, for example, to insist upon the relevance of scientific modes of procedure to the areas of historical and social investigation at least achieved the salutary effect of throwing into relief the inadequacy of previous work in these domains; moreover, it indirectly brought to the fore the entire question of the status of history as a legitimate form of thought. For, if history should prove resistant to attempts to assimilate it to other accredited branches of enquiry, it would be necessary to show why this was so and to exhibit those features of historical thinking that lent it its distinctive and irreducible character.

The new science: Vico and Herder. Among the 18th-century theorists, two writers can indeed be picked out who—while remaining firmly within the speculative tradition—at the same time possessed sufficient genius and prescience to realize that the solution to the problem of establishing history as a reputable discipline might be found by pursuing a course different from one modelled upon the methodology of the natural sciences. Partly because of the obscure and scholastic manner in which it was written, Giambattista Vico's *Scienza nuova* (3rd ed. 1744; "New Science") was a work whose importance remained for a long time wholly unrecognized, and it is only fairly recently that its significance and originality have been fully appreciated.

Central to the book is the contention that the kind of knowledge that men can achieve of their own actions, creations, and institutions is of a radically different type from the knowledge that is acquired by the observation and investigation of the nonhuman or natural world; knowledge of the former variety is, moreover, held to be in principle superior to that of the latter. For, in Vico's opinion, in order truly to know something it is necessary in some sense to have made it: it followed that, whereas the reality studied by the physical scientist is the creation of God and therefore only properly known by God, the "world of nations" that forms the subject matter of history is the creation of men and is therefore something that men can "hope to know." Thus, Vico was led to stress the differences rather than the analogies between historical and other forms of enquiry; in particular, he emphasized the need for the historian to enter imaginatively into the spirit of past ages, re-creating the outlooks and attitudes that informed them as opposed to seeking to impose upon them inappropriate or falsifying interpretations—"pseudomyths"—that derived from the cultural ethos of his own time. Vico propounded a cyclical theory of human history, according to which "nations" or societies pass through determinate stages, and he combined this with the idea that a providential principle is in some man-

Vico's notion of unique historical knowledge

ner immanent within the various forms of life that men construct. He employed such conceptions, however, in a fashion that underlined man's nature as a historical being, whose powers and capacities do not conform to a fixed or static pattern but are necessarily subject to change and development in the course of time.

In a similar vein, the German writer Johann Gottfried von Herder, in his influential *Ideen zur Philosophie der Geschichte der Menschheit* (4 vol., 1784–91; Eng. trans., *Outlines of a Philosophy of the History of Man*, 1800), implied that it was vital to view human actions and achievements from a standpoint that took proper account of "time, place and national character"—in other words, cultural milieu and the inevitable limits imposed by historical situation and circumstance. In its general direction, Herder's historical thought reflected the Enlightenment preconceptions of man as a progressive being. Herder's chief importance lies, however, in his insistence upon the misconceptions involved in treating the products of past thought and action as if they were the manifestations of an unchanging human consciousness and as if they could be explained by reference to abstract laws eternally valid for men everywhere. According to Herder, such an approach failed to recognize the complex influences that act upon human beings as members of particular historical societies; each of these societies possessed its unique life-style, subtly but inescapably determining the mentalities of those born within its confines in a manner that rendered futile all attempts to reduce human propensities and needs to the terms of some simple set of abstract formulas.

Herder's rejection of static laws

Many of Vico's and Herder's ideas appear familiar today, but it is easy to forget that the emergence of what has come to be known as the "historical sense" is a comparatively recent phenomenon, one that represents a genuine revolution in European thought. It is largely because of this revolution that social and political theories of the kind elaborated by men such as Thomas Hobbes and Benedict de Spinoza in the 17th century seem oddly artificial to 20th-century eyes, so remote are the categories in which they sought to explain human life and behaviour from those that have subsequently found acceptance.

History as a process of dialectical change: Hegel and Marx. The suggestion that there is something essentially mistaken in the endeavour to comprehend the course of history "naturalistically" and within an explanatory framework deriving from scientific paradigms was powerfully reinforced by conceptions stemming from the development of German Idealism in the 19th century. Hegel's "philosophy of the spirit" made its appearance upon the intellectual scene contemporaneously with Saint-Simonian and Comtean Positivism, rivalling the latter in scope and influence and bringing with it its own highly distinctive theory of historical evolution and change. Hegel's stress upon the "organic" nature of social wholes and the incommensurability of different historical epochs owed evident debts to Herderian ideas, but he set these within an overall view that pictured the movement of history in dynamic terms. Regularities and recurrences of the sort that typically manifest themselves in the realm of nature are foreign, Hegel maintained, to the sphere of mind or spirit, which was characterized instead as involving a continual drive toward self-transcendence and the removal of limitations upon thought and action. Man was not to be conceived according to the mechanistic models of 18th-century Materialism; essentially he was free, but the freedom that constituted his nature could only achieve fulfillment through a process of struggle and of overcoming obstacles that were themselves the expression of his own activity; it was in this sense that Hegel claimed that spirit was "at war with itself"—"it has to overcome itself as its most formidable obstacle" (*Lectures on the Philosophy of History*). In concrete terms, this meant that historical advance did not proceed through a series of smooth transitions. Once the potentialities of a particular society had been realized in the creation of a certain mode of life; its historical role was over; its members became aware of its inadequacies, and the laws and institutions they had previously accepted unquestioningly were now experienced as fetters, inhibiting further development and no longer reflecting their deepest aspirations. Thus, each phase of the historical process could be said to contain the seeds of its own destruction and to "negate" itself; the consequence was the emergence of a fresh society, representing another stage in a progression whose final outcome was the formation of a rationally ordered community with which each citizen could consciously identify himself and in which there would therefore no longer exist any sense of alienation or constraint. Somewhat curiously, the type of community Hegel envisaged as exemplifying this satisfactory state of affairs bore a striking resemblance to the Prussian monarchy of his own time.

History as a dynamic process

The notion that history conforms to a "dialectical" pattern, according to which contradictions generated at one level are overcome or transcended at the next, was incorporated—though in a radically new form—in the theory of social change propounded by Karl Marx. Like Hegel, Marx adopted a "directional" view of history; but, whereas Hegel had tended to exhibit it as representing the unfolding in time of an inner spiritual principle, Marx looked elsewhere for the ultimate determinants of its course and character. Man, according to Marx, was a creative being, situated in a material world that stood before him as an objective reality and provided the field for his activities; this primitive truth, which had been obscured by Hegel's mystifying abstractions, afforded the key to a proper understanding of history as a process finally governed by the changing methods whereby men sought to derive from the natural environment the means of their subsistence and the satisfaction of their evolving wants and needs. The productive relations in which men stood to one another, resulting in such phenomena as the division of labour and the appearance of economically determined classes, were the factors fundamental to historical movement. What he termed the superstructure of society—which covered such things as political institutions and systems of law, ethics, and religion—was in the last analysis dependent upon the shape taken by the "material production" and the "material intercourse" of human beings in their struggle to master nature: "it is not the consciousness of men that determines their being, but, on the contrary, their social being that determines their consciousness." Hence, the inner dynamic of history was held to lie in conflicts arising from changes in the means of production and occurring when modes of social organization and control, adapted to the development of the productive forces at one stage, became impediments to it at another; they were to be resolved, furthermore, not by abstract thought but by concrete action. Thus, the Hegelian conception of spirit as involved in a relentless struggle with itself and with what it had created underwent a revolutionary transformation, explosive in its implications.

Economic determinism

Marx's interpretation of the historical process, with its stress upon necessity and the operation of ineluctable laws, has often been portrayed by its proponents as being scientific in character. It has, however, more than one aspect, and it would be an error to identify its underlying methodology with that associated with Comtean Positivism. Generally speaking, the basic categories within which it was framed derived from a theory of human nature that had more in common with the postulates of German romantic thought than with those of British and French Empiricism: to this extent, the logical structure Marx sought to impose upon the data of history belonged to a tradition that stressed the differences rather than the resemblances between the human and the natural world.

Twentieth-century systems. The tendency to detect in history the presence of large-scale patterns and comprehensive uniformities continued into the 20th century in the work of a number of writers, most notably Oswald Spengler and Arnold Toynbee. Spengler's *Decline of the West* (originally published in German, 1918–22), wherein the history of mankind was presented in terms of biologically conceived cultures whose careers conformed to a predetermined course of growth and decay, was

widely acclaimed during the years of disillusionment that followed World War I; and a somewhat similar reception was given to Toynbee's massive *Study of History* (1934–59) immediately after World War II. Toynbee, like Spengler, undertook a comparative study of civilizations, thereby repudiating attempts to treat the past as if it exhibited a single linear progression: at the same time, he diverged from Spengler in suggesting that current Western society might not after all be necessarily doomed to extinction and in tempering a predominantly deterministic mode of thought with reservations that allowed a place for human free will and the possibility of divine intervention. Yet, as some of his critics were quick to point out, such qualifications were not easy to reconcile with his original insistence upon the need to adopt "a scientific approach to human affairs"; nor was it clear that his own use of inductive methods to establish the laws governing the development of civilizations was above logical suspicion or reproach. Toynbee's experiment might have been impressive as an individual achievement; nevertheless, with the multiplication of objections and in a theoretical climate that had become skeptical of speculative system-building of any kind, the very feasibility of engaging upon a project of the type he had undertaken came to be seriously questioned. It was felt increasingly that philosophy of history in the traditional sense—resting largely upon uncriticized assumptions concerning the nature of historical enquiry and its relations with other disciplines—had reached something of an impasse; if history was still to be treated as a proper subject for philosophical examination, it must be along lines quite different from those previously pursued.

Toynbee's system

ANALYTICAL PROBLEMS

The concept of history. The task of trying to delineate the specific character of historical knowledge and understanding, rather than of seeking to construct vast speculative schemes in the earlier manner, first began to attract the attention of philosophers toward the end of the 19th century. To such thinkers as Wilhelm Dilthey and Benedetto Croce, the claim that, in the absence of some all-embracing system of a teleological or quasi-scientific kind, the course of history could be regarded as constituting nothing better than a meaningless chaos appeared to be totally unacceptable. History is intelligible, they believed, in the sense that historians make it so; moreover, this was the only type of intelligibility it was either necessary or legitimate to demand. What could reasonably be looked for was a clearer and deeper insight into the conditions that render historical knowledge possible, an elucidation of the presuppositions upon which historical enquiry is founded and of the principles according to which it proceeds. It was with such an investigation in mind that R.G. Collingwood, a British philosopher who owed much to Crocean ideas, wrote in his *Autobiography* that "the chief business of twentieth-century philosophy is to reckon with twentieth-century history." By contending that the philosopher should eschew the grandiose ambition of providing a synoptic vision of the entire historical process and concern himself, instead, with the articulation and justification of existing historical procedures, Collingwood and his continental precursors made, in effect, a crucial contribution toward setting philosophy of history on a new path. Their proposals were, moreover, given additional impetus by the widespread acceptance of analytical approaches in other branches of philosophy. In consequence, contemporary thinkers have tended to focus attention upon the explication of concepts and terms that perform a key role in historical thought and description as these are actually carried on: among other things, they have been led into discussing the ways in which historians typically divide up and classify the past, the manner in which they argue for and substantiate their interpretations, and the logical structure of the explanations they are accustomed to offer.

The work of R.G. Collingwood

Explanation and understanding. Both Croce and Collingwood, in their criticisms of earlier theorists, were especially anxious to expose what they believed to be recurrent and fundamental misconceptions regarding the method and subject matter of history: central to these was the assumption that historical occurrences could be subsumed under, and explained in terms of, universal laws of the sort that played an essential part in scientific interpretations of inanimate nature. This assumption was, in their opinion, a gross error. As Collingwood put it, the moment had arrived for history to be released from "its state of pupilage to natural science." With this in mind, he went on to develop an account of historical understanding according to which the historian explains events by exhibiting them as the expressions of past thinking on the part of self-conscious purposive agents—thinking that the historian must imaginatively reconstruct or re-enact in his own mind—rather than by showing the events to be instances of general uniformities or regularities that are established by induction. In propounding this view—which Croce, though he formulated it less clearly and precisely, basically shared—Collingwood set in motion a controversy concerning knowledge and explanation in history that has been central to much subsequent discussion. As Collingwood himself was fully aware, a position similar to his own had been originally advanced (though in a very different context) by Vico; and it is indeed noteworthy that the general division, evident at the speculative level, between those who wished to comprehend historical phenomena in ways suggested by the physical sciences and those who, by contrast, argued for an altogether distinct pattern of interpretation has tended, in recent times, to re-emerge at the level of methodological and conceptual analysis.

Thus, on one side of the dispute, there have been ranged philosophers who have taken their stand upon what has been called "the unity of science" and who have insisted that the categories and procedures appropriate to the human studies do not enjoy a unique or privileged status that somehow sets them apart from those characteristic of systematic empirical enquiry in other domains. In a classical 18th-century discussion, David Hume had argued that, if two events were said to be causally related, this could only be in the sense that they instantiated certain regularities of succession that had been repeatedly observed to hold between such events in the past: to presume otherwise was to fall back upon an unacceptable belief in "intuitable" connections that had no warrant either in reason or experience. This doctrine may be said to have been given more rigorous expression among Positivist philosophers of the present century in the shape of what is variously known as the "deductive-nomological" or "covering law" theory of explanation; as originally applied to history by Carl Hempel, it amounted to the claim that explaining a given historical occurrence in terms of some other event or set of events necessarily involves an appeal, which need not be more than tacit, to laws or general propositions correlating events of the type to be explained with those of the kind cited as its causes or conditions. Although the proposed analysis has received a variety of different formulations, each designed to meet specific objections that have been raised against it, its adherents have not wavered from the conviction that some such account must be in principle correct if explanations in history are to be open to rational assessment of the sort properly demanded within any legitimate branch of empirical investigation. It is for this reason, together with others, that they have been strongly opposed to *Verstehen*, or "empathy," theories of historical knowledge, regarding the contention that historical understanding presupposes an allegedly direct identification with the mental processes of past human agents as representing at best a heuristic recommendation of doubtful utility, at worst an obscurantist doctrine that transparently fails to provide an objective criterion whereby divergent historical interpretations can be evaluated.

The unity of science approach

Resistance to the Positivist approach has come from more than one direction. To a number of practicing historians, for instance, the account offered has appeared implausible inasmuch as it overlooks the "irreducible particularity" of historical occurrences and because it postulates an unjustifiably high degree of reliance upon the presence of discernible uniformities in the sphere of

human affairs. So far as philosophers are concerned, dissatisfaction has been voiced both by those to whom the Croce–Collingwood notion of historical thinking as the "re-enactment of past experience" has seemed to contain an important element of truth and also by those followers of Ludwig Wittgenstein who have been impressed by the skepticism concerning the adequacy of scientific models apparent in his later discussions of mental concepts. A leading representative of the former group, W.H. Dray, not only constructed a series of arguments to demonstrate the deficiencies of the covering-law theory but further proposed an alternative conception of "rational explanation," which—it was suggested—fitted many of the familiar ways whereby historians seek to render the past intelligible. Thus, Dray has maintained that the function of much historical explanation consists of showing the actions of historical persons to have been "appropriate" when viewed within the perspective of their specific beliefs, aims, and principles: it was this consideration, he claims, that was uppermost in the minds of theorists who were concerned to stress the part played by imaginative or empathetic understanding in historical reconstruction, their point being primarily a "logical" one and not necessarily carrying any of the dubious epistemological implications attacked by Positivist critics. From a different standpoint, Anglo-U.S. writers influenced by Wittgenstein have challenged the entire assumption that explanations involving the notions of human intention and purpose are susceptible to a Humean pattern of causal analysis; they have also (for example, in the work of Peter Winch) stressed the extent to which historical descriptions of past behaviour require to be framed in terms the agents themselves would have recognized as giving meaning to their activities, terms embodying references to ideas and conventions that defined the social reality in which they participated.

The criterion of appropriateness

Objectivity and evaluation. Fundamental issues concerning the status of historical enquiry of the kind just mentioned have arisen in another crucial area of discussion, centring upon the question of whether—and, if so, in what sense—history can be said to be an objective discipline. Some modern philosophers have inclined to the view that the entirely general problem of whether history is objective cannot sensibly be raised; legitimate questions regarding objectivity are only in place where some particular piece of historical work is under consideration, and in that case there are accepted standards available, involving such matters as documentation and accuracy, by which they can be settled. To others, however, things have not seemed so clear, and they have drawn attention to the doubts that may be felt when history is compared with different branches of investigation, such as chemistry or biology: by contrast with such enquiries, the historian's procedure, including the manner in which he conceptualizes his data and the principles of argument he employs, may appear to be governed by subjective or culturally determined predilections that are essentially contestable and, therefore, out of place in a supposedly reputable form of knowledge. One topic that has been recurrently examined in this connection has been the role of evaluation (specifically, of moral evaluation) in historical writing—a subject, incidentally, about which historians themselves are apt to exhibit a certain uneasiness. Nevertheless, recommendations to the effect that value judgment can and should be totally excluded from history and, indeed, from the social studies as a whole have met with a mixed philosophical reception. Among Positivists and Logical Empiricists, traditionally skeptical of the rationality of value judgments and anxious in any case to reduce the differences between the human and the natural sciences, they have found some measure of support. But that has been by no means a general response. Thus, objectors have pointed out that the language the historian customarily uses, adapted as it is to the assessment and appraisal of human motives and characteristics, makes some degree of evaluation unavoidable; they argue that, even if the possibility of a drastically revised historical vocabulary allows the ideal of a *wertfrei*, or objective history, to be theoretically conceivable, such an ideal can scarcely be seriously entertained as a realizable practical goal. These considerations have been reinforced by the further point that every historian, insofar as he has to select from the mass of material confronting him, is necessarily committed to forming judgments ascribing relative importance and significance; such attributions cannot, however, be simply read off from the facts and must, rather, be said to depend upon the prior acceptance of certain critical standards. To this extent, then, one is required to acknowledge the presence in historical writing of an ineliminable evaluative component, which is liable to obtrude itself into even so "objective" a field as that of causal analysis: it is notorious that disputes between historians as to the "true" causes of occurrences such as wars or revolutions often appear to resist resolution at a purely empirical level, and it has been persuasively maintained by some philosophers that the basic grounds for such disputes may often be traced back to the historian's adherence to a moral or political standpoint not shared by his opponent.

The question of value judgments

Conclusions. Although the topics discussed above have occupied a central position in 20th-century critical discussion, they represent only a sample of the issues with which analytical philosophers of history have been concerned: other problems that have attracted attention have included the role of freedom and responsibility of historical agents, the nature and description of historical events, and the role of narrative in history. Here, as elsewhere, the approach adopted has often produced results of considerable interest, throwing a revealing light on features of historical enquiry that are easily missed or ignored by theorists in the grip of some powerful dogma or ideology. Even so, it has perhaps been accompanied by a too ready acquiescence in the view that history is "in order as it is," the philosopher's function being confined to offering a purely descriptive elucidation of typical modes of historical thought and argument. In accepting this conception of their role, analytical philosophers of history have no doubt been partly, and understandably, influenced by a desire to avoid emulating the heady ambitions of their speculative predecessors. Yet, normative questions regarding the validity or adequacy of established procedures within any domain can always be legitimately raised; in the case of history, there seems to be no compelling reason to assume that such problems necessarily lie beyond the scope of philosophical criticism and appraisal.

BIBLIOGRAPHY

Historical and critical studies: ROBIN G. COLLINGWOOD, *The Idea of History* (1946), a classical contribution to the critical theory of history; KARL R. POPPER, *The Poverty of Historicism* (1957), an influential critique of types of historical speculation; WILLIAM H. WALSH, *An Introduction to Philosophy of History* (1951), a lucid, general account; PATRICK L. GARDINER, *The Nature of Historical Explanation*, 3rd ed. (1967); WILLIAM H. DRAY, *Laws and Explanation in History* (1957); PETER WINCH, *The Idea of a Social Science and Its Relation to Philosophy* (1958); and MORTON G. WHITE, *Foundations of Historical Knowledge* (1965), four analytical discussions relating to historical knowledge and understanding; FRANK E. MANUEL, *Shapes of Philosophical History* (1965), a brief but reliable survey of the development of speculative theories.

Anthologies: PATRICK L. GARDINER (ed.), *Theories of History* (1959), readings from both speculative and critical works; WILLIAM H. DRAY (ed.), *Philosophical Analysis and History* (1966), essays by modern authors.

(P.L.G.)

Hitler, Adolf

The dictator of Nazi Germany, Adolf Hitler, was born on April 20, 1889, at Braunau am Inn, Austria-Hungary. His father, Alois (born 1837), was illegitimate and for a time bore his mother's name, Schicklgruber, but by 1876 he had established his claim to the surname Hitler. Adolf never used any other name, and the name Schicklgruber was only revived by his political opponents in Germany and Austria in the 1930s.

Early life. Adolf Hitler spent most of his childhood in the neighbourhood of Linz, the capital of Upper Austria, after his father's retirement from the Habsburg customs



Hitler reviewing troops on the eastern front, 1939.
Heinrich Hoffmann, Munich

service. Alois Hitler died in 1903 but left an adequate pension and savings to support his wife and children. Adolf received a secondary education and, although he had a poor record at school and failed to secure the usual certificate, did not leave until he was 16 (1905). There followed two idle years in Linz, when he indulged in grandiose dreams of becoming an artist without taking any steps to prepare for earning his living. His mother was overindulgent to her willful son, and even after her death in 1908 he continued to draw a small allowance with which at first he maintained himself in Vienna. His ambition was to become an art student, but he twice failed to secure entry to the Academy of Fine Arts. For some years he lived a lonely and isolated life, earning a precarious livelihood by painting postcards and advertisements and drifting from one municipal lodginghouse to another.

Hitler already showed traits that characterized his later life: inability to establish ordinary human relationships; intolerance and hatred both of the established bourgeois world and of non-German peoples, especially the Jews; a tendency toward passionate, denunciatory outbursts; readiness to live in a world of fantasy and so to escape his poverty and failure.

Military service in World War I

In 1913 Hitler moved to Munich. Temporarily recalled to Austria to be examined for military service (February 1914), he was rejected as unfit; but when World War I broke out he volunteered for the German army and joined the 16th Bavarian Reserve Infantry Regiment. He served throughout the war, was wounded in October 1916, and was gassed two years later. He was still hospitalized when the war ended. Except when hospitalized, he was continuously in the front line as a headquarters runner; his bravery in action was rewarded with the Iron Cross, Second Class, in December 1914, and the Iron Cross, First Class (a rare decoration for a corporal), in August 1918. He greeted the war with enthusiasm, as a great relief from the frustration and aimlessness of his civilian life. He found comradeship, discipline, and participation in conflict intensely satisfying and was confirmed in his belief in authoritarianism, inequality, and the heroic virtues of war.

Rise to power. Discharged from the hospital in the atmosphere of confusion that followed the German defeat, Hitler determined to take up political work in order to destroy a peace settlement that he denounced as intolerable. He remained on the roster of his regiment until April 1920 and as an army political agent joined the tiny Germany Workers' Party in Munich (September 1919).

In 1920 he was put in charge of the party's propaganda and left the army to devote his time to building up the party, which in that year was renamed the Nationalsozialistische Deutsche Arbeiterpartei (of which Nazi was an abbreviation). Conditions were ripe for the development of such a party. Resentment at the loss of the war and the peace terms added to economic chaos brought widespread discontent. This was sharpened in Bavaria, where Hitler lived throughout the 1920s, by traditional separatism and dislike of the republican government in Berlin. In March 1920 a coup d'etat by the army established a strong right-wing government. Munich became the gathering place for dissatisfied former servicemen and members of the Freikorps, which had been organized in 1918–19 from units of the German army unwilling to return to civilian life, and for political plotters against the republic. Many of these joined the Nazi Party. Foremost among them was Ernst Röhm, a member of the staff of the district army command, who had actually joined the German Workers' Party before Hitler and who was of great help in furthering his schemes for developing it into an instrument of power. It was he who recruited the "strong arm" squads used by Hitler to protect party meetings, to attack Socialists and Communists, and to exploit violence for the impression of strength it gave. In 1921 these were formally organized under Röhm into a private party army, the SA (Sturmabteilung). Röhm was also able to ensure the protection of the Bavarian government, which depended on the local army command for the maintenance of order and which tacitly accepted his breaches of law and his policy of intimidation.

Although conditions were thus favourable to the growth of the party, only Hitler was sufficiently astute to take full advantage of them. When he joined the party he found it small, ineffective, committed to a program of nationalist and socialist principles but uncertain of its aims and divided in its leadership. He accepted its program but regarded it only as a means to an end—political power. His propaganda methods and his personal arrogance caused friction with the other members of the committee, which was resolved when Hitler countered their attempts to curb his freedom by offering his resignation. Aware that the future of the party depended on his power to organize publicity and to acquire funds, they were forced to give in, and in July 1921 he became president with unlimited powers. From the first he set out to create a mass movement, whose mystique and force would be sufficient to bind its members in loyalty to him. He engaged in unrelenting propaganda through the party newspaper, the *Völkischer Beobachter* ("Racist Observer," acquired in 1920), and through a succession of meetings, rapidly growing from audiences of a handful to thousands, where he developed his unique talent for magnetism and mass leadership. At the same time, he gathered around him several of the Nazi leaders who later became infamous—Alfred Rosenberg, Rudolf Hess, Hermann Göring, and Julius Streicher.

The Munich *Putsch*

The climax in this rapid growth of the Nazi party in Bavaria came in an attempt to seize power in the Munich (Beer Hall) *Putsch* of November 1923, when Hitler and Gen. Erich Ludendorff took advantage of the prevailing lawlessness and opposition to the Weimar Republic to force the leaders of the *Land* government and the local Reichswehr commander to proclaim a national revolution. When released, however, they rescinded the proclamation. When placed on trial, Hitler, although his part in the *Putsch* had been far from glorious, characteristically took advantage of the immense publicity afforded to him. He also drew a vital lesson from the *Putsch*—that the movement must achieve power by legal means. He was sentenced to prison for five years, but served only

nine months, and that in comfort at Landsberg. He used the time to prepare the first volume of *Mein Kampf*.

Mein Kampf

Hitler's ideas included little that cannot be traced to earlier writers or to the commonly accepted shibboleths of Viennese right-wing radicalism in his youth. He regarded inequality between races and individuals as part of an unchangeable natural order and exalted the "Aryan race" as the sole creative element of mankind. The natural unit of mankind was the *Volk*, of which the German was the greatest; and the state only existed to serve the *Volk*—a mission that the Weimar Republic betrayed. All morality and truth was judged by this criterion: whether it was in accordance with the interest and preservation of the *Volk*. For this reason democratic government stood doubly condemned. It assumed an equality within the *Volk* that did not in fact exist, and it supposed that what was in the interests of the *Volk* could be decided by discussion and voting. In fact the unity of the *Volk* found its incarnation in the *Führer*, endowed with absolute authority. Below the *Führer* the party (which Hitler often called the "movement" to distinguish it from democratic parties) was drawn from the best elements of the *Volk* and was in turn its safeguard.

The greatest enemy of Nazism was not, in Hitler's view, liberal democracy, which was already on the verge of collapse. It was rather the rival *Weltanschauung*, Marxism (which for him embraced Social Democracy as well as Communism), with its insistence on internationalism and class conflict. Behind Marxism he saw the greatest enemy of all, the Jew, who was for Hitler the very incarnation of evil, a mythical figure into which he projected all that he feared and hated.

During Hitler's absence in prison the Nazi Party disintegrated through internal dissension. In the task of reconstruction after his release, he faced difficulties that had not existed before 1923. Economic stability had been achieved by currency reform and the Dawes Plan; the republic had become more respectable. Hitler was forbidden to make speeches, first in Bavaria, then in many other German states (these prohibitions remained in force until 1927–28). Nevertheless, the party grew slowly in numbers, and in 1926 Hitler successfully established his position against Gregor Strasser, who had built up a rival Nazi movement in north Germany.

Alliance with Alfred Hugenberg

The slump of 1929 opened a new period of economic and political instability. Hitler made an alliance with the Nationalist Alfred Hugenberg in a campaign against the Young Plan. Through it Hitler was able for the first time to reach a nationwide audience with the help of Hugenberg's Nationalist Party organization and the newspapers it controlled. It also enabled him to commend himself as a gifted agitator to the magnates of business and industry who controlled political funds and were anxious to use them to establish a strong right-wing, anti-working-class government. The subsidies he received from the industrialists placed his party on a secure financial footing and enabled him to make effective his emotional appeal to the lower middle class and the unemployed, based on the proclamation of his faith that Germany would awaken from its sufferings to reassert its natural greatness. Like his later intrigues with the conservatives, Hitler's dealings with Hugenberg and the industrialists exemplify his skill in using those who sought to use him.

Mass agitation and unremitting propaganda, set against the failure of the government to achieve any success in internal or external affairs, produced a steadily mounting electoral strength for the Nazis, who became the second largest party in the country, with more than 6,000,000 votes at the 1930 election. Hitler opposed Hindenburg in the presidential election of 1932, capturing 36.8 percent of the votes on the second ballot.

Placed in a very strong position by his unprecedented massfollowing, he took part in a series of intrigues for the favour of the aging president in which the other principal participants were Franz von Papen, Gen. Kurt von Schleicher, Otto Meissner, and Hindenburg's son, Oskar. In spite of a decline in the party's votes in November 1932, he held to the chancellorship as the only office he would accept, and this by constitutional, not revolutionary, methods. Throughout, he showed a unique ability to exploit conditions favourable to success. He created the Hitler myth; he propagated it by every device of mass agitation and with an actor's ability to be absorbed in the role that he created for himself. Yet all the time he remained a shrewd and calculating politician, aware of the weaknesses of his own position, perceiving more quickly than anyone else how a situation could best be turned to his own advantage. In January 1933 he reaped his reward when Hindenburg invited him to be chancellor of Germany, and he took office with the support of Papen and Hugenberg and with Field Marshal Werner von Blomberg as minister of defense.

Hitler's personal life had grown more relaxed and stable with the added comfort that accompanied the success of the party. After his release from prison, he went to live on the Obersalzberg, near Berchtesgaden. His income at this time was derived in a haphazard manner from party funds and from writing in nationalist newspapers. When he became chancellor he accepted the material comforts that followed but remained independent of them. He was indifferent to clothes and food, never smoking or drinking tea, coffee, or alcohol. He continued, even as *Führer*, to rebel against routine or regular work—a characteristic that he ascribed to his artistic temperament.

When he went to live at Berchtesgaden, his half sister Angela Raubal and her two daughters accompanied him. Hitler became devoted to one of them, Geli, but his possessive jealousy drove her to suicide in September 1931. For weeks Hitler was inconsolable. Later Eva Braun, a shop assistant from Munich, became his mistress. Hitler rarely allowed her to come to Berlin or appear in public with him and would not consider marriage on the grounds that it would hamper his career. Eva was a warmhearted girl with no intellectual ability. Her great virtue in Hitler's eyes was her unquestioning loyalty, and in recognition of this he made her his legal wife at the end of his life.

Eva Braun

Dictator: 1933–39. Once in power, Hitler proceeded to establish an absolute dictatorship. He secured the President's assent for new elections on the grounds that a majority in the Reichstag could not, after all, be obtained. The Reichstag fire, on the night of February 27, 1933 (apparently the work of a Dutch Communist, Marinus van der Lubbe), provided an excuse for a decree overriding all guarantees of freedom and for an intensified campaign of violence. In these conditions, when the elections were held (March 5), the Nazis polled 43.9 percent of the votes. The Reichstag assembled in the Potsdam Garrison Church, a theatrical gathering designed by Hitler to show the unity of his own movement with the old conservative Germany, represented by Hindenburg. Two days later an enabling bill, giving full powers to Hitler, was passed in the Reichstag by the combined votes of Nazi, Nationalist, and Centre party deputies (March 23, 1933).

Thus far successful, Hitler had no desire to carry too far a radical revolution. Conciliation was still necessary if he was to succeed to the presidency and retain the support of the army; nor had he ever intended to disappropriate the leaders of industry, provided they served the interests of the Nazi state. Ernst Röhm was the chief protagonist of the "continuing revolution"; he was also, as head of the SA, greatly distrusted by the army. Hitler tried first to secure Röhm's support for his policies by persuasion and by giving him government office but failed to win him over. Göring and Heinrich Himmler were eager to remove Röhm, but Hitler hesitated until the last moment. Finally, on June 29, 1934, he reached his decision. Röhm and his lieutenant Edmund Heines were executed without trial, together with Gregor Strasser, Schleicher, and others. The army leaders, satisfied at seeing the SA broken up, approved Hitler's actions. When Hindenburg died, on August 2, they, together with Papen, assented to the merging of the chancellorship and the presidency—with which went the supreme command of the armed forces of the *Reich*—and officers and men

took an oath of allegiance to Hitler personally. Economic recovery and a reduction in unemployment (coincident with world recovery, but for which Hitler took credit) made the regime more acceptable, and a combination of success and terrorism brought the support of 90 percent of the voters in a plebiscite.

In power, Hitler devoted little attention to the organization and running of the domestic affairs of the Nazi state. Responsible for the broad lines of policy, as well as for the system of terror that upheld the state, he left detailed administration to his subordinates. Each of these exercised arbitrary power in his own sphere, but, by deliberately creating offices and organizations with overlapping authority, Hitler effectively prevented any one of these private empires from ever becoming sufficiently strong to challenge his own absolute authority.

Foreign policy claimed his greater interest. His objectives were laid down in *Mein Kampf*, and Hitler worked toward them with consummate skill. He had early admired the pan-Germanism of the Austrian Georg Ritter von Schönerer, and the reunion of the German peoples was his first ambition. Beyond that, the natural field of expansion lay eastward, in Poland, the Ukraine, and the U.S.S.R.—expansion that would necessarily involve renewal of Germany's historic conflict with the Slav peoples, who would be subordinate in the new order to the Teutonic master race. He regarded Fascist Italy as a natural ally in this crusade against Bolshevism, provided its rivalry with Germany in central Europe could be overcome, and was ready to abandon the Germans of the Tirol to this end. Britain was a possible ally, provided it abandoned its traditional policy of maintaining the balance of power in Europe and limited itself to its interests overseas. France alone in the west was the natural enemy of Germany and must, therefore, be subdued to make expansion eastward possible.

Planning eastward expansion

Before such expansion was possible, it was necessary to remove the restrictions placed on Germany by the Treaty of Versailles. Hitler used all the arts of propaganda to allay the suspicions of the other powers. He posed as the champion of Europe against the scourge of Bolshevism and insisted that he was a man of peace who wished only to remove the inequalities of the Versailles Treaty. Germany withdrew from the Disarmament Conference and from the League of Nations (October 1933), but Hitler hastened to sign a nonaggression treaty with Poland (January 1934). Every repudiation of the treaty was followed by an offer to negotiate a fresh agreement and insistence on the limited nature of Germany's ambitions. Only once did he overreach himself: when the Austrian Nazis, with the connivance of the German embassy, murdered Chancellor Engelbert Dollfuss of Austria and attempted a coup d'etat (July 1934). The attempt failed, and as Mussolini moved troops to the frontier, Hitler disclaimed all responsibility and sacrificed those who had acted with his sanction. In January 1935 a plebiscite in the Saarland returned that territory to Germany, and Hitler took the opportunity to renounce any further claims on France. In March of the same year, he announced the introduction of conscription, and, although this provoked the united opposition of Britain, France, and Italy at the Stresa Conference, his peace propaganda was sufficiently successful to persuade the British to negotiate a naval treaty (June 1935) recognizing Germany's right to rearm. His greatest stroke came in March 1936, when he used the excuse of a pact between France and the Soviet Union to remilitarize the Rhineland—a decision that he took against the advice of his own general staff. Meanwhile, the alliance with Italy, foreseen in *Mein Kampf*, rapidly became a reality as a result of the sanctions imposed by Britain and France against Italy in the Ethiopian war. In October 1936 the Rome–Berlin axis was established; shortly afterward came the Anti-Comintern Pact with Japan; and these two were linked a year later.

By 1937–38 a new stage had been reached. In November 1937 Hitler outlined his plans of future conquest (beginning with Austria and Czechoslovakia) to a secret meeting of his military leaders. He now dispensed with the services of those who were not wholehearted in their acceptance of Nazi dynamism—Hjalmar Schacht, who declared Germany's further rearmament a danger to its economy; Werner von Blomberg and Werner von Fritsch, representatives of the caution of professional soldiers; and Konstantin von Neurath, Hindenburg's appointment at the foreign office.

Austria and Czechoslovakia

In February 1938 Hitler invited the Austrian chancellor, Kurt von Schuschnigg, to Berchtesgaden and forced him to sign an agreement giving the Austrian Nazis virtually a free hand. When Schuschnigg attempted to repudiate the agreement and announced a plebiscite on the question of an *Anschluss* with Germany, Hitler immediately ordered the occupation of Austria by German troops. The enthusiastic reception that Hitler himself received decided him to settle the future of Austria by outright annexation. He returned in triumph to Vienna, the scene of his youthful humiliations and hardships. No resistance was encountered from Britain and France. Hitler had taken special care to secure the support of Italy, and when this was forthcoming proclaimed his undying gratitude to Mussolini.

Having given assurances that the *Anschluss* would not affect Germany's relations with Czechoslovakia, Hitler proceeded at once with his plans against that country. Konrad Henlein, leader of the German minority in Czechoslovakia, was instructed to agitate for impossible demands on the part of the Sudetenland Germans, thereby enabling Hitler to justify the annexation of Czechoslovakia. But the willingness of Britain and France to compel the Czech government to cede the Sudetenland areas to Germany presented Hitler with the choice between substantial gains by peaceful agreement and even greater acquisitions by a spectacular war against Czechoslovakia. Mussolini's intervention appears to have decided him, and he accepted the Munich Agreement on September 30—only to feel resentment immediately afterward at being cheated out of an impressive military conquest.

It was to be expected, therefore, that Hitler would waste no time in provoking an occasion for occupying the whole of Czechoslovakia. This he did by fostering Slovak discontent. On March 16, 1939, from the Hradčany Castle in Prague, he proclaimed the dissolution of the state whose existence he, as an Austrian, had always regarded as unnatural. Immediately afterward the Lithuanian government was forced to cede Memel (Klaipéda), on the northern frontier of East Prussia, to Germany.

Hitler was now ready to advance toward the ultimate objective of *Lebensraum* in the East. Confronted by an uncompromising Poland, guaranteed by Britain and France, he strengthened the alliance with Italy (the "Pact of Steel," May 1939) and negotiated a nonagression pact with the Soviet Union, signed on August 24—just within the deadline set for an attack on Poland before the winter. He still disclaimed any quarrel with Britain, but to no avail, and the invasion of Poland (September 1) was followed two days later by a British and French declaration of war.

Invasion of Poland

In his foreign policy Hitler combined complete opportunism in means and timing with unwavering pursuit of the objectives laid down in *Mein Kampf*. He showed astonishing skill in judging the mood of the democracies and exploiting their weaknesses—in spite of the fact that he had scarcely set foot outside Austria and Germany and spoke no foreign language. Up to this point every move had been successful—even his anxiety over British and French entry into the war was dispelled by the rapid success of the war in Poland. The result was to convince him more and more of his own infallibility and to induce him to push ahead still faster with his plans for conquest.

World War II. Hitler from the first had assumed direction of the major strategy of the war. When the success of the campaign in Poland failed to lead to the peace negotiations with Britain for which he had hoped, he ordered the army to prepare for an immediate offensive in the west. Bad weather, however, provided the reluctant generals with the opportunity to postpone the western

offensive, and this in turn led to two major changes in planning. The first, on the suggestion of Adm. Erich Raeder, commander in chief of the navy, was Hitler's order to occupy Denmark and Norway in April 1940. Hitler took a close personal interest in the operation, and from this time his intervention in the detail of military operations was to grow steadily greater.

The second was the adoption of Gen. Erich von Manstein's plan for an attack through the Ardennes (opened May 10), instead of through the Low Countries. Against his generals' advice, Hitler held back Gen. Heinz Guderian's tanks south of Dunkirk, thus enabling the British to organize the evacuation of their army. But the campaign as a whole was a brilliant success, and Hitler could claim the major credit for its overall planning. On June 10 Mussolini entered the war on the side of Germany, and at the end of June Hitler avenged the Treaty of Versailles by signing an armistice with France on the site of the Armistice of 1918.

The next step was the subjugation of Britain by aerial bombardment, followed by invasion. But in the summer of 1940, long-term preparations were begun for the invasion of the Soviet Union, and, as the expected surrender of Britain still failed to materialize, the eastern campaign quickly came to dominate Hitler's conception of the grand strategy of the war to the exclusion of everything else. The Soviet Union had occupied eastern Poland and Bessarabia, and Hitler sought to counter any further moves by forcing the governments of Hungary and Romania to accept an agreement that he dictated and by urging the abandonment of Mussolini's plans for the invasion of Greece. Mussolini, however, piqued at being kept in ignorance of Hitler's intentions, invaded Greece; and the lack of success of the Italian armies made it necessary for German forces to come to their aid in the Balkans and North Africa. Hitler's plans were further disrupted by a coup d'etat in Yugoslavia in March 1941, overthrowing the government that had made an agreement with Germany. Regarding this as an insult to Germany and himself, Hitler immediately ordered his armies to subdue Yugoslavia. The campaigns in the Mediterranean theatre, although successful, remained subordinate to the eastern offensive, with which Hitler was so preoccupied that he lost the opportunism and flexibility that he had shown in political affairs. Even when Raeder and Erwin Rommel urged Hitler to destroy the whole British Middle East position by a final blow at Suez, he would spare no more forces from Operation "Barbarossa"—the planned invasion of the Soviet Union.

The attack against the U.S.S.R. was launched on June 22, 1941, with Hitler so confident of success that he refused to provide winter clothing and equipment for his troops. The German Army advanced swiftly into the Soviet Union, but failed to destroy its Russian opponent. Hitler became completely overbearing toward his generals. He disagreed with them about the object of the main attack and he wasted time and strength by failing to concentrate on a single objective and by frequently reversing his own decisions. In December 1941 an unexpected Russian counterattack made it clear that Hitler's hopes of a single campaign would not be realized.

War with the United States

The next day came the Japanese attack on Pearl Harbor. Hitler precipitately declared war on the United States—although the pact with Japan was purely defensive and he had not been informed of the Japanese intentions. Misled by an essentially central European view of world politics, he apparently took no account of the force that a mobilized United States could bring to bear in Europe.

Hitler's conduct throughout 1942 was marked by further errors of judgment—he paid insufficient attention to the Mediterranean and the Atlantic at a time when a relatively small additional effort in those theatres might have been decisive. In the Soviet Union his continued unreadiness to concentrate on a single objective probably forfeited the opportunity to capture Stalingrad while it was still lightly defended.

Meanwhile, he directed Himmler to prepare the ground for the "new order" in Europe. The concentration camps were expanded, and there were added to them extermination camps such as Auschwitz and Mauthausen, as well as mobile extermination squads. The Jews of Germany, Poland, and the Soviet Union were most numerous among the victims; in German-occupied Europe between 4,500,000 and 5,500,000 had been killed by the end of the war as the only solution in Hitler's view of the Jewish "problem." The sufferings of other races were only less when measured in numbers killed. Such barbarism was indiscriminate, even where, as in the Ukraine, Hitler might have encouraged nationalist feelings to his own advantage.

At the end of 1942, defeat at el-Alamein and at Stalingrad brought the turning point in the war, and Hitler's character and way of life began to change. Hitherto, the success that he had imagined had been largely realized, but to preserve the world of fantasy from defeat and failure he isolated himself more and more from reality. Directing operations from his headquarters in the east, he refused to visit bombed cities or to read reports of setbacks; those close to him, especially Martin Bormann, his secretary, took care that only pleasing information reached him; and he became increasingly dependent on his physician, Theodor Morell, and the injections that he supplied. Even so, he had not yet lost the power to react vigorously in the face of misfortune. After the arrest of Mussolini in July 1943 and the Italian armistice, he not only directed the occupation of all important positions held by the Italian army but ordered the kidnapping of Mussolini, with the intention that he should head a new Fascist government. On the eastern front, however, the refusal to withdraw led only to greater losses without any possibility of holding up the Soviet advance. Inevitably, relations with his army commanders grew increasingly strained, the more so with the growing importance given to the SS divisions, directly responsible to Hitler. Meanwhile, the failure of the U-boat campaign and the bombing of Germany made more evident how reduced were the chances of victory.

Assassination attempt

All these factors made more desperate the few soldiers and civilians who were ready to remove Hitler and negotiate a peace. Several attempts were planned in 1943–44; the most nearly successful was made on July 20, 1944, when Col. Claus von Stauffenberg exploded a bomb at a conference at Hitler's headquarters in East Prussia. But Hitler escaped with superficial injuries, and, with few exceptions, those implicated in the plot were executed. The reduction of the army's independence was now made complete, and National Socialist political officers appointed to all military headquarters.

Thereafter, Hitler was increasingly ill and fatigued; but he did not relax or lose control over the Nazi Party or the army, and he continued to exercise an almost hypnotic power over his close subordinates, none of whom was able to wield any independent authority. In December 1944 he moved his headquarters to the west to direct an offensive in the Ardennes for which the last reserves of manpower were mobilized. When it failed, his hopes for victory became ever more visionary, based on the use of new weapons or on the breakup of the grand alliance, especially after the death of Roosevelt. Far from trying to save what could be rescued from defeat, he ordered mass material destruction and condemned his armies to death by refusing to allow surrender.

From January 1945 Hitler never left the chancellery in Berlin or its bunker, abandoning a plan to lead a final resistance in the south as the Soviet forces closed in on Berlin. In a state of extreme nervous exhaustion, prematurely senile if not insane, he at last accepted the inevitability of defeat and thereupon prepared to take his own life, leaving to its fate the country over which he had taken absolute command. Before this, two further acts remained. In the small hours of April 29, he married Eva Braun. Immediately afterward he dictated his political testament, justifying his career and appointing Karl Dönitz as head of the state and Josef Goebbels as chancellor.

Hitler's death

On April 30 he said farewell to Goebbels, Martin Bormann, and the few others remaining, then retired to his suite and either shot or poisoned himself. His wife took poison. In accordance with his instructions, their bodies were burned.

Hitler's success must be attributed to the susceptibility of postwar Germany to his own unique talents as a political leader. His rise to power was not inevitable, and any change in a complex conjunction of circumstances might have relegated him to the obscurity and failure of his youth; yet there was no one who equalled his ability to exploit and shape events to his own ends. The power that he wielded was unprecedented, both in its scope and in the technical resources at its command; but he made no permanent contribution, moral or material, to mankind. His originality and distinctiveness lay in his methods rather than his ideas and purpose, which were shared in whole or in part by millions of people, in Germany and elsewhere. By the time he was defeated, he had broken down the whole structure of the world in which he lived and left behind a Germany and a Europe that have remained divided ever since.

BIBLIOGRAPHY

Biographical studies: ALAN BULLOCK, *Hitler: A Study in Tyranny*, rev. ed. (1964); FRANZ JETZINGER, *Hitlers Jugend* (1956; Eng. trans., *Hitler's Youth*, 1958); BRADLEY F. SMITH, *Adolf Hitler: His Family, Childhood and Youth* (1967); KONRAD HEIDEN, *Adolf Hitler: Das Zeitalter der Verantwortungslosigkeit* (1936; Eng. trans., *Der Fuehrer: Hitler's Rise to Power*, 1944, reprinted 1967), covers the period up to 1934; H.R. TREVOR-ROPER, *The Last Days of Hitler*, 2nd ed. (1950).

Hitler's own words: Hitler's speeches have been collected and edited by MAX DOMARUS, *Reden und Proklamationen, 1932–1945*, 2 vol. (1962–63); a selection of his speeches in English for the years 1922–39 has been edited by NORMAN H. BAYNES, *The Speeches of Adolf Hitler, April 1922–August 1939*, 2 vol. (1942); *Secret Conversations, 1941–44* (Eng. trans. 1953; British title, *Table Talk, 1941–44*, 1953); *Mein Kampf*, 2 vol. (1925–27; Eng. trans. by R. MANNHEIM, 1969).

Reminiscences of Hitler: E.F.S. HANFSTAENGL, *Hitler: The Missing Years* (U.S. title, *Unheard Witness*, 1957), covers the years 1922–34; ALBERT SPEER, *Erinnerungen* (1969; Eng. trans., *Inside the Third Reich*, 1970).

(A.B./W.F.Kn.)

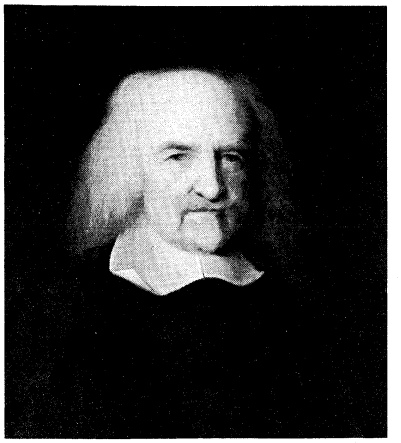
Hobbes, Thomas

Thomas Hobbes, whose long life actively covered most of the 17th century, was one of the greatest of British political philosophers. Paradoxically, he was a political absolutist while qualifying also as one of the founders of Liberalism. His importance lies in his having insisted that the first requirement of political and moral institutions is that they should provide the citizens with security. Thus, Hobbes's starting point was the individual, his rights and need for security—from which he passed to the social contract by which a sovereign is invested with absolute authority. In ethics he was an egoist verging on psychological hedonism (pleasure is the good) and in metaphysics a Materialist, to whom all is body in motion.

Hobbes was born at Westport (now part of Malmesbury), Wiltshire, on April 5, 1588, the son of Thomas Hobbes, vicar of Westport and Charlton. Reports of the Spanish Armada were filling England with alarm at the time of his birth, and Hobbes afterward attributed his own love of peace to the fact that fear and he were twins. His father was "a choleric man," however; and he disappeared after engaging in a brawl at his own church door and abandoned his three children to the care of his brother, a well-to-do glover in Malmesbury.

Early life. At the age of four Hobbes was sent to the church school at Westport, then to a private school, and finally, at 15, to Magdalen Hall, Oxford, where he devoted most of his time to books of travel and the study of maps and charts. On graduating (February 5, 1608), Hobbes became a private tutor to William Cavendish, afterward 2nd earl of Devonshire, and so began a lifelong connection with the Cavendish family. He grew very fond of his pupil, who was only a little younger than himself. In 1610 Hobbes visited France and Italy with his pupil. There he probably found that the Aristotelian philosophy that he had been taught at Oxford was beginning to crumble before the discoveries of Galileo and of Johann Kepler, who formulated the laws of planetary motion. But Hobbes was not yet a philosopher. On returning home he decided to make himself a classical scholar.

Studies and travels



Hobbes, detail of an oil painting by John Michael Wright (1625–1700). In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

At some time between 1621 and 1625, he came into contact with Francis Bacon, known for his philosophy of the inductive method. John Aubrey, a contemporary antiquary, related that Hobbes occasionally acted as Bacon's amanuensis and helped him to render a few of his *Essays* into Latin. But the chief fruit of Hobbes's classical studies was his translation of Thucydides, the Greek historian. Its publication in 1629 was inspired by the troubles of the time, for Hobbes saw in the fate of ancient Athens a salutary warning against democracy. It was also in 1629, the year after the death of the 2nd earl of Devonshire, that Hobbes went abroad again as travelling companion to the son of Sir Gervase Clifton.

The turning point in his intellectual history occurred about the time of this trip, when, in Euclid's *Elements*, he traced the proofs back through proposition after proposition and was thus demonstratively convinced of their truth. As Aubrey wrote, "This made him in love with geometry." The method of proceeding by clear steps from equally indubitable premises seemed thereafter to be the only possible one for science and philosophy.

In 1630 Hobbes was recalled from Paris to teach the young earl of Devonshire, William Cavendish, son of his late patron. In his prose autobiography, Hobbes himself related how he was in a gathering of learned men when the question was asked, "What is sense?" No one appeared to know the answer, but it occurred to Hobbes that, if material things and all of their parts were always at rest or in uniform motion, there could be no distinction of anything and consequently no perception; thus, the cause of all things must lie in diversity of motion. He was therefore driven to geometry to gain insight into the principles of motion. Hobbes laid out these ideas in his first known philosophical work, *A Short Tract on First Principles*, which dates from this period. During a third trip abroad, this time with the younger Cavendish, Hobbes's interest in science and philosophy was stimulated by his contact with the leaders of the new thought in Europe. He became obsessed with the idea of motion. He decided that the basic reality is matter in motion, and he aimed to deduce from this fact, by strict demonstrative arguments (as in geometry), the nature of everything else. He was able to discuss his ideas in Paris with the circle of Marin Mersenne, a learned mathematician and theologian, and, in 1636, with Galileo. He then planned a philosophical trilogy: *De Corpore* (1655; "Concerning Body") was to show that physical phenomena are explicable in terms of motion; *De Homine* (1658; "Concerning Man") was to show what specific bodily mo-

tions are involved in human cognition and appetition; and *De Cive* (1642; "Concerning Citizenship") was to deduce from what had already been established the proper organization of men in society.

Political interests

In 1637 Hobbes returned to England to find the country in the political ferment that preceded the Civil War, and he decided because of this threat to publish the last part of his planned philosophy first. He set out to prove that the royal powers and rights called in question were inseparably annexed to sovereignty, which at that time was admitted to reside in the king. *The Elements of Law, Natural and Politic*, part I on man and part II on citizenship, was circulated in manuscript in 1640. It already embodied Hobbes's characteristic doctrine that men can only live together in peace if they agree to subject themselves to an absolute and undivided sovereign, and it contained most of the political and psychological doctrines for which Hobbes is famous and which reappeared in *De Cive* and *Leviathan* (see below). It differed from his *Leviathan*, however, by stressing that primary democracy is the first form of commonwealth by institution, monarchy and aristocracy arising when the democratic sovereign created by the social contract between individuals annihilates itself by transferring its power absolutely to one man or to a few. Unfortunately, Hobbes antagonized both parties in the current constitutional struggle: the believers in the divine right of kings were irritated by his social contract theory, the Parliamentarians by his advocacy of absolute monarchy.

Exile in Paris. When strife became acute in 1640, Hobbes feared, perhaps unjustifiably, for his safety and retired to Paris, "the first of all that fled." He was soon in contact with later fugitives from England. He rejoined the Mersenne circle, wrote "objections" to the *Meditationes* and the *Dioptrique* of Descartes, and in 1642 published *De Cive*, which expanded the argument of the second part of *The Elements of Law* and concluded with a section on religion that dealt more fully with the relation between church and state. A Christian church and a Christian state, he held, were one and the same body; of that body, the sovereign was the head; he therefore had the right to interpret Scripture, decide religious disputes, and determine the form of public worship. Hobbes spent the next four years working on optics and on *De Corpore*. In 1646 the young prince of Wales, later to become Charles II, sought refuge in Paris, and Hobbes accepted an invitation to instruct him in mathematics. Contact with exiles from England made it increasingly difficult to concentrate on natural philosophy, and he turned once more to political theory. In 1647 he published a second, augmented edition of *De Cive* and, in 1651, an English version. In 1650 the manuscript of *The Elements of Law* was published in two parts, as *Human Nature* and *De Corpore Politico* ("Of the Body Politic").

Contents of the *Leviathan*

Hobbes's masterpiece, however, was the *Leviathan, or the Matter, Form, and Power of a Commonwealth, Ecclesiastical and Civil* (1651). In the first two parts, "Of Man" and "Of Commonwealth," he reworked the ground already covered in the earlier treatises; in the last two, "Of a Christian Commonwealth" and "Of the Kingdom of Darkness," he embarked upon a discussion of Scripture and made a vigorous attack on the attempts of papists and Presbyterians to challenge the right of the sovereign.

Hobbes's reputation as a thinker rests mainly on his contributions to the philosophy of man, in which he propounded an influential egoistic psychology. In moral theory he is generally regarded as a pioneer of the Utilitarian school. He justified obedience to moral rules on a purely secular basis, as the means to "peaceable, social, and comfortable living." Yet he also said that the laws of nature were God's commands.

In his political theory Hobbes first analyzed the conditions necessary for peace and security and then, in his version of the social contract, provided a recipe for constructing an ideal state in which these conditions could be satisfied. His fundamental concept was natural right rather than natural law. It is essentially a right to self-preservation. No man is obliged to act in accordance with the law of nature if he thinks such conduct inimical to his own security. Yet peace cannot be achieved unless the law of nature is generally observed. Hobbes's solution was to give everyone a guarantee of the good behaviour of his fellows by creating a power sufficient to keep them in awe. This power will be created if each individual promises every other individual that he will carry out whatever commands some selected person (or an assembly) shall consider necessary for the peace and defense of all. A sovereign so established may survive even if all the subjects desire to depose it. The sovereign's right will be as absolute as its power; it is responsible only to God. It cannot be unjust to its subjects, since these have authorized its actions. Nor is it bound by any covenant with the people.

By 1651 Charles I was dead and the Royalist cause seemed hopelessly lost; accordingly, at the end of *Leviathan*, Hobbes attempted to define the circumstances under which submission to a new sovereign became legitimate. He had always maintained that a subject had the right to abandon a ruler who could no longer protect him and to transfer his allegiance to one who could; but the statement of this view in *Leviathan* gave serious offense to Prince Charles's advisers, who concluded that Hobbes was trying to curry favour with the new regime in England in order to facilitate his own return. Barred from the exiled court and under suspicion by the French authorities for his attack on the papacy, Hobbes thus found his position in Paris becoming daily more intolerable. At the end of 1651, he returned to England and made his peace with the new regime.

Controversies. Though Hobbes was now 63 years of age, he was to retain his vigour for another quarter of a century. For a time he worked quietly on *De Corpore*, but in 1654 he became involved in one of the controversies that were to play so large a part in his later life. In 1646 he had discussed the problem of free will with John Bramhall, bishop of Londonderry; a written interchange ensued. One of the philosopher's admirers then published an unauthorized copy of Hobbes's answer under the title *Of Liberty and Necessity*. Bramhall, incensed, proceeded to issue the whole correspondence in a work entitled *A Defence of True Liberty from Antecedent and Extrinsical Necessity* (1655), and Hobbes retorted with *The Questions Concerning Liberty, Necessity, and Chance* (1656), in which he reprinted the correspondence with animadversions on the Bishop's replies. In 1658 Bramhall issued *Castigations of Mr. Hobbes, His Last Animadversions*, which contained a charge of atheism, to which Hobbes replied ten years later (published 1682).

Disputes on free will and mechanism

Hobbes had made enemies at Oxford by the publication of *Leviathan*, which attacked the university system as having supported the pope and as still working social mischief by adherence to the old learning. Oxford was therefore quick to avail itself of the opportunities for criticism offered by *De Corpore*, which was published at last in 1655.

Hobbes had been so impressed by Galileo's achievements in mechanics that he sought to explain all phenomena and, indeed, sense itself in terms of the motion of bodies. Thus his a priori mathematical approach to natural philosophy separated him decisively from Francis Bacon, who had advocated an experimental–inductive method. Hobbes's main antagonists, opposed to such a mechanistic metaphysics, were Seth Ward, professor of astronomy, and John Wallis, author of the great treatise *Arithmetica Infinitorum*, both of them much abler mathematicians than Hobbes. He replied to their attacks in *Six Lessons to the Professors of Mathematics in the University of Oxford* (1656). After more rough thrusts on both sides, Hobbes retired to complete his *De Homine* (1658), which consisted for the most part of an elaborate theory of vision, with a brief account of human psychology.

In the spring of 1660 Hobbes published an onslaught on the newfangled methods of mathematical analysis in six dialogues. In *Dialogus Physicus, sive de Natura Aeris* (1661; "Dialogue on Physics, or on the Nature of Air") he fulminated against Robert Boyle and other friends of

Wallis, who were then forming what became (in 1660) the Royal Society, dedicated to experimental research, in opposition to the deductive method of *De Corpore*. Wallis retorted in a scathing and devastating satire, accusing Hobbes, quite unjustly, of having written *Leviathan* in support of the Puritan leader Oliver Cromwell's title and of having deserted his royal master in distress. Hobbes answered these charges in a letter to Wallis, published under the title *Mr. Hobbes Considered in His Loyalty, Religion, Reputation, and Manners* (1662). In this piece, which is of great biographical value, he told his own and Wallis' "little stories during the . . . rebellion" so effectively that Wallis attempted no reply.

After a time Hobbes began a third period of controversial activity, which did not end on his side until his 90th year. His *De Principiis et Ratiocinatione Geometrarum* (1666) was designed to humble the professors of geometry by showing that their works contained much uncertainty and error. *Quadratura Circuli, Cubatio Sphaerae, Duplicatio Cubi* (1669; "The Squaring of the Circle, The Cubing of the Sphere, The Doubling of the Cube") gave Hobbes's solutions to these famous problems—solutions promptly refuted by Wallis. In 1678 appeared his last piece of all, *Decameron Physiologicum* ("Ten Questions of Physiology"), a new set of dialogues on physiological questions.

Favour at court amid charges of atheism

Meanwhile, from the time of the Restoration in 1660, Hobbes enjoyed a new prominence. Charles II, whom he had tutored in mathematics, received Hobbes again into favour. Though Hobbes's presence at court scandalized the bishops and the prim virtue of the Chancellor, the King relished his wit. He even granted Hobbes a pension of £100 a year and had his portrait hung in the royal closet. But "Hobbism" was frequently identified with freethinking and even with atheism because of his attack on the church. His enemies were many and powerful. Though he was pressed by the hostile church party as early as 1662, it was not until 1666, when the House of Commons prepared a bill against atheism and profaneness, that he felt seriously endangered; for the committee to which the bill was referred was instructed to investigate *Leviathan*. Hobbes, then verging upon 80, burned such of his papers as he thought might compromise him and set himself to inquire into the state of the law of heresy.

The results of his investigations appeared in three short dialogues and in a tract entitled *An Historical Narration Concerning Heresy and the Punishment Thereof* (published posthumously in 1680), in which he maintained that since the abolition of the high court of commission there was no court of heresy to which he was amenable and that, in any case, nothing was to be declared heresy but what was at variance with the Nicene Creed—as the doctrine of *Leviathan* was not.

Although Parliament dropped the bill on atheism, Hobbes could never afterward get permission to print anything on subjects relating to human conduct, the King apparently having made it the price of his protection that no fresh provocation should be offered to popular sentiment. The most important of the works thus withheld from publication was the spirited dialogue *Behemoth, The History of the Causes of the Civil Wars of England* (composed about 1668). To the same period probably belongs the unfinished *Dialogue Between a Philosopher and a Student of the Common Laws of England* (1681), a trenchant criticism of the constitutional theory of English government as upheld by Edward Coke. His *Historia Ecclesiastica* (1688), consisting of Latin elegiacs exposing the methods through which ecclesiastics encroached upon the civil power, may also date from this period.

Last years and historical influence. Though he was impugned by enemies at home, no Englishman of the day stood in such high repute abroad as Hobbes, and distinguished foreigners who visited England were always eager to pay their respects to the old man, whose vigour and freshness of intellect remained unquenched.

In his last years Hobbes amused himself by returning to the classical studies of his youth. The autobiography in Latin verse with its playful humour, occasional pathos, and sublime self-complacency was brought forth at the age of 84. In 1675, he produced a translation of the *Odyssey* in rugged English rhymes, with a lively preface, "Concerning the Virtues of an Heroic Poem." A translation of the *Iliad* appeared in the following year. As late as August 1679 he was promising his publisher "somewhat to print in English." He died at Hardwick Hall in Derbyshire on December 4 of that year and was buried in the neighbouring church of Hault Hucknall.

Personal appearance and character

Hobbes was tall and erect in figure. Believing that "a beard did not make a philosopher," he had himself shaved close except for a little tuft under his lip. He lived a temperate life. His favourite sport was tennis, which he continued to play occasionally, even at the age of 75. Socially he was genial and courteous, though in argument he sometimes lost his temper. Cautious in practical matters but intellectually bold in the extreme, he claimed to have read little and boasted that he would have known as little as other men if he had read as much. He is said to have had an illegitimate daughter, for whom he made generous provision.

Hobbes was not only the architect of a grand metaphysical design, but he was also a critical philosopher with a lively interest in language and a keen eye for its snares. Indeed, his account of the sources of absurdity, which provided him with a potent weapon against the Scholastics, gives him some title to be regarded as a forerunner of modern logical analysis.

Hobbes's reputation was soon overshadowed by that of Locke; and when the theory of the social contract went out of fashion, he suffered a period of neglect in Great Britain. But in the 19th century interest in him was revived by the Utilitarians—*i.e.*, by the followers of Jeremy Bentham, a moralist, and in particular by John Austin, a jurist, who made a modified version of Hobbes's doctrine of sovereignty and law the basis of his jurisprudence —and since Austin's day, Hobbes has gradually been accorded recognition as one of the greatest English political thinkers.

MAJOR WORKS

IN LATIN: *Elementorum Philosophiae Sectio Tertia: De Cive* (1642); *Elementorum Philosophiae Sectio Prima: De Corpore* (1655); *Elementorum Philosophiae Sectio Secunda: De Homine* (1658).

IN ENGLISH: *Human Nature; or, The Fundamental Elements of Policie* (1650); *De Corpore Politico; or, The Elements of Law, Moral and Politick* (1650); *Leviathan; or the Matter, Forme, and Power of a Commonwealth, Ecclesiasticall and Civil* (1651); *The Questions Concerning Liberty, Necessity, and Chance* (1656).

BIBLIOGRAPHY. The standard edition of Hobbes's complete works is that by SIR WILLIAM MOLESWORTH (1839–45, reprinted 1962), comprising *English Works*, 11 vol., and *Latin Works*, 5 vol. Notable modern editions of separate works are: *Leviathan*, ed. by MICHAEL OAKESHOTT (1946); *The Elements of Law, Natural and Politic* (including *A Short Tract on First Principles*) and *Behemoth*, ed. from early manuscripts by FERDINAND TONNIES (1889; 2nd ed., 1969); and the English version of *De cive*, ed. by STERLING LAMPRECHT (1949). For details of original editions, see HUGH MACDONALD and MARY HARGREAVES, *Thomas Hobbes: A Bibliography* (1952).

Further studies: GEORGE C. ROBERTSON, *Hobbes* (1886); FERDINAND TONNIES, *Thomas Hobbes, der Mann und der Denker* (1912); FRITHIOF BRANDT, *Den mekaniske naturopfattelse hos Thomas Hobbes* (1921; Eng. trans., *Thomas Hobbes' Mechanical Conception of Nature*, 1928); LEO STRAUSS, *The Political Philosophy of Hobbes* (1936, reprinted 1952); RICHARD S. PETERS, *Hobbes*, 2nd ed. (1967); HOWARD WARRENDER, *The Political Philosophy of Hobbes* (1957); FRANCIS C. HOOD, *The Divine Politics of Thomas Hobbes* (1964); J.W.N. WATKINS, *Hobbes's System of Ideas* (1965); KEITH C. BROWN (ed.), *Hobbes Studies* (1965); M.M. GOLDSMITH, *Hobbes's Science of Politics* (1966); F.S. MCNEILLY, *The Anatomy of Leviathan* (1968); DAVID P. GAUTHIER, *The Logic of Leviathan* (1969). There are also substantial treatments of Hobbes in CRAWFORD B. MACPHERSON, *The Political Theory of Progressive Individualism* (1962); and JOHN PLAMENATZ, *Man and Society*, vol. 1 (1963). A delightful biographical account of Hobbes is included in *Aubrey's Brief Lives*, ed. by OLIVER L. DICK, 3rd ed. (1960).

(Ed.)

Hobbies

Hobbies, broadly defined, are constructive leisure-time activities that may have a variety of goals: pleasure, relaxation, therapy, self-improvement, travel with a purpose, community service, achieving recognition and popularity, or making new friends and social contacts. Other goals may include developing an avocation or making money. Hobbies may also be defined by focussing on the individual's attitude toward any particular leisure-time activity: if he thinks of it as a hobby, if it fills his nonworking hours in a relaxing and rewarding fashion that gives him pleasure or satisfaction, to him it is a hobby. The term hobbies thus covers thousands of leisure-time activities.

Distinction between hobbies and sports and games

For the purposes of this article, sports and games are not included as hobbies, although they may have similar characteristics and goals, and, indeed, there is unresolved disagreement as to whether sports are hobbies. Athletes and sportsmen are not to be considered precluded from having hobbies, however. Outdoor enthusiasts of small-game hunting and fishing, for example, are noted for enjoying such hobbies as taxidermy, fly tying, writing, photography, and collecting books, movies, art, and accessories that reflect their special interests.

This article covers the history of hobbies and the evolution and continued growth of the many kinds of hobbies available to and pursued by modern man, including hobbies relating to art, hobbies relating to nature, hobbies relating to the past, and hobbies relating to handicrafts and skills. For specific information regarding hobbies, the reader should consult the sources listed in the *Bibliography*.

Evolution and growth of hobbies

The present-day interest in and encouragement of hobbies throughout the world has no counterpart in any other period of civilization. Once the prerogative of the leisured classes, of royalty and wealth, in the second half of the 20th century hobbies have achieved the status of a necessity for practically everyone. No age group is exempted. Hobbies are no longer considered something for the very young, the idle rich, the retired, or the infirm. They are accepted as a natural corollary of free time directly tied in with shorter working hours, timesaving and laboursaving computerized technology, more and longer holidays, and early retirement from regular work, as well as man's longer life-span. Some economists foresee a 30-hour or even a 25-hour workweek before the end of the 20th century. More and more, the need to learn how to use nonworking hours in a way that will afford pleasure, repose, and satisfaction becomes increasingly urgent.

EARLY DEVELOPMENTS

Origins. Some of today's favourite hobbies were not always associated with free time and pleasure. The activity of collecting, for example, is rooted in the origins of human society. It preceded the arts of metallurgy and agriculture, for the earliest economy was based on food-gathering and the instinct for making oneself as comfortable as humanly possible. Early man lived by picking and choosing, identifying, sampling, and exploring, spurred on by his curiosity and aided by the developing powers of his memory and his ingenuity. The importance in the modern world of this oldest manifestation of human culture is revealed not only by the existence of museums and institutions filled with the results of man's urge to collect but also by the widespread and persistent popularity of collecting—of stamps, coins, books, seashells, autographs, figurines, antique furniture, automobiles, glassware, weapons, and inumerable other items and artifacts.

Many hobbies date back through centuries, some to prehistoric times, hence it is not possible to pinpoint the origins of hobbies per se as being the exclusive discovery of any one people or the product of any one period in history. A sampling of familiar hobbies that date back to the beginnings of early civilization in one form or another would include: music, dance, poetry and prose, drawing and painting, sculpture, carving and whittling, weaving, embroidery, basketry, pottery making, beadwork, kite, toy, and doll making, leathercraft, raising pets, making hunting decoys and fishing lures, making jewelry and miniatures, and folklore, magic, and astrology. They were not referred to as hobbies, of course; the word did not even exist.

Hobbies of early civilizations

There is a popular notion that cavemen were the first known creative hobbyists. Prehistoric cave paintings and drawings authenticated as 10,000 to 15,000 years old, some even older, were discovered in 1879 near the village of Santillana, about 20 miles west of Santander, Spain, and at Lascaux in central France in 1940. The paintings and drawings, considered excellent by any critical standards, were often found in the farthest recesses and most forbidding corners of the caves. Some of the caves were entirely uninhabitable. There is mixed anthropological opinion as to whether these drawings were done for magical, religious, or decorative purposes. They could have been drawn to identify the animal to be slain and in what manner. It is to be hoped that the caveman enjoyed his task, but there are questions that becloud any certainty that it was a free-time pursuit for pleasure.

While many rulers in ancient times collected art treasures and rare manuscripts and other documents by conquest or threats of force, Aristotle is probably the most famous of the earliest hobbyists. Strabo, the Greek historian and geographer, declares that Aristotle was the first person who collected a library in Greece, and that it was he who communicated the taste to the Egyptian sovereigns. Cicero and Pliny are known to have been collectors of the writings of their famous predecessors and contemporaries. Any collector who has searched for some choice elusive items will sympathize with Pliny's complaint that letters of Julius Caesar were very scarce and hard to come by even in his own time.

Contributions of the Middle Ages. Doing something for pleasure in one's spare time was certainly not encouraged during the Middle Ages, although it was a period of great productivity of fabulous *objets d'art*, paintings, sculpture, book illumination, and great tapestries, mostly with religious connotations. The themes of the tapestries were broader in scope than any other contemporary works of art and included ancient legends, mythology, sport, chivalry, and history—such as the famous "Bayeux Tapestry" depicting the Battle of Hastings in 1066, which was, in fact, embroidered. Many of the finest examples of these works now repose in the museums of the world.

In those periods of history that divided men into nobles, commoners, and slaves, few were permitted to have a hobby. That was the prerogative only of the noble, the rich, and the venturesome. Free time was not a commodity available to everyone, and, although museums and private collections house notable examples of ancient decorative arts and handcrafts, little of it can be attributed solely to peoples' hobbies. Perhaps the one exception would be children's toys.

In the 16th century, conceivably earlier, a hobby referred exclusively to one thing—a hobbyhorse—eventually shortened to hobby. The hobbyhorse was a favourite and universal toy for children of all ages. In appearance the hobbyhorse, or hobby, ranged from a simple stick to a gaily caparisoned framework or the like, surmounted by an imitation of a horse's head. Either way, it served adequately for the meaningful and relevant games of the times that involved war, knighthood, chivalry, the good men and the bad. In time, the popularity of the hobbyhorse as a toy declined, but the fervour in pursuit of the pastime that had historically marked children's enjoyment of the hobbyhorse remained in the language to signify unusual interest in any subject outside routine daily activities.

Origin of the term hobby

GROWTH IN MODERN TIMES

"Renaissance man." Before the age known as the Renaissance, however, there had appeared many examples of "Renaissance man"—"complete men," many-sided individuals who pursued their wide and varied interests in

many fields—history, art, literature, the natural sciences, and philosophy. Later, before the fields of physics, chemistry, biology, geology, and meteorology were divided into separate sciences, many men who were physicians, barristers, tax collectors, farmers, or engaged in other professions and trades made surveys and field trips; travelled to other countries; collected minerals, fossils, plants, and artifacts; formulated and carried out experiments; studied, wrote, and read papers; and were known as natural philosophers.

This article does not attempt to identify the broadly ranging interests and activities of such men as prototypes of modern hobbies. It does intend to suggest that the sense of pleasure and satisfaction such men obtained in the pursuit of interests and activities outside of their main line of work is similar to the sense of pleasure and satisfaction sought and obtained by modern hobbyists. Before the term hobby came into general use, many men engaged in constructive and satisfying activities outside the routine of their principal occupations.

The arts-and-crafts movement. Specific inclusion of the creative arts and crafts as hobbies both on a participating as well as a collecting basis would appear to have gained concrete impetus in the second half of the 19th century with the publication of the English magazine *The Hobby Horse*, which publicized the arts-and-craft movement that centred around the artist and craftsman William Morris and his contemporaries.

Appearance of the magazine *The Hobby Horse*

Following public exhibitions of handicrafts held in England between 1888 and 1896, similar exhibitions were held in Boston and Chicago in 1897.

Victorian ladies whiled away their winter evenings making household elegancies that were certainly of the hobby class: the 1870 shadow box with dried flowers, leaves, and grasses; the hand painting of cheap, undecorated pottery called Albert ware; the wax fruits; Berlin woolwork (brilliant worsteds on canvas) for needlecraft novelties; transparencies mounted on glass—all are treasured as charming examples of the Victorian lady's hobbies.

Unlike the ladies, Victorian gentlemen, although many made collections of minerals, coins, or other items, did not engage in handicraft hobbies to any great extent, at least not until fretsaws, or scroll saws, began to appear during the 1850s. Thereafter, for about 20 years, thousands of men produced whatnots, fancy picture frames, bracket shelves, even scroll-work cutouts of inscriptions and quotations, although the supplier of printed patterns warned that only experienced fretsawyers should attempt the prodigious task of sawing out the Lord's Prayer.

Kinds of hobbies

Choosing a hobby may be an extension of many things: a person's habits, character traits, and personality, daily work, tastes, and ambition. It may be in direct contrast to the hobbyist's daily job as an escape from routine, fragmented work, and boredom. It may fill what would otherwise be idle time with an interesting activity, in the case of an older person with time on his hands. Chance and opportunity may be factors. Available time for a hobby, the probable cost, equipment if any, and the facilities that may be necessary are all considerations.

Popularity of work-around-the-house hobbies

Large numbers of people have been asked in surveys what they would like to do if they had more time. The largest proportion said that they would do more work around the house, and the second largest that they would spend more time with their families. Work around the house incorporates such hobbies as gardening and landscaping, gourmet cooking and canning, woodworking, making and restoring furniture, decorating, remodelling and repairing small appliances, toys, china, and bric-a-brac. For those who would spend more time with their families there are hiking tours and nature walks, enjoyment of the arts as spectators or participants, astronomy, model building, and raising pets, to indicate but a few of the possibilities.

Information about any hobby may be obtained through books, magazines, and newspaper articles devoted to the subject; stores, shops, galleries, and other places offering related items; instruction and discussion in schools, colleges, adult-education programs, and community centres; clubs and organizations actively interested in specific hobbies; private and public exhibitions, special sales, and auctions of items in a particular field. Many hobbies are promoted in large-scale exhibitions and sales with participating dealers and exhibitors. Those presented with the greatest frequency are antiques, coins, stamps, photography and sound equipment, vintage automobiles and other nostalgic interests, rare and old books, horticulture and flower arrangements, bottles, model-building kits, home improvement, and works of art in all categories.

HOBBIES RELATING TO THE ARTS

The arts are usually accepted as the most significant and important of all hobbies but not necessarily the most popular. It is possible to be active in both the visual and performing arts on many levels of participation: the role of artist or craftsman itself, sponsor, entrepreneur, student and scholar, critic, collector, viewer, and auditor.

Painting and sculpture. The arts have been enjoyed without benefit of money, formal education, prior knowledge, or experience, instruction, or guidance. A modicum of each is useful, however, and not prejudicial to enjoyment. The amateur status of the hobbyist in art provides a latitude for doing something badly without too much skill and still enjoying it enormously. Amateur painters have included such distinguished hobbyists as U.S. president Dwight D. Eisenhower and British prime minister Sir Winston Churchill. Churchill was the first concededly amateur artist ever to have a retrospective show of his own at New York City's Metropolitan Museum of Art. As an amateur's show, it was accorded a most favourable and enthusiastic reception; by professional standards it would not have been as graciously received.

Those who would eschew formal training in the arts will find encouragement in Lloyd Goodrich's comment on early American primitive painting: "For the fullest pictorial record of America in the eighteenth and early nineteenth centuries, one looks more to naive than professional art." A notable example is the American primitive Grandma Moses, who, in her 70s, too frail to work any longer about the farm, started painting and achieved both a critical and financial success.

For the tyro there are packaged beginner's kits with an assemblage of materials and instruction for painting by the numbers, collage, decoupage, papier-mâché, the making of mobiles, sculpture, linoleum- and wood-block printing, and other art forms.

Decoupage and collage

Decoupage and collage have been described as the age-old kindergarten pastimes of cutouts and paste-ups on a modern adult scale. In simplest terms, decoupage, which originated in France in the late 17th century, is the cutting out of designs and patterned materials and applying them permanently to any surface with adhesives to create an attractive decoration. Decoupage and collage are frequently exhibited framed. Decoupage as a decoration for furniture, screens, wall panels, and trays is at least several centuries old. In Italy it was once called the poor man's art; those who could not afford an artist's services did their own decorating with decoupage. Still one of the great appeals of this activity, which costs practically nothing to enjoy, is that many plain but useful everyday items, such as containers, frames, and metalware of all kinds, can be made to look appealingly distinctive and can be easily sold in boutiques and other specialty shops. An offshoot of decoupage, collage, a 20th-century art form, employs the technique of decoupage usually within the limitations of a frame, like a picture, to create an artist's statement with the same authority as a painting done with oils, watercolour, or other mediums of expression. The mediums may be mixed: the Spanish painter Pablo Picasso was one of the first modern painters to use the technique, combining materials with paint on canvas.

Sculpture and painting are usually presented together, but the actual molding of clay under one's hands with the thumbs as tools has an appeal for some not matched by the application of paint with a brush. Modelling with clay can be begun with a packaged kit of all the materials

and instructions. Many famous sculptor-teachers have thought that preliminary training is unnecessary; according to them, the only way to learn is to take the tools in hand and begin. The beginner is advised to start with clay as a medium before advancing to the more difficult carving of stone or wood. Not all sculpture is done with traditional materials. Hobbyists have undertaken sculpture with such items as soap, bread dough, vegetables, wax, nuts and bolts, pieces of scrap metal, various plastics, automobile-radiator cores, and heaps of discarded junk. Sculpture can stir the imagination; its limitations are only the sculptor's own.

Music. Music has become for many people a passionate interest combining hobby and avocation. Business and professional men and women, housewives, and others seeking to sooth emotions tensed by the increasingly hectic pace of modern living have learned to play guitars, recorders, and many, many other instruments. Those who enjoy vocal music participate in choirs or choruses or form duets, trios, quartets, octets, or other groups to study and sing folk music, popular songs, the music of particular or many cultures, nations (*e.g.*, Slavic; Israeli), or periods (*e.g.*, the Renaissance), or other categories of music. Many persons, of course, enjoy learning and singing solo, for their own satisfaction or to entertain others. Many also take up an instrument, the guitar, for example, to accompany themselves or others in song.

Group instruction for musical amateurs

For amateur music makers, group instruction has many advantages. Class or group rates are much lower per student, and there is also an element of stimulation in a class of music makers that speeds up the learning processes and increases enjoyment.

Some instrumental music-oriented hobbies that have a large following are: playing in amateur chamber-music groups, participating in community concert events, collecting ancient or unusual musical instruments, and collecting duplicate recordings of symphonies, concertos, and other performances by different artists.

Theatre. The theatre has always been a stimulating intellectual hobby and retains its popularity, despite the competition in many countries of television and in all countries of radio and other activities that compete for people's free time and interest.

Anyone who is even marginally associated with a civic theatre or a similar group becomes aware of the great range of opportunities, not to say needs, for hobbyists in many fields, from the inception and execution of stage and set design and lighting and costuming to directing and acting in the final production. Information as to current activities of performing-arts groups usually is available in libraries, bookshops, and local newspapers.

Not all performing-arts groups develop the same activities; the very differences from one group to another offer a choice for prospective participants. What an amateur group produces usually reflects the tastes and interests of the group's membership. It may range from folk theatre, Greek drama, Shakespeare, Ibsen, or Gogol to original works written by a member or members of the group and various forms of spontaneous theatre. A current trend is toward meaningful plays that relate to solutions of contemporary problems. Some groups stage seasonal or annual pageants or mystery or morality plays, including passion plays or revivals, especially in Great Britain and on the Continent, particularly in France, Germany, The Netherlands, and Austria. Taking part in various aspects of these productions is a hobby for at least some of the participants.

A delightful hobby related to the theatre is going to the theatre itself; for some, even a bad play is better than no play at all, and a great play is an experience long remembered. Collecting playbills, scenes from the plays, and reviews and studying the history and origins of particular types and forms of theatre are offshoot hobbies that engage many people who enjoy the theatre.

Photography. Photography is certainly the best served of all hobbies. Books, magazines, newspaper columns, and special shops are devoted to it, and many stores that never sell cameras have films for sale and a film-developing service for colour and black-and-white pictures. Taking pictures has seemingly been reduced to remarkable simplicity, with cameras so automatic as to relieve the picture taker of any thoughts other than to make certain that what he wants to take a picture of shows in the viewfinder and that there is enough film on hand.

Photography as a hobby commands the respect and interest of those who look upon photography as an art form and those who use it as a second hobby to document another hobby or interest. Photography as an art form derives from an understanding of what every phase of picture production can contribute to the finished work from the concept of an idea to the framing. As in all art forms, each artist must find his own answers. A camera club or an experienced photographer companion can accelerate the processes of learning.

Photography as art and to document another hobby

Some hobbyists will work only in black-and-white, some only in colour. Some are interested only in slides or transparencies, some only in prints, some in the arts of developing, cropping, printing, and enlarging their own photos, and some in the many aspects of making motion pictures.

As for subject matter, one enthusiast will photograph anything and everything that appeals to him; another will specialize in portraits, wild flowers, scenes of the city, birds, temples, cathedrals, or any other subject of special interest. Some of the subjects that have later developed into books include celebrities, architectural oddities, wild animals, insects, ghost towns, unusual craft skills, covered bridges, gardens, flower arrangements, and advertising incongruities. Other persons connect photography with travel: the camera has become the badge of the tourist around the world.

Writing. Whether writing is an art or a craft has never been definitely resolved. A newspaperman who can write a story while the composing room is waiting to set it up would be embarrassed by any reference to himself other than as a craftsman. A poet who polishes each bit of imagery until it takes wings and soars considers himself an artist and not a craftsman.

Many people take up writing—poetry or prose, fiction or nonfiction—as a hobby because they think they will enjoy doing it or that it will fill a void in their lives. Writing as a hobby has enriched the field of hobbies itself. Many hobbyists have added to their hobby's enjoyment by keeping a written record, as in the form of a journal, notebook, or diary, of their progress in collecting coins or antiques or in creating a garden. Such a record, indeed, is considered by many to be an essential part of the hobby of bird watching. Also, many of the books, pamphlets, and articles available on practically every hobby that exists were prepared by other hobbyists who over the years specialized to a degree that they could write with authority. This represents a valuable supplement to the professional writing in these areas.

Types of literary hobbies

Another field of interest that has attracted many hobbyists is local history. Local historical data is collected and written about by residents of that region whose routine jobs are unrelated to writing. Local historical societies encourage the writing of regional history and reminiscences. An example of a small city newspaper in the United States that publishes local history by amateur writers is the Kingston (N.Y.) *Daily Freeman*. It has run outstanding articles on forgotten industries of the past—the Hudson River steamboats, the Delaware and Hudson Canal, and the vanishing Catskill Mountains resorts—all contributed by readers with no professional writing experience.

Finally, the family trees of many countries would be missing many of their branches but for the amateur genealogists who made a permanent record of their research and findings.

Workshops in the arts. Where practicable, workshops in both the arts and crafts are to be preferred to working alone with the aid of a textbook. The usual workshop provides equipment, supplies, and reference materials, demonstration, criticism, a new environmental experience, the possibility of making new friends, and working with others, all of which can be stimuli in crystallizing the participant's objectives.

Collecting objets d'art. The doors to collecting *objets d'art* as a hobby were never open more widely to everyone than they are today. The traditional *objets d'art* that have attracted amateur collectors have been porcelains, tapestries, silver, china, antiquities, jewelry, sculpture, and bronzes. Miniatures, playing cards, clocks, buttons, paperweights, toys, and thousands of other classes of objects also have been the subjects of collections.

Influence of Pop culture

In the 1960s Pop culture, which had a profound influence on all the arts, made its impact on what are *objets d'art* and what are collectables. Examples of Pop art were not rare, not precious, not beautiful, nor skillfully crafted—even "incredible" to some critics—and yet they were accepted and sought after by the start of the 1970s. The artistic maelstroms of the 1950s and 1960s brought about change and sometimes reversals of many entrenched attitudes about art as well as other aspects of culture.

A dateline for the new standards and inclusions of articles of artistic worth would be the early 1960s. The new identity was Pop art, exemplified in the United States by the transfer of a can of Campbell's soup from the shelves of a supermarket to the walls of New York City's Museum of Modern Art. For the collector of *objets d'art*, the trickle of items that were sometimes indeed strained to qualify as having any artistic worth other than the exorbitant price tag became a flood of new possibilities. In addition, the new regime opened a haven to the collecting of "found" art objects. An assemblage of shells, pebbles, bits of glass, metals, pieces of driftwood, stones in abstract shapes, round and flowing like the nonfigurative sculpture of Henry Moore, the work of time and nature to be wondered at and valued only in aesthetic terms, may now grace any collection without apology.

This has in no way diminished the demand or weakened the market for the collectors' pieces of the pre-Pop-culture era. The supply of traditional *objets d'art* is not inexhaustible, and it diminishes as demand grows. When a rarity comes on the market, say, one of an existing known number of pieces, the bidding usually goes to a new high.

Many people enjoy collecting paintings, sculpture, or other works of art as a hobby. For an overall view of this field of endeavour, see the article ART COLLECTING.

HOBBIES RELATING TO THE PAST

Two popular fields of interest relating to the past have been mentioned above under writing—local history and genealogy. Many categories of *objets d'art* also relate to the past, as do all historical aspects of painting and sculpture, music, and the theatre. People's interests in the past are seemingly inexhaustible and form the basis of many rewarding hobbies centred on periods of history, historic and prehistoric peoples and cultures, historic personages and events, and social movements, to name a few. Hobbyists of history usually read everything available relating to their particular area of interest. They may also visit museums, historic sites, or restorations. They may re-enact battles, for example, with armies of toy soldiers or armadas of model ships. Or they may retrace in person the routes of campaigns or explorations or migrations. But whatever else the pursuit of hobbies relating to the past may entail (or offer), they nearly always include the collection of items either from the past or directly relating to it.

Three of the oldest and most persistently popular hobbies relating to the past are collecting antiques, collecting coins, and collecting stamps. These three are briefly dealt with below, as are three examples of the collecting of memorabilia—of transportation, of sports, and of military history.

Antiques. To the collector, antique means old, but it also carries connotations of aesthetic, historic, and financial value. Thus the decorative arts of all past eras have come to be considered antiques.

Many countries permit objects that qualify as genuine antiques to be imported duty free. In 1952 the so-called Florence agreement, sponsored by UNESCO, agreed to "facilitate the free flow of educational, scientific and cultural materials," and antiques were affected by subsequent legislation adopted in the participating countries. By the 1970s, 50 countries were applying the agreement, including the United States, which in 1966 passed a new tariff act permitting the duty-free importation of "antiques made prior to 100 years before their date of entry" (previously, under the U.S. Tariff Act of 1930, an object had to predate 1830 to qualify).

The antiquity of collecting antiques

The collecting of antiques goes back almost as far as history, beginning with preservation of temple treasures. In England, concern for antiques led to collections as early as the 16th century. Such collecting subsequently increased, stimulated by Romanticism, growing interests in antiquities in general, and the start of archaeology. In the second half of the 19th century, museums and collections of decorative arts, intended to stimulate designers as well as collectors, were established in London, Vienna, Paris, and New York. Antique collecting in the United States was given its first strong impetus by the Philadelphia Centennial Exhibition, 1876, which awakened interest in the national past.

When antiques are collected as a hobby, investment, although antiques have been recommended by financial publications as an investment, should be only part of the consideration for buying. Every collector has experienced at some time or other pleasure in these aspects of the activity: the search and discovery of a much-wanted addition for a collection; the possibility of a bargain; the reluctant straining of the purse strings when an item is too good to pass up and the avowal to economize on something else; the moment of adding a new treasure to a collection; the moment of showing it to others, especially other collectors; the study of books that may establish an acquisition as rare.

The search for antiques is a favourite travel hobby, and many vacation itineraries are planned for antique hunting. The numerous published price lists for popular antiques rarely leave a dealer in doubt as to what his wares are worth, but then the established price of today's particularly choice piece may be tomorrow's bargain.

Coins. Coin collecting, or numismatics, the scientific term for coin collecting, is related to many fields of study including history, archeology, languages, religion, geography, and economics. Coin collecting has always been recognized as a favourite hobby. Together with stamp collecting, it has been widely recommended for its educational value, the good habits it promotes, the friendships that develop with fellow collectors, and the benefits of quiet relaxation that are inherent in this activity. The thought of financial return at some time in the future when the collection will be ultimately sold is always a pleasant consideration but rarely is of primary consideration with the beginning collector.

The collection of ancient coins was stimulated by the Italian Renaissance, and many collections of Greco-Roman coins were made during the 15th and 16th centuries. In the 17th century, numismatic scholars undertook analysis of existing great collections, of which Italy had more than 350, France and the Low Countries about 200 each, and Germany not many less. Many collections had been uncritically accumulated and most included spurious pieces. From the 18th century onward greater discrimination was exercised.

Coin collecting stimulated by the Renaissance

The number of lesser collectors grew in the 19th century, aided by the publication of reliable handbooks and authoritative catalogs, such as the British Museum series from 1873. Numismatic societies were founded in many European countries, beginning with the Numismatic Society of London, now the Royal Numismatic Society, in 1836. London emerged and remains as the world's largest numismatic market, serving public and private collectors in many lands. International cooperation is promoted through the International Numismatic Commission (founded 1936) associated with the International Committee for Historical Sciences.

A coin collector may gather coins from all over the world; may limit himself to the coins or paper money of one country; or concentrate on current coins or on

one period, one denomination, or one particular coin. In general, private collectors of the 20th century, greatly increased in numbers, have had to specialize in order to build a satisfactory collection.

Stamps. Stamp collecting, or philately, attracts the greatest number of collecting hobbyists. Most daily newspapers carry some stamp news: at a minimum, post-office news releases that tell of new commemorative issues, dates of issue, where these are to be first placed on sale, and something of the historical background of the stamps themselves. Many newspapers have weekly stamp columns, occasionally expanded into full departments with dealer advertising. There are the weekly, semiweekly, and fortnightly publications for both collectors and dealers.

Entertaining, fascinating, engrossing, challenging, an armchair adventure in prospecting for rarities and bargains—these are all valid descriptions of stamp collecting. Obviously it may be pursued at almost any age and anywhere by anyone.

The first postage stamps for the prepayment of letter postage were issued in England in 1840; stamp collecting began almost immediately thereafter. In 1841 an advertisement in *The Times* of London asked "if any good-natured person" would assist a young lady in collecting cancelled postage stamps to cover her dressing room, and in 1842 *Punch* labelled stamp collecting "a new mania."

Beginnings of stamp collecting

From the practically simultaneous appearance of catalogs in the 1860s, it seems clear that collecting was spontaneous, with Great Britain, France, Germany, Belgium, and the United States apparently having active collectors before 1862.

The Philatelic Society (from 1906 the Royal Philatelic Society) was founded in 1869 in London, the American Philatelic Society in 1886, and the Society of Philatelic Americans in 1894. In addition to the many national societies there are clubs or societies of collectors in most cities of the world.

With standard catalogs listing more than 200,000 items, the tendency toward specialization is even greater among stamp collectors than among coin collectors. Speciality clubs dealing with the philately of, for example, France, Germany, Italy, Luxembourg, Japan, Switzerland, Israel, and the United Nations as well as special subjects, such as airmails, exist to serve almost any interest.

Thematic or topical philately attracts many people as an extension of some other activity. Some familiar topical interests are transportation, music, literature, statesmen, arts and crafts, naturalists, inventors, liberators, and other heroic figures. Whatever a collector's other interests, there are stamps pictorially reflecting some historical or promotional phase of the subject.

Memorabilia. Progress and the changing face of society bring thousands of new hobbyists into the collecting field each year. Nostalgia, sentiment, a desire for continuity with the past, and a reminder of a way of life that has passed are reflected in the things that interest and appeal to collectors. Many collectors treasure memorabilia relating to antique automobiles, ferries, the circus, apothecaries, steam engines, steamboats, and other symbols of a past that is disappearing.

Collecting activities frequently start with the accumulated contents of an attic or cellar of any house, or a garage, barn, or shed used for storage. The greater the accumulation of discarded items and the longer they have been untouched the greater the possibility of converting trash into treasure. Before doing anything with an accumulation of forgotten potential treasures, it is worthwhile to attend a few antiques shows and flea markets to see what is collected and sold. Interest in nostalgic items is not just in genuine antiques more than 100 years old. Nostalgia may encompass any period—a collector's youth or childhood, his parents' or grandparents' time. It may be as recent as the 1920s, 1930s, or 1940s, which already have their collecting vogues. Collectibles of these periods are hopefully referred to as tomorrow's antiques.

Collecting tomorrow's antiques

Typical atticana includes collections of postcards, boxes of assorted buttons, stereoptican slides, books, newspapers, periodicals, advertising cards, paper fans, political buttons, tokens, road maps, license plates, automobile-operating manuals, home-remedy bottles, valentines, phonographs and records, fruit jars, mail-order catalogs, kerosene and carbide lamps, travel souvenirs, sheet music, tin boxes, scrapbooks, inkwells, dolls and paper cutouts, toys, sewing machines, pictures and frames, coin banks, tokens and badges, war mementos, circus, theatre, vaudeville, and sports programs, paperweights, miniatures, odd pieces of china and silver, player-piano rolls, brass beds, comic books, figurines, hatpins, buttonhooks, watch fobs, pocketknives, banners and posters, family photo albums, local directories, and old almanacs and atlases.

All of these items may be worth money and are constantly sought and sold through collector and hobby publications.

Transportation. All means of transportation, defunct or still in existence, such as railroads, sailboats, steamboats, ferries, trolley cars, stagecoaches, and public liveries, are of interest to collectors. Their memorabilia include contemporary photographs, prints, paintings, advertising posters, and travel circulars, stock certificates, signs, bells and whistles, lamps and other fixtures, imprinted hardware, furniture, silverware, passes, copies of construction plans and specifications, waybills, handmade models or toys, and route maps and other paper ephemera. Books and articles on these subjects frequently represent a lifetime of interest and might otherwise never be written.

Nautical memorabilia frequently has more than just sentiment to recommend it. Sailors of the mid-19th century—1820–70—produced some notable examples of scrimshaw and carvings in wood. Scrimshaw is engraving on the bones and teeth of the sperm whale and on walrus tusks. The etched portion was usually inked or rubbed with pigment to bring out the design detail. Sailors' carvings of eagles and figureheads sometimes equalled the skills of professional carvers of the period. Much of this work was whittled out with only a jackknife.

Scrimshaw

Sports. Sports enthusiasts create much of their own memorabilia in the form of autographs and photographs of sports figures, scrapbooks devoted to a particular sport or to individual players or contestants. Some hobbyists collect souvenirs, items of clothing, or other articles bearing their team's name, emblem, or colours, or miniature replicas of players, equipment, or trophies. Some sports created a legacy of antiquities for collectors, gutta-percha- and feather-stuffed golf balls, for example, and antique golf clubs, guns, and fishing rods and reels. For many sports, especially those that maintain established collections of their own, such as the baseball and football halls of fame in the United States, it is difficult to obtain any genuine items except under circumstances that authenticate the source. The triumphs of hunting and fishing, however, can be reflected in mounted trophies of the collector's own.

Military history. In the United States, the years 1861–65 attract more leisure-time interest than any other four-year period in American history. The war between the states has broad areas of specific interest: historical reading and writing; stamps and money; photography; music; weapons; discovery and invention; military souvenirs; and Lincolniana. Stamps and money of the Confederacy have increased in demand and value. It is still possible to collect a complete set of Confederate stamps at modest cost. Some Confederate coinage is scarce, though the paper money is plentiful. The state notes, which were limited in quantity and for use only within the issuing state, are in some instances more valuable but still to be had. The field of collecting toy soldiers of many lands is well represented by examples of Civil War soldiers in addition to the ordnance of both sides. In all countries, World War I posters and captured enemy matériel of World War II are the more prominent collectibles of those periods, although books about various battles and campaigns and the various generals on all sides of the conflicts have been best sellers.

The American Civil War as a hobby

Military figures are widely collected by historical china hobbyists. The Napoleon crest on china, porcelain, glass-

ware, and other materials is a collector's item, just as in the same manner elephants, owls, eagles, monkeys, occupational memorabilia and miniatures in all materials are collected by other specialists.

HOBBIES RELATING TO NATURE

Nature is man's oldest hobby. It is now his greatest concern. Environmental deterioration has affected many aspects of such simple pleasures as watching and studying birds and other wildlife, hiking and nature walks along trails and streams, even the breathing of "country air." Conservation and improvement of the environment have become corollaries of interest to every nature lover who has witnessed the corroding deterioration of the natural resources. Conservation is in itself an activity of growing involvement for everyone. The study of conservation and participation in improved environmental projects can involve important public service and encourage others to enlist in environmental improvement programs over the long term.

Conservation as a hobby

The nature story is as involved and varied as the human story. It is more than trout fishing in the spring, camping outdoors in the summer, and deer hunting in the fall. The nature story invites the study of wild flowers, trees, and shrubs, birds and bird sanctuaries, small game, brooks, and rocks and fossils, all of which should be seen in their environment.

The city or suburban family can enjoy 52 weekends a year in the country without venturing more than 50 to 100 miles from its doorstep. Not everyone has the time or facilities for travel. Most cities and suburban areas have developed the local resources of parks, nature trails, wildlife sanctuaries, zoos, and museums within their own area, however, so that much can be seen and learned about nature within walking distance or a local bus ride. Enjoying the outdoors may be a family or group activity, while the experienced nature hobbyist will not object to travelling alone; in fact he may prefer to. But the novice or beginner usually will find companionship more pleasurable, safer, and more instructive.

Gardening. Gardening is a popular hobby activity that has undergone a change of emphasis and attracted new adherents in the process. The 20th-century exodus from city to the suburbs and the hinterlands of intensively industrialized and urbanized areas has both new and traditional gardeners thinking in practical terms of better home and garden planning to ensure maximum green areas and the best land use, something beyond the expanse and expense of a well-kept lawn.

The thought now is to have the trees, shrubs, and other plants work together, the purpose being to improve the environment with its resulting benefits. Emphasis, for example, is on soil improvement through green manures and other organic materials that were once burned up and polluted the air or were hauled away as garbage. Fresh vegetables organically grown on a home plot have a status of involvement, the unmatched taste of freshness, and the saving of time and money in shopping. The home canning or freezing of fresh fruits and vegetables produced in one's own garden is an expanding activity in the same tradition.

Flowering and fruit-bearing trees do more than contribute to the variety of each season. They provide shade, play a cleansing role in air and water pollution, and contribute to an appealing environment for birds, squirrels, and other small animals.

City gardening

No space or land should be considered too small if there is adequate light and water. Soil and containers can be brought to the site if necessary. City gardeners, for example, find the challenge of the pocket-size garden on terraces, rooftops, backyards, and miniature parks in open lots a rewarding experience when the urban handicaps are overcome and the rewarding green oasis is used and enjoyed.

The best gardens are usually planned gardens, demanding no more time, labour, and expense than the hobbyist is prepared to put into the gardening project. It may be a family affair or a personal activity.

The creation of gardens for pleasure is a most ancient art. Egyptian wall and tomb paintings, Minoan frescoes, descriptions in the *Odyssey*, the records of classical Rome, the writings of ancient China, and the culture of Japan all testify to the enduring history of gardens as havens of quiet and peace in a troubled world.

Modern gardening hobbyists usually become interested in one or more of several broad categories encompassing flower, woodland, rock, water (or pool), scented, herb, and vegetable gardens. Other interests include arboretums, roof gardens, greenhouse or conservatory gardens, and indoor plants and window boxes. (See also the article GARDENING.)

Bird watching. Public awareness of the environment has contributed to a vast increase in the number of bird watchers, and indeed changes in the environment have contributed to changes in bird ranges. One of the many by-products of concern over the environment has been man's curiosity to find out what its components are. One of these components, of course, is the bird life of a country. Birds are sensitive indicators of change.

In order to be able to observe these changes, the bird watcher must first be able to identify the birds and the different components that make up their habitat. As the bird watcher must go where the birds are to observe them, he becomes familiar with these habitat components and subsequent changes, whatever the cause.

Equipment for bird watching

The bird watcher's equipment is modest—a pair of binoculars or a spotting scope, a field guide, notebook, and pencil; many also carry cameras. Early in the morning, soon after sunrise, is the best time for bird watching. The next best time is late afternoon, when the flocks are feeding and almost ready to roost. An attractive feature of bird watching is that it takes practically no effort at all to get started. Activities in bird watching are well established and organized almost everywhere. Local museums, libraries, and nature study clubs are excellent points of beginning for inquiries.

Nature walks. Hiking on a planned trip with a specific destination or walking leisurely and observantly through a wooded area or across fields can be more enjoyable when one recognizes the geological causes of the terrain features or perhaps the more familiar wild flowers or even a single plant that is scarce or not indigenous to the area. In the course of travel, one may discover footprints or other evidence of the presence of animals in the vicinity. For the beginner, guided instructional tours heighten the enjoyment of the first experiences and accelerate the learning processes.

Nature collecting hobbies. Along the trails, butterflies and moths are frequently seen and captured and added to private collections. Sometimes the specimens are preserved just for viewing and identification or used in arrangements with dried wild flowers reminiscent of Victorian displays. An extension of interest in collecting butterflies and moths would include the gathering of wild flowers, leaves, branches, bark, and driftwood where available. Wooded and marshy areas are ideal sources of this material.

Collecting rocks and minerals. Collecting rocks and minerals is probably one of the least expensive hobbies to start and maintain. Rocks and minerals are probably the most plentiful and unevenly distributed substances known. In some localities anywhere from 75 to 100 different minerals may be found within the space of a few miles, whereas in other areas possibly as few as five or ten. Mountains, stone quarries, gravel pits, and road cuts through hilly areas are among the favourite locations for collecting rocks, and old mines and dumps are common sources of minerals. Most collectors agree, however, that the hobbyist's immediate vicinity will generally yield a sufficient variety with which to start a collection. Collectors of stones may look for particular colours or varieties of colours, shapes, or textures or may concentrate on places of origin. The amateur collector may become a member of a local mineral or gem society, attend regular meetings, and go on conducted tours and trips to obtain specimens in other locations.

Sources of rocks and minerals

Shell collecting. A hobbyist who collects and studies seashells is a conchologist or a malacologist. Experts say

that there are 75,000 known species of seashells. One basic appeal of seashells is that they are plentiful and frequently may be obtained without cost. Seashells can be found at almost any hour of the day, but, to obtain the best specimens available, experienced collectors search for shells at low tide when they have the entire area between tides to explore. Shells of animals that live in shallow water but beyond the tide line are washed up on the beach only after storms and must be hunted then. Shells of animals that live in deeper water have attracted many of the growing numbers of skin divers to the hobby of shell collecting.

The rarity of certain specimens is of greater interest and importance to the collector than to the craftsman. Some shell dealers specialize in shells for collectors, others sell shells in bulk for shell craft. Needless to say, bulk shells are less rare and cost less. The opportunity to earn money with shell craft attracts many people, especially in those areas where shells are to be had for the gathering. There are numerous salable items that are made with shells: jewelry, trays, picture frames, bookends, greeting cards, sometimes with shells as the sole material and sometimes in combination with other materials.

Collecting artifacts and fossils. Other hobbies related to nature include searching for artifacts of historic and prehistoric man and for fossils of animals that inhabited the earth millions of years ago. These outdoor hobbies reflect widespread interests in archaeology and paleontology and are similar in many respects to the hobbies of collecting rocks and minerals and shells.

Raising animals. Raising horses, dogs, cats, monkeys, canaries, parakeets, or other cage birds, tropical fish, snakes, or any of the numerous common four-legged animals such as rabbits, guinea pigs, and hamsters has long been a popular hobby. Some people raise herds of purebred cattle or raise sheep, goats, or other animals as a hobby. Some hobbyists raise animals to sell as a leisure-time source of income. Some plan to sell only their surpluses. The owner of a pet dog, cat, or other animal for companionship is not usually considered a hobbyist in this sense.

To raise animals, including pets, as a hobby requires knowledge and observance of all the improved methods of animal care. Location—in city, suburb, or country—influences the type of pets that may be raised, as local ordinances may limit the raising of pets in quantity except under supervised conditions and may forbid such activity entirely.

HOBBIES RELATING TO HANDICRAFTS

Goals of handicraft hobbies

In handicraft hobbies, the ultimate goal is to fuse the intangible qualities of creativity, aesthetic sensibility, and skill into a work of art in its own right. It is most desirable for the hobbyist to examine the techniques and finishing that distinguish one example of handicraft from another of similar character and to note the differences that make one piece worth many times more than its counterpart. Art in craftsmanship is often unobtrusive while it brings the beauty of a work to the fore. Details of style and craftsmanship can be financially rewarding in those fields often regarded as women's domain: dressmaking, ladies' accessories, knitting, embroidery, quilting, and other forms of needlework and rug making.

Some handicrafts require more equipment than others. Experts advise starting with a minimum of simple tools and then acquiring others as skill and need grow. Tools used for one craft are often suitable for others as well.

Some "work alone" crafts have a special appeal for people who prefer to work unhurriedly in the quiet of their own company. Anyone who develops a reasonable proficiency can, if he so desires, attract a clientele from far beyond the limits of his own city or town in such fields as taxidermy, clock repair, ship-model building, cleaning and repair of prints and paintings, doll repairs and restoration, bookbinding, antique auto restoration, caning, and the repair of antiques, glass, china, and bric-a-brac. The degree of skill a craftsman has acquired and his knowledge of the craft should determine the area of repairs, restoration, or new work he should consider undertaking. Age is not a handicap in these crafts in any sense. Most people who adopt these activities have many of the characteristics displayed by professionals in these fields, the only difference being in the degree of experience. Often retired from regular work, they are usually unhurried, knowledgeable, patient, like to handle a job through from start to finish, and are satisfied with nothing less than the best job possible. The best instruction is through apprenticeship.

Packaged kits. The manufacture of hundreds of millions of dollars worth of hobby materials each year is the largest single segment of supply and inspiration for hobby activities. Model making and the occupational arts and crafts are the major aspects of the hobby material industry. Advances in technology, science, and educational methods as well as changes in public taste are quickly reflected in the manufacture of hobby materials.

Model and hobby kits

In the model-building group of hobbies, airplanes, rockets, racing cars, boats, and ships provide the action of speed, remote control, and competition with specific rules and regulations as to class of model and cost. The upsurge in popularity of track auto racing of all classes is reflected in model-car racing, where performance and the competitive elements are stressed. Model-railroad building and operation retains its popularity with both young people and adults. These action hobbies all have a strong father-and-son appeal.

At the start of the 1970s, plastic models for cars, planes, ships, boats, military vehicles, human anatomy, birds, and animals comprised the most popular type of hobby kits sold in terms of dollar volume.

Second in demand were package and bulk materials for painting by the numbers, mosaic kits, and crafts utilizing plastics, wood, leather, textiles, metal, cork, clay, beads, and sequins, shells and stone, raffia, ribbon, wool for rug making, yard goods, and laces. This list is representative but by no means inclusive of the materials available in hobby kits.

The sciences are represented by hobby kits in chemistry, biology, physics, astronomy, including telescopes, and the earth sciences, including weather stations.

Some hobbyists resist the idea of using prepared kits; some welcome the challenge of using discarded waste materials, such as tin cans and old newspapers. Practically every operation practiced with silver, copper, brass, or pewter may start with tin. Pieces may be finished off with decoration reflecting the fashions in tinware sold by the early 19th-century peddlers. Newspapers, paste, and powdered glue are the start of modelling in papier-mâché, a craft of old traditions and many beautiful *objets d'art*. The technique is easy to learn, the possible uses extensive: masks, puppets, theatrical props, float displays, model-railroad tunnels, figurines, and parlor accessories reminiscent of the Victorian era, when papier-mâché was widely used.

Of the many hobbies relating to handicrafts, three involving ancient crafts—needlework, ceramics, and woodworking—are covered in slightly greater detail below.

Needlework. Many people enjoy needlework in their spare time. The most attractive features are the many forms of needlework possible—embroidery, sewing, quilting, to mention only a few—and the fact that all forms of needlework may be taken up and put down at will. Needlework may be done alone or with others at sewing circles, craft guilds, church groups, farm and home bureaus, and rehabilitation centres.

Varieties of needlework

There is certainly no craft that requires less equipment than embroidery: needle, thimble, and thread. The methods follow the most ancient of traditions. Examples of precious fragments of embroidered pieces that date back many centuries show that buttonhole, matting, square, darning, double-running, chain, cross, and split stitches, all in use today, were used then. Although the tools and techniques are centuries old, this is not a restrictive art form. In embroidery the design need not come in repeats, even when based on counted threads. With needle and thread, ideas are given design and provide texture and form on the ground weave.

Sewing—especially the making of clothes—is as popular as ever. It is a highly functional craft. Sewing centres, adult-educational classes, simplified patterns, the trend to do-it-yourself activities, all contribute to the continued popularity of sewing.

Quilts of all types—patchwork, crazy, hit-and-miss, appliqué, and all the variations with different names—are among the best selling items at antique shows. They need not be the finished quilts, just tops. A well-designed quilt, well executed, is destined to become a treasured heirloom. Quilts are made and used because they represent warmth, comfort, and time-binding to a degree unmatched by any other single household article. Many quilts are signed and dated by their makers; all quilts should be, because few things more truly catch the personality of the maker forever.

The boundaries of needlework are being pushed forward with unusual uses of thread. Jack Lenor Larsen, a noted contemporary designer, says "Weaving is almost as obsolete a term as tapestry is for what is happening today." Fibre forms" or "art fabrics" are current terms for the ultramodern creations. The modern artisan is said to find inspiration in many places—in eastern European peasant tapestries, in modern painting and sculpture, in the pre-Columbian weavings of Peru. In contemporary exhibitions it is not unusual to see works sometimes as tall as 20 feet (six metres) and combining as many as 20 massive elements into environments, usually without figural pattern.

The birth of this innovative art-craft was proclaimed in the United States in exhibitions at the Museum of Contemporary Crafts in New York City at its 1963 show, with further developments seen at the Museum of Modern Art five years later. Since then artists have increasingly combined weaving with other nonwoven techniques, or they have given up the loom entirely to concentrate on knotting, twisting, looping, braiding, or needlework. It is possible to be startlingly innovative with the classic forms of needlework and the related craft-arts and find a ready audience for the work.

Ceramics. The making of pottery and other ceramic ware is a craft of unique extremes. The hobbyist can become active in it on so many levels of participation that it is understandably one of the most popular of all crafts, not necessarily in numbers participating in it but certainly in intensity of enthusiasm and fervour.

As a major art field, ceramics involves creative design; the techniques of building pottery forms by hand, wheel, and casting methods; the chemistry and application of glazes; and the techniques of kiln operation. At its best, pottery is one of the most controversial of all crafts. Contemporary art institutes and museums join the designer-craftsman in insisting that despite the utilitarian uses to which pottery may be put, it must be evaluated as art. In exhibitions, pottery is frequently coupled with modern art.

Pottery making and ceramics for beginners

Fortunately there is a less controversial and more conventional but satisfying range of pottery making and ceramics for those who are beginners and intermediate craftsmen. As an example of the basic level of this craft field, there are nonfired pottery methods of using clays that will dry at room temperature and others that may be baked in a kitchen oven at about 300° F (about 150° C) and later decorated with liquid glaze. These materials and these methods are used in many situations: schools, camps, and hospitals, and in homes where it is not possible to have a kiln available for firing pottery, (although groups of ceramacists frequently have a jointly owned kiln). The kitchen is often used with satisfying success for the making of one-of-a-kind pottery and ceramic pieces. Self-instruction books have been found most useful in this field and are constantly used as a reference by even the most skilled artisans. Catalogs of ceramic supply houses, which advertise in hobby and craft publications, should not be overlooked as a source of simple instruction on methods of using their materials and ideas for items to make. The two largest items of equipment generally used in a ceramic and pottery workshop are a potter's wheel and a kiln. It is possible to make a potter's wheel and to build a kiln outdoors that will be serviceable and cost little more than the time and effort.

Interest in ceramics and pottery may centre around the making of one type of item in which a person acquires a dexterity, skill, and individual design style. This may be in the form of jewelry, tiles, animal sculpture, personalized pottery, and other items ranging from highly utilitarian to merely decorative.

A natural extension of an interest in making pottery and ceramics is the collecting of unusual and outstanding pieces made by others, organizing classes and group instruction, reading and writing on the subject, becoming a member of craft museums, and enlarging one's knowledge of the art.

Woodworking. This has been traditionally a man's craft, but women are increasingly doing outstanding woodworking themselves. In the United States this is especially true in New England, where women cabinetmakers have distinguished themselves.

Many things are still made of wood not of necessity but because no other material or craft is so satisfying. The first tools that practically every man acquires, no matter what his favourite craft may ultimately be, are those for woodworking. Hammer, saw, try square, ruler, chisel, vise, knife—these are essential in every shop. Those who use power tools start with a drill, table saw, and sander.

What to make depends upon the degree of skill that can be applied, the purpose for which wood items are intended, the availability of needed wood (especially of the kiln-dried variety), and the possible use of other skills and interests. The one obvious advantage of woodworking is that everything a person can make is in demand, either for his own home, as a gift, or for sale. Among the many possible divisions of woodworking are carving and whittling, cabinetry, repairs, and restoration. The experienced woodworker's prime interest in his craft is to make things that are better styled and better executed than anything that is available in the ready-to-buy market. A few examples in addition to tables, chairs, and other items of furniture would be bird and game decoys, antique reproductions, plaques and panels, toys, miniatures such as doll houses and furniture for doll houses, ship models, and cabinets for the display of minerals and other collectibles.

REWARDS OF HOBBIES

Some rewards of hobbies can be measured by material achievements: money, medals and ribbons, citations, testimonials, a scrapbook of clippings.

Intangible rewards

Beyond the tangibles are the intangibles. These intangibles include added self-esteem, release from monotony and boredom, the revitalization of interests and of capacity to work, the aesthetic and intellectual stimulation missing from so many essential routine tasks. Any exact expression of the intangible rewards of hobbies must be personal and must be experienced. In effect, time spent in pursuing a hobby is an escape to an alternate way of life. A person may work hard at a hobby, harder perhaps than at a job, but such work is free of the tensions, accountability, and possibly the responsibilities that attend much economic work. Unlike daily, committed routines, the hobbyist's knowledge that he may stop at any time and turn to something else is part of the freedom of the alternate way of life. Conversely, much that is routine is made more endurable because of the prospect of getting back to a hobby.

From the very moment of decision to take up a hobby there is stimulation in embarking on a search for an activity that will provide the pleasure and change sought. It is stimulating to find new items and add to any collection; it is aesthetically stimulating to become so expert at an adopted craft that it wins recognition as an art; it is intellectually stimulating to know that it is still possible to assemble an important collection in many categories of books, art, or antiques.

The money that may be made from a hobby can be important money; it is unlike any other money earned. This is money from pleasure that can generate new pleasures.

The actual amount may be small; the use to which it is put, very large. It can be the difference between a tight budget and small luxuries.

Financial rewards can vary greatly from one hobby to another, but to select an activity on the basis of its possible monetary rewards invalidates both the spirit and intent of pursuing a hobby. It then becomes a form of working on a second job. There is no specific formula that can be applied with absolute assurance to determine just how much money can be earned from any given activity. One thing is certain: expertise and authority in any area, based on ability, whether it is in the collecting field, in the arts or crafts, will eventually bring requests for advice and services. Talks, lectures, demonstrations, and workshops bring in supplemental fees to many hobbyists. A substantial percentage of dealers in most hobby and collecting fields were once hobbyists and collectors. This is especially true in antiques, where the system of trading up from a poor example of an item to a superlative one leads to the accumulation and the filling in and enlarging of a collection until the collector must become a dealer to sell off unwanted surplus items. Most hobbyists who turn their hobby into an avocation for a secondary income or a new business entirely find that specialization in at least one art, one craft, or one collecting field will attract potential customers from far beyond the boundaries of their own community.

BIBLIOGRAPHY

General Works: WILBERT D. NEWGOLD, *Newgold's Guide to Modern Hobbies, Arts and Crafts* (1960), a standard reference work that still introduces some popular hobbies and crafts not found elsewhere in a single volume; LUCILE ROOD, *How to Find Leisure Time and Use It Creatively* (1968), a practical, common-sense approach to an increasingly urgent subject requiring solution; ARTHUR LIEBERS, *Fifty Favorite Hobbies* (1968), an experienced hobbyist and researcher provides an introduction and sampling of some all-time favorites.

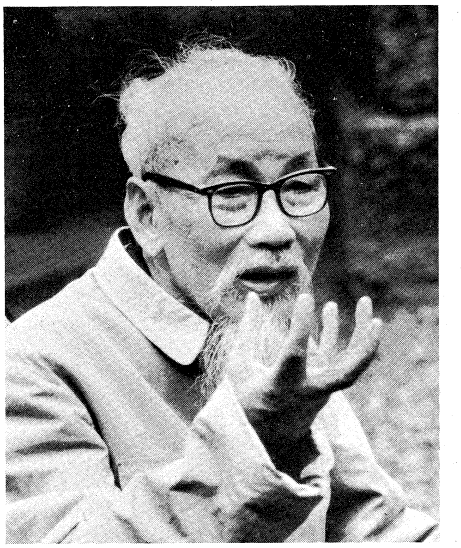
Selected references to specific hobbies: ALINE B. SAARINEN, *The Proud Possessors* (1958), an authoritative, well-researched, and well-written account of the beginnings and making of some of America's famous art collections; WESLEY TOWNER, *The Elegant Auctioneers* (1970), an informative and entertaining insight into the world of auctioneers and collectors; JOHN MEBANE, *The Poor Man's Guide to Antique Collecting* (1969), for those who believe the real point of collecting is the thrill of finding a bargain; DOROTHY JENKINS, *A Fortune in the Junk Pile: A Guide to Valuable Antiques* (1963), on finding treasure in trash; WILLIAM HILLCOURT, *The New Field Book of Nature Activities and Hobbies* (1970), a handy reference manual that has been updated since its original publication in 1961 as the *Field Book of Nature Activities and Conservation*; F.F. ROCKWELL and ESTHER C. GRAYSON, *The Rockwells' Complete Guide to Successful Gardening* (1965), an excellent reference work for gardeners; L.J. HUSSEY and C. PESSINO, *Collecting Small Fossils* (1970), makes the subject inviting and enjoyable; DONA Z. MEILACH and LEE ERLIN SNOW, *Creative Stitchery* (1970), a highly usable, easy-to-follow introduction to an increasingly popular hobby; DANIEL RHODES, *Clay and Glazes for the Potter* (1957), a practical book for the potter, student, teacher, designer, collector, or industrial ceramist; JOHN G. SHEA, *Woodworking for Everybody*, 4th ed. (1970), a good starting point on this subject.

(W.D.N.)

Ho Chi Minh

Ho Chi Minh (real name Nguyen That Thanh), leader of the Vietnamese national movement for nearly three decades and president of the Democratic Republic of Vietnam (North Vietnam) from 1945 to 1969, was one of the prime movers of the post-World War II anticolonial movement in Asia and one of the most influential Communist leaders of the 20th century.

Early life. The son of a poor country scholar, Nguyen Sinh Huy, Ho Chi Minh was born on May 19, 1890, in the hamlet of Hoang Tru, in the Nghe An province of present North Vietnam, then French Indochina. He was brought up in the village of Kim Lien. He had a wretched childhood, but between the ages of 14 and 18 he was able to study at a grammar school in Hue. He is next known to have been a schoolmaster in Phan Thiet and then was apprenticed at a technical institute in Saigon. In 1911,



Ho Chi Minh, 1968.
Marc Riboud—Magnum

under the name of Ba, he found work as a cook on a French steamer. He was a seaman for more than three years, visiting various African ports and the American cities of Boston and New York. After living in London from 1915 to 1917, he moved to France in the middle of World War I. There he worked, in turn, as a gardener, sweeper, waiter, photo retoucher, and oven stoker.

Years in France

During the six years that he spent in France (1917–23), he became an active Socialist, under the name Nguyen Ai Quoc (Nguyen the Patriot). He organized a group of Vietnamese living there and in 1919 addressed an eight-point petition to the representatives of the Great Powers at the Versailles Peace Conference that concluded World War I. In the petition, Ho demanded that the French colonial power grant its subjects in Indochina equal rights with the rulers. This act brought no response from the peacemakers, but it made him a hero to many politically conscious Vietnamese. The following year, inspired by the success of the Communist revolution in Russia and Lenin's anti-imperialist doctrine, Ho joined the French Communists when they withdrew from the Socialist Party in December 1920, and he denounced the evils of French and British colonialism in his journal *Le Paria* ("The Outcast"; Paris).

After his years of militant activity in France, where he became acquainted with most of the French working class leaders, Ho went to Moscow at the end of 1923. In January 1924, following the death of Lenin, he published a moving farewell to the founder of the Soviet Union in *Pravda*. Six months later, from June 17 to July 8, he took an active part in the fifth Congress of the Communist International, during which he criticized the French Communist Party for not opposing colonialism more vigorously. The text of his statement to the congress is noteworthy because it contains the first formulation of his belief in the importance of the revolutionary role of oppressed peasants (as opposed to industrial workers). From then on he was an important figure in world Communism.

In December 1924, under the assumed name of Ly Thuy, Ho went to Canton, China, a Communist stronghold, where he recruited the first cadres of the Vietnamese nationalist movement, organizing them into the Vietnam Thanh Nien Cach Mang Dong Chi Hoi (Association of Young Vietnamese Revolutionaries), which became famous under the name Thanh Nien. Almost all of its members had been exiled from Indochina because of their political beliefs and had gathered together in order to participate in the struggle against French rule over their country. Thus, Canton became the first home of Indochinese nationalism.

When Chiang Kai-shek, then commander of the Chinese Army, expelled the Chinese Communists from Can-

ton in April 1927, Ho again sought refuge in the Soviet Union. In 1928 he went to Brussels and Paris and then to Siam (now Thailand), where he spent two years as a representative of the Communist International, the world organization of Communist parties, in Southeast Asia. His followers, however, remained in South China. Meeting in Hong Kong at the beginning of May 1929, members of the Thanh Nien decided to form an Indochinese Communist Party. Others, in the Vietnamese cities of Hanoi, Hue, and Saigon, began the actual work of organization, but some of Ho's lieutenants were reluctant to act in the absence of their leader, who had the confidence of Moscow. Ho was brought back from Siam, therefore, and on February 3, 1930, he presided over the founding of the party. At first it was called the Vietnamese Communist Party, but after October 1930, Ho, acting on Soviet advice, adopted the name Indochinese Communist Party (PCI). Thus, the movement received greater international recognition and avoided the suggestion of "petit-bourgeois nationalism" implied by the earlier title. In this early phase of his career, Ho acted more as an arbiter of conflicts among the various factions, allowing the organization of revolutionary action, rather than as an initiator. His prudence, his awareness of what it was possible to accomplish, his care not to alienate Moscow, and the influence that he already had achieved among the Vietnamese Communists can be seen in these actions.

The founding of the PCI

The creation of the PCI coincided with a violent insurrectionary movement in Vietnam. Repression by the French was brutal; Ho himself was condemned to death in absentia as a revolutionary. He sought refuge in Hong Kong, where the French police obtained permission from the British for his extradition, but friends helped him escape, and he reached Moscow via Shanghai.

In 1935 the seventh Congress of the International, meeting in Moscow, which he attended as chief delegate for the PCI, officially sanctioned the idea of the Popular Front (an alliance with the non-Communist left against Fascism)—a policy Ho had advocated for some time. In keeping with this policy the Communists in Indochina moderated their anticolonialist stance in 1936, allowing for cooperation with "antifascist colonialists." The formation of Premier Léon Blum's Popular Front government in France in the same year allowed leftist forces in Indochina to operate more freely, although Ho, because of his condemnation in 1930, was still not permitted to return from exile. Repression returned to Indochina with the fall of the Blum government in 1937, and by 1938 the Popular Front was a dead letter.

World War II and the founding of the Vietnamese state. In 1938 Ho returned to China and stayed for a few months with Mao Tse-tung at Yen-an. When France was defeated by Germany in 1940, he and his lieutenants, Vo Nguyen Giap and Pham van Dong, plotted to use this turn of events to advance their own cause. About this time he began to use the name Ho Chi Minh (He Who Enlightens). Crossing over the border into Vietnam in January 1941, the trio and five comrades organized in May the Viet Nam Doc Lap Dong Minh Hoi (League for the Independence of Vietnam), or Viet Minh; this gave renewed emphasis to a peculiarly Vietnamese nationalism.

The new organization was forced to seek help in China from the government of Chiang Kai-shek. But Chiang distrusted Ho as a Communist and had him arrested. Ho was then imprisoned in China for 18 months, during which time he wrote his famed *Notebook from Prison* (a collection of short poems written in classic Chinese, a mixture of melancholy, stoicism, and a call for revolution). His friends obtained his release by an arrangement with Chiang Fa-k'uei, a warlord in South China, agreeing in return to support Chiang's interests in Indochina against the French.

In 1945 two events occurred that paved the way to power for the Vietnamese revolutionaries. First, the Japanese completely overran Indochina and imprisoned or executed all French officials. Six months later the United States dropped the atomic bomb on Hiroshima, and the Japanese were totally defeated. Thus, the two strongest adversaries of the Viet Minh and Ho Chi Minh were destroyed.

Ho Chi Minh seized his opportunity. Within a few months he contacted U.S. forces and began to collaborate with the OSS (Office of Strategic Services, a U.S. undercover operation) against the Japanese. Further, his Viet Minh guerrillas fought against the enemy in the mountains of South China.

At the same time, commandos formed by Vo Nguyen Giap, under Ho's direction, began to move toward Hanoi, the Vietnamese capital, in the spring of 1945. After Japan's surrender to the Allies, they entered Hanoi on August 19. Finally, on September 2, before an enormous crowd gathered in Ba Dinh Square, Ho Chi Minh declared Vietnam independent, using words ironically reminiscent of the U.S. Declaration of Independence: "All men are born equal: the Creator has given us inviolable rights, life, liberty, and happiness . . . !"

All obstacles were not removed from the path of the Viet Minh, however. According to the terms of an Allied agreement, Chiang Kai-shek's troops were supposed to replace the Japanese north of the 16th parallel. More significantly, France, now liberated and under the leadership of Charles de Gaulle, did not intend to simply accept the *fait accompli* of an independent Vietnam and attempted to reassert its control. On October 6 Gen. Leclerc landed in Saigon, followed a few days later by a strong armoured division. Within three months, he had control of South Vietnam. Ho had to choose between continuing the fight or negotiating. He chose negotiations, but not without preparing for an eventual transition to war.

Ho Chi Minh's strategy was to get the French to make the Chinese in the north withdraw and then to work for a treaty with France in which recognition of independence, evacuation of Leclerc's forces, and reunification of the country would be assured. Negotiations began in late October 1945, but the French refused to speak of independence, and Ho was caught in a stalemate. In March the deadlock was broken: on his side, Ho Chi Minh allowed parties other than the Viet Minh to be included in the new government, in an attempt to gain a wider base of support for the demands made on the French; at the same time, the French sent a diplomatic mission to China to obtain the evacuation of the Chinese soldiers. This was done, and some of Leclerc's troops were also removed from Haiphong, in the north. Having secured the withdrawal of the Chinese, Ho signed an agreement with the French on March 6. According to its terms, Vietnam was recognized as a "free state with its own government, army, and finances," but it was integrated into a French Union in which Paris continued to play the key role. Twelve days later, Leclerc entered Hanoi with a few battalions, which were to be confined to a restricted area.

The agreement was unsatisfactory to extremists on both sides, and Ho Chi Minh went to France for a series of conferences (June to September 1946) and concluded a second agreement with the French government. But the peace was broken by an incident at Haiphong (November 20–23, 1946) when a French cruiser opened fire on the town after a clash between French and Vietnamese soldiers. Almost 6,000 Vietnamese were killed, and hope for an amicable settlement ended. Sick and disillusioned, Ho Chi Minh was not able to oppose demands for retaliation by his more militant followers, and the first Indochina war began on December 19.

The first Indochina war

After a few months, Ho, who had sought refuge in a remote area of North Vietnam, attempted to re-establish contact with Paris, but the terms he was offered were unacceptable. In 1948 the French offered to return the former Annamese (Vietnamese) emperor Bao Dai, who had abdicated in favour of the revolution in August 1945. These terms were more favourable than those offered to Ho Chi Minh two years earlier, because the French were now attempting to weaken the Viet Minh by supporting the traditional ruling class in Vietnam. But his policy was not successful. The Viet Minh army, commanded by Giap, was able to contain the French and

Bao Dai's forces with guerrilla tactics and terrorism, and, by the end of 1953 most of the countryside was under Viet Minh control, with the larger cities under a virtual state of siege. The French were decisively defeated at Dien Bien Phu on May 7, 1954, and had no choice but to negotiate.

The Geneva Accords and the second Indochina war. From May to July 21, 1954, representatives of eight nations—with Vietnam represented by two delegations, one composed of supporters of Ho Chi Minh, the other of supporters of Bao Dai—met in Geneva in an attempt to find a peaceful solution. They concluded with an agreement according to which Vietnam was to be divided at the 17th parallel until elections, scheduled for 1956, after which the Vietnamese would establish a unified government.

It is difficult to assess Ho's role in the Geneva negotiations. He was represented by Pham van Dong, his most faithful associate. The moderation exhibited by the Viet Minh in accepting a partition of the country despite their triumph at Dien Bien Phu and in accepting control of less territory than they had conquered during the war follows the pattern established by the man who had signed the 1946 agreements with France. But this flexibility, which was also a response to pressures exerted by the Russians and Chinese, did not achieve everything for the Viet Minh. Hanoi lost out because the elections that were to guarantee the country's reunification were postponed indefinitely by the United States and by South Vietnam, which was created on a de facto basis at this time.

North Vietnam, where Ho and his associates were established, was a poor country, cut off from the vast agricultural areas of the south. Its leaders were forced to ask for assistance from their larger Communist allies, China and the Soviet Union. In these adverse conditions Ho Chi Minh's regime became repressive and rigidly totalitarian. Attempted agricultural reforms in 1955–56 were conducted with ignorant brutality and repression. "Uncle" Ho, as he had become known to the North Vietnamese, was able to preserve his immense popularity, but he abandoned a kind of humane quality that had distinguished some of his previous revolutionary activities despite ruthless "purges" of Trotskyists and bourgeois nationalists in 1945–46.

The old statesman had better luck in the field of diplomacy. He travelled to Moscow and Peking (1955) and to New Delhi and Jakarta (1958), skillfully maintaining a balance between his powerful Communist allies and even, at the time of his journey to Moscow in 1960, acting as a mediator between them. Linked by old habit, and perhaps by preference, to the Soviet Union, but aware of the seminal role China had played in the revolution in Asia, preoccupied with using his relations with Moscow to lessen China's influence in Asia, and, above all, careful to assert Vietnamese rights, Ho Chi Minh skillfully maintained a balance between the two Communist giants. When the war was resumed, he obtained an equal amount of aid from both.

Beginning about 1959, North Vietnam again became involved in war. Guerrillas, popularly known as the Vietcong, were conducting an armed revolt against the U.S.-sponsored regime of Ngo Dinh Diem in South Vietnam. Their leaders, veterans of the Viet Minh, appealed to North Vietnam for aid. In July 1959, at a meeting of the central committee of Ho Chi Minh's Lao Dong (Worker's Party), it was decided that the establishment of socialism in the North was linked with the unification with the South. This policy was confirmed by the third congress of the Lao Dong, held shortly thereafter in Hanoi. During the congress, Ho Chi Minh ceded his position as the party's secretary general to Le Duan. He remained chief of state, but from this point on, his activity was largely behind-the-scenes. Ho certainly continued to have enormous influence in the government, which was dominated by his old followers Pham van Dong, Truong Chinh, Vo Nguyen Giap, and Le Duan, but he was less actively involved, becoming more and more a symbol to the people. His public personality, which had never been the object of a cult comparable to that of Stalin, Mao, or even Tito, is best symbolized by his popular name, Uncle Ho. He stood for the essential unity of the divided Vietnamese family.

Aid to the Viet Cong

This role, which he played with skill, did not prevent him from taking a position in the conflict ravaging his country, especially after American air strikes against the North began in 1965. On July 17, 1966, he sent a message to the people ("nothing is as dear to the heart of the Vietnamese as independence and liberation") that became the motto of the North Vietnamese cause. On February 15, 1967, in response to a personal message from U.S. President Lyndon Johnson, he announced: "We will never agree to negotiate under the threat of bombing." Ho lived to see only the beginning of a long round of negotiations before he died on September 3, 1969.

The removal of this powerful leader undoubtedly damaged chances for an early settlement. His funeral was held in Hanoi's Ba Dinh Square, where, 14 years earlier, he had proclaimed the resurgence of the state of Vietnam. Le Duan, Ho's successor as party head, read his remarkable last will and testament. In it, Ho spoke of the Vietnamese people as his "nephews," reflecting a peculiarly Vietnamese relationship suggested by the words *bac* (a paternal uncle even more highly respected than a father) and *nghia* (a term implying self-sacrifice and the necessity of helping one's kin). Equally characteristic are his references to the "plains and mountains" of Vietnam and his insistence that Vietnam must again be united under one "roof." These phrases are suggestive of the realities of peasant life and its relationship to the land and hearth, which alone give meaning to Vietnamese national feeling. But he links this national feeling to the hope for a "worldwide revolution," which is again characteristic of the man who devoted 50 years of his life to Marxism–Leninism.

Ho Chi Minh left no family. He is not known to have had a wife, and his brothers and sisters died before him. His death caused no great changes either in the conduct of the Indochina war or in the power structure of his country. The collective leadership he fostered during his lifetime seems to have filled the gap left by his death.

Ho Chi Minh's importance

Among 20th-century revolutionaries, Ho waged the longest and most costly battle against the colonial system of the great powers. One of its effects was to cause a grave crisis in the national life of the mightiest of capitalist countries, the United States. As a Marxist, he stands with the Yugoslav leader Tito as one of the progenitors of the "national Communism" that developed in the 1960s, and (at least partially) with Communist China's Mao Tse-tung in emphasizing the role of the peasantry in the revolutionary struggle.

Most of Ho Chi Minh's writings are collected in the two-volume *Selected Works*, published in Hanoi in 1960, in the series of Foreign Language Editions. (J.La.)

BIBLIOGRAPHY. The best known biography of Ho Chi Minh is that of the French journalist JEAN LACOUTURE, *Hô Chi Minh* (1967; Eng. trans., 1968). A short, well-written account is that by the Pulitzer Prize-winning American journalist, DAVID HALBERSTAM, *Ho* (1971). Two somewhat more scholarly works are CHRISTIANE PASQUEL RAGEAU, *Ho Chi Minh* (1970); and N. KHAC HUYEN, *Vision Accomplished? The Enigma of Ho Chi Minh* (1971). JEAN SAINTENY, *Face à Ho Chi Minh* (1970), tells of Ho's life from the standpoint of a noted French diplomat who was a close personal friend of Ho and had many long conversations with him, which he reports in the book. The Committee for the Study of the History of the Vietnamese Worker's Party has compiled an English-language book, *Our President Ho Chi Minh* (1970), which has an introduction by North Vietnamese Prime Minister PHAM VAN DONG. HO CHI MINH, *Selected Works*, 4 vol. (1960–62), has also been published in English by the Vietnamese government, as has his *Prison Diary*, 2nd ed. (1965). Finally, CHARLES FOURNIAU, *Ho Chi Minh, notre camarade* (1970), has assembled recollections of Ho Chi Minh from members of the French Communist Party.

Hockey (Field)

Field hockey is an outdoor game played with a hard ball by two opposing teams of 11 players each, using hooked

or bent sticks with which each side attempts to drive the ball into the other's goal. Often called simply hockey, the term field hockey is used to distinguish it from ice hockey (*q.v.*). It is played much more widely throughout the world than the latter game.

This article is intended for the general reader who may have no knowledge of the sport. Therefore, in addition to tracing its history and outlining its present status, the article is designed to help a spectator understand a game he may be watching. For information on the specific rules or how to play the reader should consult the works listed in the bibliography.

History. Hockey is believed to be the oldest game in the world played with a stick and ball and, according to historians, had its roots in Persia around 2000 BC. The game was acquired by the Greeks, who in turn passed it on to the Romans. A discovery made at Athens in 1922 gives reason to believe that a form of stick game came from the East. This was a bas-relief found in a wall built by Themistocles (*c.* 514–449 BC), which depicts six youths taking part in a game resembling hockey and shows what is termed a face-off, or bully, in the modern game, but with the hooked sticks pointed downward instead of upward. Traces of a sort of stick game played by the Aztec Indians have also been found, and evidence indicates that probably most of the American Indian tribes played a rough stick game for centuries. The sport can also be identified with early games called shinty, hurling, and bandy. During the Middle Ages the game was played in France under the name *hoquet* (the French word for shepherd's crook). When the British took it up it is said this word was anglicized to hockey. In England and other countries, at one time, it was forbidden along with certain other sports because it interfered with archery, a part of the national defense.

Modern hockey

About 1875 a game resembling modern hockey began to be played in England. No goal could be scored if hit from a distance of more than 15 yards from the goal, but players did not for a time realize the need of a definite, marked-out striking circle. A landmark in the progress of hockey was the formation of the famous Wimbledon Club in 1883. A few more clubs soon came into being in the London area and the game spread throughout England. The real birthday of modern hockey was, however, Jan. 18, 1886, the date of the formation in London of the Hockey Association and of the adoption of the striking circle and other rules. In 1895 the first "international" match was played, in which England defeated Ireland 5–0. About 1900 the need for a body to frame the rules for the entire British Isles was realized and the International Hockey Board consisting of England, Ireland, Wales, and Scotland, was formed.

The modern game spread rapidly, particularly on the Continent and in India, due there largely to the influence of the British Army. Although not played to the same extent as association football (soccer), the game is popular in many countries. The national associations of these countries, including those making up the rest of the British Commonwealth and also the United States, became members of the Fédération Internationale de Hockey (FIH, founded 1924). Hockey was included in the Olympic Games in 1908 and 1920, and in each Olympics from 1928. Great Britain, after an absence of several decades, again competed in hockey in the Olympic Games held in London in 1948, and a close working arrangement was instituted at that time between the International Hockey Board and the FIH whereby observance of the same playing rules by all countries was assured. The first Western Hemisphere country to compete in Olympic hockey was the United States, in 1932.

The modern game emanates from England and the rules are to a large extent determined there. Until recent years India has dominated Olympic and other international competition, but more recently Pakistan and a number of non-Asian nations have improved greatly. The leading countries in this sport in the second half of the 20th century have been India, The Netherlands, Great Britain, Germany, and Pakistan. Australia, New Zealand, Spain, Japan, and a number of African countries, however, have developed great proficiency and, with interest in hockey increasing everywhere, the standard of play has risen and international matches tend to be more closely contested. In some countries (*e.g.*, India and Pakistan) hockey is the national sport. Sports writers covering the Olympic Games of 1932 at Los Angeles voted the performance of the Indian team the outstanding exhibition of skill in any sport.

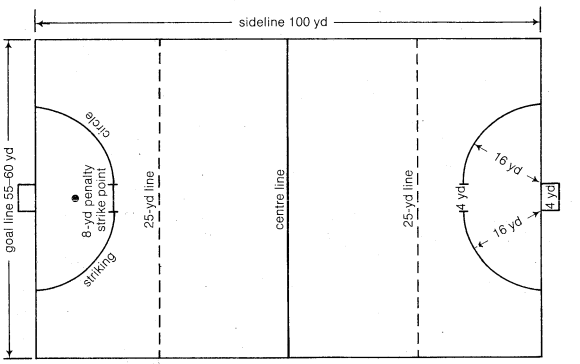
Styles of play and rules. The Indian style of play features short, controlled passes as opposed to the European type, in which the tendency is to hit the ball about the field more. The short-toed stick, developed in India and Pakistan, is now used by most players. By using this stick and employing a left-handed grip well to the rear of the stick they are able to reverse the stick (*i.e.*, play the ball with the toe of the stick pointing downward) without changing the grip and thus attain remarkable control of the ball.

The game has been progressing rapidly in the Western Hemisphere, where formerly only the United States was a member of the FIH. In 1967 it was played for the first time in the Pan-American Games, with eight countries competing. The order of finish was Argentina, Trinidad-Tobago, the United States, and Canada. The Olympic rules now provide that the winner of the Pan-American Games shall be one of the countries included in the Olympic competition, which is limited to 16 teams.

A version of the game to be played indoors has been devised, utilizing special rules and a limited number of players.

Playing the game

The game is played by two teams of 11 players on a rectangular ground (see the accompanying diagram). The goals are four yards wide and seven feet high. For a goal (which counts one point) to be scored the ball must go into the goal and while within the striking circle must have been touched by the stick of an attacker. The ball originally was a cricket ball (cork centre, string-wound, and covered with leather), but plastic balls are also approved. The stick is usually 36 to 38 inches long and weighs 18 to 28 ounces. The striking surface is flat on the left side only and only the flat surface may be used.



Field hockey playing field for men.

The usual composition of a team is five forwards, three halfbacks, two fullbacks, and a goalkeeper. A game consists of two halves of 35 minutes each, with a 5-minute intermission. Under international rules, no substitution is allowed, but in the United States, Bermuda, and some other countries it is permitted by agreement between the teams. A time-out is called only in case of injury. The goalkeeper wears heavy pads and while in the striking circle is allowed to kick the ball or stop it with the foot or body. All other players, however, may stop the ball with the hand or stick only. They wear no particular protective gear.

Play is started (and restarted after a goal is scored and after half-time) by a face-off, or bully, in the centre of the field. Two players, one from each team, face each other with the ball on the ground between them. After alternately tapping the ground and then his opponent's stick three times, each player tries to strike the ball, thus putting it into play. Until the ball is in play all other players must be onside (*i.e.*, between the ball and their own goal

line). There are various provisions for putting the ball in play in case it goes off the field. Most fouls are penalized by giving the opposition a free hit from the point of infraction, but more severe penalties (penalty corners and penalty strokes) can be awarded for fouls by the defense in the vicinity of the goal. An umpire—there is one for each half of the ground—may decline to enforce any penalty if, in his opinion, enforcing it would give an advantage to the offending team. Minor infractions, where in the opinion of the umpire no advantage results to the offender, are not penalized.

The following is a brief explanation of some of the fouls that might seem puzzling to the uninitiated spectator. The off-side rule, which is similar to that of soccer, is designed to prevent a player from getting an advantage by staying up the field ahead of the ball and ahead of less than three members of the opposing team. Raising the heavy stick above the shoulder while playing the ball is obviously dangerous. Advancing the ball with the hand instead of stopping it dead is contrary to the purpose of the game, as is stopping it with the body or foot. Finally, the obstruction rule: in most sports one of the most effective ways of preventing an opponent from getting at the ball (or other playing object) is to interpose the body between the ball and the opponent. Since the hockey stick is heavy and is often used with considerable force, to expect a player to defend the ball in that manner would surely invite injuries. Accordingly, if an opposing player is within striking distance of the ball, it is a foul to obstruct him, intentionally or otherwise.

Women's field hockey

Field hockey is also played extensively in many countries by women. Although women do not compete in this sport in the Olympic Games, they do engage in international competition. The international governing body, the International Federation of Women's Hockey Associations, was formed in 1927. Teams representing England have, for the most part, dominated international competition. The game was introduced in the United States in 1901 by Constance M.K. Applebee, a member of the British College of Physical Education, and thereafter became the most popular outdoor team sport among women in the United States. It is played in schools, colleges, and clubs, and sectional and national tournaments are held.

The rules used by women are similar to those used by men, the more important exceptions including: the striking circle measurement is 15 yards; the dotted lines are 5 yards from the sidelines; and playing time does not exceed two 30-minute halves.

BIBLIOGRAPHY. R.Y. FISON and R.L. HOLLANDS, *Hockey* (1951); NORMAN BORRETT and PEGGY LODGE, *Hockey for Men and Women* (1955); J.T. LEES, *Field Hockey for Girls* (1942), and with B. SHELLENBERGER, *Field Hockey for Players, Coaches and Umpires* (1957); MARJORIE POLLARD, *Hockey for All* (1957); A.L. DELANO, *Field Hockey* (1966). For information on coaching and umpiring, see the HOCKEY ASSOCIATION, *Hockey Coaching* (1968); G. SINGH, *How to Play Hockey* (1959), and *Hockey Umpiring* (1963). Rules of the game may be found in the INTERNATIONAL HOCKEY RULES BOARD, *Rules of the Game of Hockey* (annual), and *Official Field Hockey Guide for Women and Girls* (annual), also includes articles for coaches and players, notes on umpiring, and bibliography. For current articles and statistical data, see the magazine *World Hockey*.

(H.K.Gr.)

Hogarth, William

An invigorating satirist and critic of art and society, William Hogarth made significant contributions in the fields of portraiture, genre, and historical painting. His fame rested, however, on his comic narrative engravings, which, until relatively recently, obscured his great originality as a painter and theorist. His career marked a new departure in English art. Fiercely independent and chauvinistic, he became the focal point of artistic opposition to the conservative, academic tradition at home and the first native-born artist to inspire notable admiration abroad. Pugnacious and opinionated, he incited controversy; and while his prolific talents never went unrecognized, a series of disappointments embittered his later life and inhibited the full expression of his mature powers as a painter. Hogarth attached enormous importance to his public career, while his private life has remained in a shadow. Apart from what his works reveal of their author, an assessment of his character and experience must rely chiefly on two sources: autobiographical notes written shortly before his death in 1764 in a spirit of self-vindication and the equally selective comments of an engraver, George Vertue, which betray the grudging admiration that characterized contemporary criticism of Hogarth.

By courtesy of the trustees of the Tate Gallery, London



Hogarth, self-portrait, "The Painter and His Pug," oil on canvas, 1745. In the Tate Gallery, London. 91.17 cm × 69.85 cm.

Youth and early career. Hogarth, the only son of Richard Hogarth, a minor classical scholar who made a living as a schoolmaster and publishers' hack, was born on November 10, 1697, in London and grew up with two younger sisters, Mary and Ann, in the heart of the noisy, teeming city. Richard's evident abilities as a classicist brought him scant reward but provided an educated and industrious, if not prosperous, home. Looking back on this period in his autobiographical notes, Hogarth dwelt almost exclusively on his father's shabby treatment at the hands of printers, booksellers, and wealthy patrons. The notes were compiled in a jaundiced mood, but it is evident that he was oversensitive to neglect of parental worth. Apart from confirming his distrust of learning, his resentment at his father's disappointing experiences fostered the boy's self-assertiveness and independence of character.

As a boy with little inclination to scholarship but gifted with a lively perception of the busy world around him, he enjoyed mimicking and drawing striking characters he observed, interests that were encouraged by visits to a local painter's workshop. While not discouraging his artistic inclinations, his father, Hogarth later complained, could do little more "than put me in a way of shifting for myself." He consequently sought the security of a solid craftsman's training and became apprenticed, at about the age of 15, to Ellis Gamble, a silversmith. Hogarth presumably moved to his master's house, where he learned to engrave gold and silver work with armorial designs—in his own phrase, the "monsters of heraldry." Valuable years lost on, what Vertue aptly termed, "low-shrubb instructions" had crucial bearing on Hogarth's

Apprenticed to a silversmith

subsequent development. Apart from the insecurity they bred, Hogarth's frustration with his training led him to exploit unorthodox methods of self-instruction in order to make up for lost time. His great originality and flexibility as an artist owed much to this pragmatic and unconventional approach to his career.

Hogarth's years of apprenticeship were by no means devoted exclusively to hard work, however. Sociable and fond of fun, a keen and humorous observer of human behaviour, with a special love of the theatre and shows of all kinds, he was evidently a gay companion. Never prudish, he knew the exuberant life of the London streets, bawdy houses, fairs, and theatres at first hand and derived from them a fertile appreciation of the vitality of popular tradition. At the same time, he felt drawn to the coffeehouses and taverns frequented by writers, musicians, actors, and liberal professionals, forming lasting friendships in such lively intellectual circles. His sympathies rested with the middle classes and, specifically, with the critical, enlightened element—rational, tolerant, and humanitarian—that played such a prominent role in the cultural life of Hanoverian England.

George I had been king for six years when Hogarth set up shop on his own at the age of 23, resolving to escape the rigid limitations of his trade. He began by attending a private drawing school in St. Martin's Lane, where he joined other students drawing from casts and live models. He had a natural distaste for copying, however, likening it to emptying water from one vessel into another, and this instinctive rejection of formal training, combined with a natural waywardness, convinced him that the best method of learning to draw lay in direct attention to actual life. An intuitive realist, primarily concerned with expressive rather than formal values, he developed a kind of visual mnemonics: "the retaining in my minds eye without drawing on the spot whatever I wanted to imitate." From close observation of the everyday scene ("Nature" for this confirmed city dweller), Hogarth trained his unusual visual memory until, unlike the majority of artists, he could dispense with preliminary studies, committing his ideas directly to paper or canvas. This inspired improvisation, adopted partially as a means of reconciling work and pleasure, was supplemented by a formidable knowledge of the European tradition in art, acquired through familiarity with a vast range of reproductive engravings. Meanwhile, he earned his living as a copper engraver, executing trade cards, tickets, and book illustrations that assumed an increasingly authoritative stamp as fresh influences transformed the decorative mannerisms of his apprenticeship. His growing success as an illustrator brought Hogarth little satisfaction, for it entailed unwelcome dependence on the booksellers who had exploited his father; he later insisted that engraving "did little more than maintain myself in the usual gaities of life but (was) in all a punctual paymaster." He had long been an admirer of Sir James Thornhill's fluent, if provincial, adaptation of the late Baroque style, and in 1724 he joined a free drawing school, newly opened in Thornhill's house. It was the start of a critical association. Holding the official post of serjeant painter to the King and being the first knighted English-born artist, Thornhill in his career affirmed the vitality of native art and the social respectability of the artist. Hogarth cared passionately about both, primarily for personal reasons but also because he believed in art as a vital creative force in society. He despised the connoisseurs' exclusive admiration for the old masters and their snobbish prejudice in favour of foreign artists. In his first major work, published independently of the booksellers in 1724, "Masquerades and Operas," Hogarth attacked the degeneracy of contemporary taste and expressed attitudes that were vigorously sustained throughout his life. Boldly questioning the standards of an influential clique that was supported by the Earl of Burlington, an influential art lover and patron, Hogarth's first blow in a private war with the connoisseurs was shrewdly designed to appeal to his hero, Thornhill, who was himself, at this time, suffering the hostile effects of Burlington's Neoclassical revival. Thus, Hogarth made powerful enemies at the start of his career, and when they retaliated around 1730 by nullifying royal interest in his early work, he was cruelly disappointed. Indeed, despite his own intransigent frankness, Hogarth was always discouraged and offended when his opponents hit back.

Development of visual mnemonics

Association with Sir James Thornhill

A lawsuit he brought in 1728 against Joshua Morris, a tapestry weaver, throws eloquent light on his susceptibilities. The details of the case reveal that, by the age of 30, Hogarth felt sufficiently confident of his abilities to embark on a painting career. Morris failed to share this confidence and rejected a painting he had ordered on grounds that it was not finished. Hogarth indignantly sought and obtained public vindication with the help of professional witnesses, including Thornhill. Their testimony was amply justified by his first dated painting, "The Beggar's Opera" (1728), a scene from John Gay's popular farce on lowlife, which emphasized Hogarth's prevailing interests: his involvement with the theatre and with down-to-earth, comic subjects. Closely attentive to realistic detail, he recorded the scene exactly as it appeared to the audience and included portraits of the principal actors and spectators. He thus anticipated both his later narrative paintings and the small, informal group portraits, or "conversation pieces," that occupied him in the years immediately after this auspicious debut.

"The Beggar's Opera"

Encouraged by his success and, possibly, also by his close friendship with Thornhill's son John, Hogarth eloped in March 1729 with Thornhill's daughter Jane. The marriage proved stable and contented, though childless, but Thornhill initially showed little enthusiasm for the match; and the couple took lodgings in South Lambeth. There Hogarth met Jonathan Tyers, who was planning to re-open pleasure gardens at Vauxhall. The project naturally appealed to Hogarth, and he encouraged Tyers to invite artists to decorate the pavilions with genre scenes. Although his direct involvement was minimal, Hogarth's cheerful sympathies influenced the whole scheme, especially the subject matter.

Reputation and success. A few months after Hogarth's marriage, Vertue remarked on his public success with "conversations," and in the next few years, these small paintings, which acknowledged a great debt to the early 18th-century painter Antoine Watteau and the elegance of French Rococo art, brought Hogarth an appreciative and wealthy clientele. For a particularly important patron, the Duke of Richmond, he painted "Conquest of Mexico" (1731), a brilliant and charming portrayal of the amateur performance of Dryden's *Indian Emperor* before the royal children. His deft and fastidious handling of the paint and graceful disposition of the child players in relation to their audience are quite unmatched in English art of the period and demonstrate Hogarth's masterly control of the painting medium at the very outset of his career.

Though he displayed remarkable energy at the time, Hogarth quickly tired of these little paintings, which involved numerous portraits for relatively poor remuneration. For his own enjoyment he began, while still executing "conversations," to record humorous scenes from everyday life. The crowded canvas of "Southwark Fair" (1733) captures the noisy and exuberant vigour of a popular festival and shows Hogarth feeling his way toward a completely new kind of narrative art based on vivid appreciation of contemporary life. Friends he made in the theatrical world, the actor-manager David Garrick and writer Henry Fielding, shared his enthusiasm for honest naturalism in art. The theatre continued to play a central role in Hogarth's life, not only supplying rich enjoyment and visual material for his paintings but, in a deeper sense, shaping his ideas on the scope and purpose of art. Literary and theatrical associations governed his approach to his "modern moral subjects" initiated by "A Harlot's Progress" (1731–32), painted and engraved in six parts. As he explained, in consciously Shakespearean terms:

"A Harlot's Progress"

> Subjects I consider'd as writers do. My Picture was my Stage and men and women my actors who were by Mean of certain Actions and express[ions] to Exhibit a dumb shew.

Like his great predecessor, the 16th-century Flemish

painter Pieter Bruegel the Elder, Hogarth wanted to extract entertaining and instructive incidents from life. In telling the story of a young country girl's corruption in London and her consequent miseries, he not only ridiculed the viciousness and follies of society but painted an obvious moral. The engravings were aimed at a wide public, and their tremendous success immediately established Hogarth's financial and artistic independence. He was henceforth free, unlike most of his colleagues, to follow his own creative inclinations. Private subscriptions to the prints brought Hogarth over £1,200, far more than he could hope to earn annually from painting, but, even so, numerous pirated editions absorbed further profits. To safeguard his livelihood from unscrupulous printers, he fought to obtain legislation protecting artist's copyright and held back the eight-part "Rake's Progress" until a law of that nature, known as the Hogarth Act, was passed in 1735.

In an informal moment in May 1732, at the close of a merry evening's drinking, Hogarth embarked down the Thames with four equally inebriated companions on a madcap tour of Kent. They spent five lighthearted days roistering about the countryside and produced an amusing record of this "peregrination" for their friends, in the form of a burlesque on the continental excursions of the gentry.

In the following year Hogarth moved into the house in Leicester Fields that he was to occupy until his death and, with a teasing glance at the art pundits, installed a gilded head of the 17th-century painter Sir Anthony Van Dyck over his door. Too little a flatterer to become a fashionable portrait painter, with provoking gestures of this kind as well as assiduous newspaper publicity, Hogarth strenuously cultivated his public image, never missing an opportunity of self-advertisement. His critics called him conceited and remarked with distaste on his astute business tactics.

Historical and portrait painting. After Thornhill's death, in 1734, Hogarth re-established his drawing school on a cooperative basis, and it became an important arena for artistic discussion and experiment. He also was evidently tempted to assume Thornhill's mantle as a painter in the grand manner. "Before I had done anything of much consequence in this walk," he recalled, disingenuously dismissing the two "Progresses," "I entertained some hopes of succeeding in what the puffers in books call 'the great style of history painting.'" In 1735, in line with the humanitarian concern that occupied enlightened opinion of the day, he was elected a governor of St. Bartholomew's Hospital; and he seized this opportunity to decorate the main staircase with two large religious works, "Pool of Bethesda" and "The Good Samaritan." In abandoning comic narrative and genre for history painting, he was generally held to have overreached himself, though Vertue conceded that the results were "by everyone judged to be more than could be expected of him." Modern critics have tended to endorse this opinion. Hogarth himself deeply resented this attitude and attached great importance to his historical works. He wanted desperately to be taken seriously as a painter and realized that the popularity of his topical engravings interfered with his ambitions. His comic narrative pictures had broken radical new ground in English art but stereotyped him as a caricaturist. Conscious of his originality and encouraged by Fielding's distinction between comic realism and the grotesque, Hogarth claimed the respect traditionally paid to the old masters. Without this recognition, his successes appeared meaningless to him, and he fought ceaselessly to confound his critics and establish himself as the leader of a vigorous, new school of English painting.

Paintings for St. Bartholomew's

Around 1740 he turned once again to painting portraits, chiefly of middle class sitters, and retrospectively remarked, comparing himself to Rembrandt with less modesty than justice, that these were said "by some Nature itself by others exicrable." He derived special enjoyment from painting the full-length, seated portrait of his friend, the philanthropist Captain Thomas Coram—a compelling and deeply sympathetic image that injected the dead aristocratic tradition with forthright realism and carried far-reaching implications for European portraiture. Hogarth, well aware of its importance, judiciously placed it on semipublic display at the Foundling Hospital, a benevolent institution for orphan children established by Coram in 1739. From the start Hogarth played a very active role in the affairs of this charitable venture, and when the buildings were completed in 1745, he persuaded a group of fellow artists to join him in contributing paintings as edifying decoration. Their cooperative effort produced the first public exhibition of contemporary art in England and was a vital step toward the foundation of the Royal Academy in 1768.

Self-portrait

The famous self-portrait of 1745, a year that marked, in many ways, the high point of Hogarth's career, was also an artistic manifesto. He mischievously juxtaposed his own blunt and intelligent features with those of his sturdy pug-dog, Trump, and placed volumes of the great English writers William Shakespeare, John Milton, and Jonathan Swift beside a palette inscribed with the sinuous "line of beauty," his shorthand symbol for the variety, intricacy, and expressiveness of Nature. In the same year he published the long-announced prints of "Marriage à la Mode," censuring the marriage customs of the upper classes. He had completed the paintings in May 1743, when he made a brief visit to Paris to secure French engravers for the plates, as he wished them to be particularly elegant. Vertue implied that the extra cost entailed in this venture dissuaded Hogarth from repeating the practice, though he probably chose to execute his own plates from an understandable impatience with "translators." On his return in June, he arranged an auction of his earlier serial and narrative paintings. The ticket of admission, "The Battle of the Pictures," depicted a contest between the "ancients" (admirers of old pictures regardless of their merits or state of preservation) and the "moderns" (himself), underlining the defiant spirit in which he viewed the affair. As a further thrust at the "lovers of dark pictures," he excluded dealers and artists and conducted the bidding by the clock, two eccentric circumstances that probably contributed to the extremely low prices fetched. Hogarth interpreted the sale as a stinging defeat, which, combined with unenthusiastic notice of his Foundling picture, caused him to withdraw from painting temporarily.

Return to prints. Apart from a gratifying commission for a large history piece, which he won from the lawyers of Lincoln's Inn (one of the four legal societies and schools in London), and a further visit to Paris in 1748, Hogarth concentrated for the next few years on simple, didactic prints aimed at an unrefined public, executed from drawings not paintings. "Beer Street," "Gin Lane," and "Four Stages of Cruelty" (1751) he cut deliberately crudely on wood blocks to make them cheaper and facilitate a wide distribution. "Industry and Idleness" (1747) contains, in addition to its obvious moral message, a good deal of self-dramatization, depicting the virtuous apprentice made good in a hostile world. In these years Hogarth's uncertainty and frustration expressed themselves in a number of unfinished paintings. In several spontaneous sketches, he succeeded where he had failed in his heroic pictures and synthesized dynamic elements of the 17th-century Baroque style with an uncompromising realism and fully expressive handling of the paint. Unregarded in his lifetime, it is only in the wake of the 19th-century Impressionist movement that such sketches have received serious attention.

In 1751 Hogarth again organized a sale even more bizarre than the first. When the six paintings for "Marriage à la Mode" went to the only bidder for £120, Vertue complacently observed that Hogarth, in anger and mortification, took down the Van Dyck sign. He retreated into aggrieved isolation, pursuing his philanthropic interests but adopting, in public, a defiant and defensive pose that involved him in increasingly rancourous debate on artistic matters. He expounded his own theories in *The Analysis of Beauty* (1753), combining practical advice on painting with criticism of the art establishment. He expressed his belief in the "beauty of a composed intricacy

The Analysis of Beauty

of form," which "leads the eye a kind of chace" and advocated variety, irregularity, movement, and exaggeration in the interests of greater expressiveness. Though his ideas were respectfully received, especially on the Continent, the book inspired much adverse comment from his opponents. He even quarrelled with his colleagues in his academy over a proposal to adopt the French academic system, which Hogarth opposed as he did rules and rigidity of any kind. Yet he must have been painfully aware that the future now lay with Joshua Reynolds and other artists unsympathetic to his ideals.

Appointment as serjeant painter

His large "Election" series (1754–58), painted with elaborate care, was a last attempt to prove the dignity of "comic history painting," and thereafter he painted little of importance. His appointment as serjeant painter to George III, contrived in 1757, revived some interest in portraiture, but his last years, when he probably suffered considerable ill health, were dominated by the acrimony attendant on his painting "Sigismunda" (1759) and his print "The Times, I" (1762). The first represented a characteristic bid by Hogarth to outdo an old master painting recently sold for £400. His pride suffered a harsh blow when his patron, expecting a light genre subject, rejected the finished picture.

An ill-advised departure into political satire, "The Times, I" attacked William Pitt the Elder and the extremely popular war party that supported the war with France. Hogarth claimed it was executed "to stop a gap in my income," but general irritability and a desire to pose as a persecuted heretic and lone defender of truth lay behind this reckless challenge to public opinion. The storm of abuse it excited wounded him deeply and soured his last years. Obsessive to the last, a few months before his death on October 26, 1764, he executed an engraving sardonically titled "Tail-Piece" or "The Bathos," in which he sombrely depicted the demise of his own artistic world. In a sense it was prophetic, for as the 19th-century English painter Constable rightly remarked, "Hogarth has no school, nor has he ever been imitated with tolerable success." His immediate influence had been more strongly felt in literature than painting, and after his death it was significantly the Romantics, many of whose ideas Hogarth had anticipated, who first recognized his greatness. Though never neglected, Hogarth was chiefly remembered for his satiric engravings, and like that other lonely pioneer, the 19th-century painter J.M.W. Turner, the implications of his work were better understood on the Continent than in England.

MAJOR WORKS

PAINTINGS: "The Beggar's Opera" (1728; Tate Gallery, London); "Captain Woodes Rogers and His Family" (1729; National Maritime Museum, Greenwich, London); "The Wedding of Stephen Beckingham and Mary Cox" (1729–30; Metropolitan Museum of Art, New York); "A Musical Party" (*c.* 1730; Fitzwilliam Museum, Cambridge); "The Fountaine Family" (*c.* 1730; Philadelphia Museum of Art, Philadelphia); "A Fishing Party" (*c.* 1730; Dulwich College Picture Gallery, London); "The Conquest of Mexico (The Indian Emperor)" (1731; Earl of Ilchester's Collection, London); "The Cholmondeley Family" (1732; Marquis of Cholmondeley's Collection); "A Rake's Progress" (eight scenes, begun 1732; Sir John Soane's Museum, London); "Southwark Fair" (1733; Lady Oakes' Collection); "The Distressed Poet" (*c.* 1735; City Museum and Art Gallery, Birmingham, England); "The Good Samaritan" (1735; St. Bartholemew's Hospital, London); "The Four Times of the Day" (*c.* 1736; Viscount Bearsted and the Earl of Ancaster Collection); "Captain Thomas Coram" (1740; Foundling Hospital, London); "Miss Mary Edwards" (*c.* 1740; Frick Collection, New York); "The Graham Children" (1742; National Gallery, London); "Marriage à la Mode" (six scenes, 1743; Tate Gallery, London); "Garrick in the Character of Richard III" (1745; Earl of Feversham Collection); "Mrs. Elizabeth Salter" (1744; National Gallery, London); "Self-Portrait" (1745; National Gallery, London); "Lord George Graham in His Cabin" (1745; National Maritime Museum, Greenwich, London); "Moses Brought Before Pharoah's Daughter" (1746; Foundling Hospital, London); "An Election" (four scenes, 1754–58; Sir John Soane's Museum, London); "The Ascension" (1756; St. Mary Redcliffe, Bristol); "Hogarth's Servants" (*c.* 1758; National Gallery, London); "Picquet, or Virtue in Danger" (1758–59; Albright-Knox Art Gallery, Buffalo); "The Shrimp Girl" (*c.* 1759; National Gallery, London).

ENGRAVINGS: "Masquerades and Operas" (1724; Burlington Gate); "A Harlot's Progress" (1731–32); "A Rake's Progress" (1735); "The Strolling Actresses Dressing in a Barn" (1738); "Marriage à la Mode" (1745); "Industry and Idleness" (1747); "Beer Street" and "Gin Lane" (1751); "The Four Stages of Cruelty" (1751); "The Invasion" (two plates, 1756); "The Bench" (1758); "The Times, I" (1762).

BIBLIOGRAPHY. WILLIAM HOGARTH, *The Analysis of Beauty with the Rejected Passages from the Manuscript Drafts and Autobiographical Notes*, ed. by J. BURKE (1955), also includes a critical essay on Hogarth's aesthetics; J. NICHOLS, *Biographical Anecdotes of William Hogarth*, 2nd ed. (1782), a gossipy outline based partly on contemporary reminiscences; J. IRELAND, *Hogarth Illustrated*, 2 vol. (1791), a commentary on the prints supplemented by a third volume in 1798 giving a tidied version of the *Notes;* J.B. NICHOLS (ed.), *Anecdotes of William Hogarth Written by Himself* (1833), a useful catalog of the prints; A. DOBSON, *Hogarth*, 7th ed. (1907), the most reliable biography to date and an important bibliographical source; S.E. READ, *A Bibliography of Hogarth Books and Studies, 1900–1940* (1941); F.D. KLINGENDER (ed.), *Hogarth and English Caricature* (1944), with valuable plates; R.E. MOORE, *Hogarth's Literary Relationships* (1948); A.P. OPPE, *The Drawings of William Hogarth* (1948), a critical catalogue raisonné; R.B. BECKETT, *Hogarth* (1949), the only study of the paintings; E.K. WATERHOUSE, *Painting in Britain, 1530–1790* (1953), a useful summary of Hogarth and contemporary figures; P. QUENNELL, *Hogarth's Progress* (1955), a light and chatty account; F. ANTAL, *Hogarth and His Place in European Art* (1962), an analysis of Hogarth's cultural milieu and sources with a Marxist bias; R. PAULSON, *Hogarth's Graphic Works*, 2 vol. (1965), an excellent catalog with an illuminating introduction and bibliography; see also Paulson's *Hogarth: His Life, Art, and Times*, 2 vol. (1971); catalog of the *William Hogarth Exhibition at the Tate Gallery, December 1971/February 1972* (1971); and RONALD PAULSON, *Hogarth: His Life, Art, and Times*, 2 vol. (1971).

(S.E.B.)

Hōjō Family

The Hōjō family, as hereditary regents to the military government of Japan, were the actual rulers of Japan from 1199 to 1333. During that period, nine successive members of the Hōjō family held the regency. The Hōjōs took their name from their small estate in the Kanogawa Valley in Izu Province.

Hōjō Tokimasa (1138–1215), the first known member of the family, was charged by the Japanese ruler Taira Kiyomori with the co-wardenship of the exiled Minamoto Yoritomo in 1160. In 1180, however, when Yoritomo rallied the armed men of the Kantō, a region in Central Japan, against Taira rule, Tokimasa fought with him. Yoritomo acquired all power in Japan by 1189 and ruled as shogun (military commander); Tokimasa became the warden of Kyōto, while his daughter Masako married Yoritomo, with whom she had long had a liaison. At Yoritomo's death in 1199 Tokimasa became the guardian of the heir Yoriie and in effect regent, although Masako governed in the name of her son. The Hōjō family kept in good order and improved the simple but effective machinery of rule that Yoritomo had established. Yoritomo had received permission from the Emperor to place his own men as constables (*shugo*) and tax collectors (*jitō*) in each province. These appointees were responsible to the Samurai *dokoro*, or private military staff of the shogun, at Kamakura. The staff was headed by the *shikken*, or regent to the shogun. Thus, this office controlled the law, the peace, and the revenues of Japan, and the Hōjō family came to monopolize the office of *shikken* and to make it hereditary among them.

Rise to power

By 1247, when members of the house and clan held, through appointment, dominion over half the provinces of Japan, Hōjō rule tended to become authoritarian, and the regency was run not from its titular office but from Hōjō headquarters as a family council. This assumption of power, beginning with Tokimasa, was not difficult because the armed class did not wish to relinquish the peace, profits, and stability the *bakufu* (military government) had brought it. They were reluctant to permit the heir Yoriie, a youth of uncertain temper and strong ap-

petites, to become shogun. Yoriie attempted the murder of Tokimasa but was himself exiled and killed. When the remaining heir, Sanetomo, was murdered (1219) the last impediment to Hōjō domination was gone. The final accretion of Hōjō power came in 1221, when the emperor Toba II raised the Taira of western Japan against the Hōjō. The revolt (Shōkyū no Ran) not only failed but in its failing the Hōjō were able to confiscate thousands of estates and place them in the hands of landless adherents and friends. Many landless warriors, created by the litigious system of family inheritance in Japan, had little love for the Hōjō but less for hunger and dispossession. Their number, as it rose and fell, was an indication of the stability of the *bakufu*, and until the late 13th century the Hōjō kept their numbers small. The first three Hōjō regencies—Yoshitoki, who succeeded Tokimasa in 1205, was murdered in 1224 and replaced by his brother Yasutoki (1183–1242)—were the apex of capable feudal rule in Japan. Dependable cadastral records were created in 1222–23. In 1232 a brief and workable code (Jōei shikimoku) for the conduct and regulation of the armed class in a feudal society was promulgated. Slowly, between 1221 and 1232, the simple military system of Yoritomo was transformed by the Hōjō family into a capable private government.

Relationship with the court and the aristocracy

Essentially, this meant maintaining a cordial but careful relationship with the court and its complex system of reigning, retired, and cloistered emperors and with the great aristocracy of Kyōto, who wished an end to the *bakufu* system. A Hōjō commander and garrison were stationed in Kyōto; but the property, revenues, and ceremonials of the Imperial family and nobility were protected. The powerful Buddhist clergy were kept in hand by strict auditing of their accounts. The vassals of Hōjō were kept solvent, peaceful, and apart from the court. The peasant was protected in his freedom and tenure. The regency drew its income from the Hōjō estates, which comprised nearly the whole of the Kantō. The family adhered firmly to Yoritomo's dictum that the simple warrior life would best preserve this class from the pervasive decadence of the Kyōto aristocracy. Yasutoki died in 1242 and was succeeded by the Hōjō regents Tsunetoki (1224–46) in 1242, Tokiyori (1227–63) in 1246, and Tokimune (1215–84) in 1256. Tokimune's regency was the last stable and powerful epoch of the Hōjō. Tokimune refused the Mongol Kublai Khan's demand (1271) that Japan pay tribute to him. The result was an unsuccessful Mongol–Chinese–Korean assault on Hakata Harbour in Kyushu. In 1281 a massive second joint assault on Kyushu was again beaten back; but the cost of preparing the defense, of the two months' battle around Hakata, and of maintaining a war footing until Kublai died in 1294 was deadly. For 20 years Hōjō resources had been under great strain in the defense of Japan; the resources of their vassals had been consumed in the war.

Decline of Hōjō power

When Sadatoki (1270–1311) became regent in 1284, he found himself so embroiled in a succession dispute between two powerful factions of the Imperial family—a struggle beginning to split all Japan—that he secluded himself in a temple, from where he continued to administer Japan during the last ten years of his life. His successor, the ninth and last Hōjō regent, Takatori (1303–33), passed his minority dissolutely and extravagantly. On attaining his majority (1316) he left the affairs of the regency in the hands of inept men at a time when only a severe and powerful man could have managed the difficult economic and political situation. In 1331, because of the continuing quarrel over the Imperial succession, Takatori exiled the emperor Daigo II. Escaping from exile, the Emperor found it easy to raise war against the Hōjō. Takatori was betrayed by his own general, Ashikaga Takauji, who seized Kyōto from its Hōjō garrison. The *bakufu*'s own domain of the Kantō rose in revolt under Nitta Yoshisada (the opposition to the Hōjō was, in part, a revolt of the family's own constables and stewards, who had become locally powerful). Nitta sacked Kamakura, and on May 22, 1333, the last Hōjō regent committed suicide. But the foundation the Hōjō had laid was enduring. Daigo's attempt to restore a civil Imperial government lasted only three years. Ashikaga Takauji declared himself shogun in 1336, and from then until 1868 a form of *bakufu*—as created by Yoritomo and refined by the Hōjō—ruled Japan.

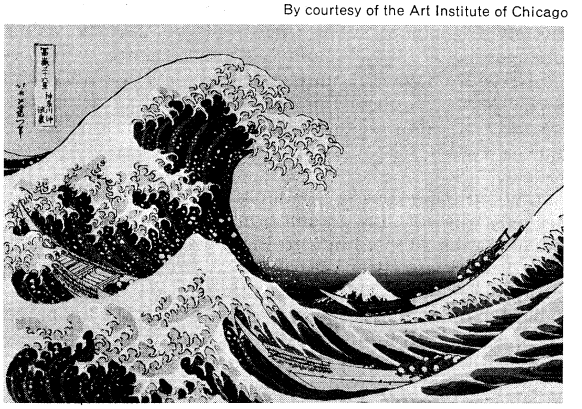
BIBLIOGRAPHY. GEORGE SANSOM, *A History of Japan to 1334*, vol. 1 (1958), an excellent general history of the period; SHINODA MINORU, *The Founding of the Kamakura Shogunate, 1180–1185* (1960), a prime source for the feudal period, based on the *Azuma Kagami*—explains the origin of military government; FREDERIC JOUON DES LONGRAIS, *Age de Kamakura: Sources (1150–1333)*, vol. 3 (1950), a detailed guide to feudal law, government, and documents; JOHN HALL, "Japanese Feudal Laws: The Hojo Code of Judicature," *Tasj*, 1st Series, vol. 34 (1906).

(J.A.Ha.)

Hokusai

A famed Japanese artist, Hokusai embodied in his long lifetime the essence of the *Ukiyo-e* (Pictures of the Floating World) school of art during its final century of development. His stubborn genius also represents, in its 70 years of continuous artistic creation, the prototype of the single-minded artist, striving only to complete a given task. Moreover, Hokusai constitutes a figure who has, since the later 19th century, impressed Western artists, critics, and art lovers alike, more, possibly, than any other single Asian artist.

By courtesy of the Art Institute of Chicago



"The Breaking Wave off Kanagawa," woodblock print by Hokusai, from the series "Thirty-six Views of Mt. Fuji," 1826–33. 36.4 cm × 25.4 cm.

Birth and youth

Born in 1760, in the Honjo quarter just east of Edo (Tokyo), Hokusai became interested in drawing at the age of five. He was adopted in childhood by a prestigious artisan family named Nakajima but was never accepted as an heir—possibly supporting the theory that, though the true son of Nakajima, he had been born of a concubine.

Hokusai is said to have served as a youth as clerk in a lending bookshop; and from 15 to 18 was apprenticed to a wood-block engraver. This early training in the book and printing trades obviously contributed to Hokusai's artistic development as a printmaker.

The earliest contemporary record of Hokusai dates from the year 1778, when, at the age of 18, he became a pupil of the leading *Ukiyo-e* master, Katsukawa Shunshō. The young Hokusai's first published works appeared the following year—actor prints of the Kabuki theatre, the genre that Shunshō and the Katsukawa school practically dominated.

To judge from the ages of his several children, Hokusai must have married in his mid-20s. Possibly under the influence of family life, from this period his designs tended to turn from prints of actors and women to historical and landscape subjects, especially *uki-e* (semi-historical landscapes using Western-influenced perspective techniques), as well as prints of children. The artist's book illustrations and texts turned as well from the earlier themes to historical and didactic subjects. At the same time, Hokusai's work in the *surimono* genre during the subsequent decade marks one of the early peaks in his career. *Surimono* were prints issued privately for special

occasions—New Year's and other greetings, musical programs and announcements, private verse selections—in limited editions and featuring immaculate printing of the highest quality.

Hokusai's early 30s were to prove years of personal change. His master Shunshō died early in 1793, and somewhat later his young wife passed away, leaving a son and two daughters. In the year 1797 he remarried and adopted the name Hokusai. This change of name marks the beginning of the golden age of his work, which was to continue for a half century.

In format, Hokusai's *oeuvre* from this period covers the gamut of *Ukiyo-e* art: single-sheet prints, *surimono*, picture books and picture novelettes, illustrations to verse anthologies and historical novels, erotic books and album prints, and hand paintings and sketches. In his subject matter, Hokusai only occasionally (in a few notable prints, in paintings, and erotica) chose to compete with Utamaro, the acknowledged master of voluptuous figure prints. Aside from this limitation, however, Hokusai's work encompassed a wide range, with particular emphasis on landscape views and historical scenes in which figures were often of secondary interest. Around the turn of the century he experimented for a time with Western-style perspective and colouring.

From the early 19th century Hokusai commenced illustrating *yomihon* (the extended historical novels that were just coming into fashion). Under their influence, his style began to suffer important and clearly visible changes between 1806 and 1807. His figure work becomes more powerful but increasingly less delicate; there is greater attention to classical or traditional themes (especially of samurai, or warriors, and Chinese subjects) and a turning away from the contemporary *Ukiyo-e* world.

In about the year 1812, Hokusai's eldest son died. This tragedy was not only an emotional but also an economic event, for as adopted heir to the affluent Nakajima family, the son had been instrumental in obtaining a generous stipend for Hokusai, so that he did not need to worry about the uncertainties of income from his paintings, designs, and illustrations, which at this period were paid for more with "gifts" than with set fees.

The *Hokusai manga*

Whether for economic reasons or not, from this time on Hokusai's attention turned gradually from novel illustration to the picture book and, particularly, to the type of wood-block-printed copybook designed for amateur artists (including the famous *Hokusai manga*). Very likely his intention was to find new pupils and hence new patronage, and in this he succeeded to some degree.

Though famed for his detailed prints and illustrations, Hokusai was also fond of displaying his artistic prowess in public—making, for example, huge paintings (some fully 200 square metres [about 2,000 square feet] in area) of mythological figures before festival crowds, in both Edo and Nagoya. He was once even summoned to show his artistic skills before the shogun (the military leader who, although theoretically subordinate to the emperor, was in fact the ruler of Japan).

In the summer of 1828, Hokusai's second wife died. The master was then 68, afflicted intermittently with paralysis and left alone, evidently with only a profligate grandson, who had proved to be an incorrigible delinquent. It is probably no coincidence, therefore, that before long Hokusai's favourite daughter (and pupil), O-ei, broke her unhappy marriage with a minor artist named Tōmei and returned to her father's side, where she was to stay for his remaining years.

An energetic artist, Hokusai rose early and continued painting until well after dark. This was the customary regimen of his long, productive life. Of Hokusai's thousands of books and prints, his "Thirty-six Views of Mt. Fuji" is particularly notable. Published from about 1826 to 1833, this famous series (including supplements, a total of 46 colour prints) marked a summit in the history of the Japanese landscape print; in grandeur of concept and skill of execution there was little approaching it before and nothing to surpass it later—even in the work of Hokusai's famed late contemporary Hiroshige (*q.v.*).

Hokusai's frequent changes in domicile (more than 90 dwellings) and of his own name are indicative of the artist's restless nature. Besides his principal *noms d'artiste* (roughly one per decade)—Shunrō, Sōri, Kakō, Hokusai (often with the prefix Katsushika), Taitō, Gakyō, Iitsu, Manji—the artist had also some two dozen other occasional pseudonyms, though these were normally used as adjuncts to his principal name of a given period.

Despite his appeals to heaven for "yet another decade—nay, even another five years," on the 18th day of the fourth month (*i.e.*, mid-May in the Western calendar) in the year 1849 "the old man mad with painting," as he called himself, breathed his last. He was 89 but still insatiably seeking for an ultimate truth in art—as he had written some 15 years earlier:

> From the age of five I have had a mania for sketching the forms of things. From about the age of 50 I produced a number of designs, yet of all I drew prior to the age of 70 there is truly nothing of any great note. At the age of 73 I finally apprehended something of the true quality of birds, animals, insects, fishes, and of the vital nature of grasses and trees. Therefore, at 80 I shall have made some progress, at 90 I shall have penetrated even further the deeper meaning of things, at 100 I shall have become truly marvelous, and at 110, each dot, each line shall surely possess a life of its own. I only beg that gentlemen of sufficiently long life take care to note the truth of my words.

MAJOR WORKS

PRINTS AND ILLUSTRATIONS: "Festivals of the Green Houses" (*c.* 1790); "Festivals for the Twelve Months" (*c.* 1790); "Foreigners Observing Japanese Customs" (1796); *Chūshingura* series (I) (*c.* 1800); "Brocade Prints of the Thirty-six Poetesses" (1801); "Fifty Fanciful Poets, Each with One Poem" (1802); "Fuji in Spring (1803); "A Picture Book of Kyōka" (1803–04); "Fifty-three Stations on the Tōkaidō" (1804); *Chūshingura* series (II) (1806); *Suikoden* (1807); "Portraits of Six Poets" (*c.* 1810); "Quick Lessons in Simplified Drawing" (1812); *Hokusai manga*, vol. 1 (1814); *Hokusai gashiki* (1819); *Hokusai sōga* (1820); "Paintings with One Stroke of the Brush" (1823); "Thirty-six Views of Mt. Fuji" (*c.* 1826–33); "Views of Famous Bridges" (*c.* 1827–30); "Snow, Moon, and Flowers" (*c.* 1827–30); "Flowers and Birds" (*c.* 1827–30); "The Poems of China and Japan Mirrored to Life" (*c.* 1828–33); "Tōshi-sen" (1833–36); "Hundred Views of Mt. Fuji" (1834–35); "Hundred Poems Explained by the Nurse" (*c.* 1845).

BIBLIOGRAPHY. J.R. HILLIER, *Hokusai: Paintings, Drawings, and Woodcuts* (1955), the best general appreciation of Hokusai in English, though the biographical material is based on outdated sources—includes a detailed listing of his illustrated books; JAMES A. MICHENER (ed.), *The Hokusai Sketchbooks: Selections from the Manga*, with translations by RICHARD LANE (1958), a comprehensive sampling of the *Hokusai manga*, with commentary, and translation of all prefaces; THEODORE R. BOWIE, *The Drawings of Hokusai* (1964), a pioneer study, but flawed by being based largely on forgeries and school copies; RICHARD LANE, *Masters of the Japanese Print* (1962), includes a critical survey of Hokusai's work and times, based on original sources.

(Ri.L.)

Holbein, Hans, the Younger

Hans Holbein, the Younger, 16th-century German painter, was one of the greatest portraitists and most exquisite draftsmen of all time. He was both prolific and versatile, painting religious panels, portraits, and miniatures, designing woodcuts and stained-glass windows, and executing large interior and exterior murals. It is the artist's record of the court of King Henry VIII of England, as well as the taste that he virtually imposed upon that court, that was his most remarkable achievement.

Holbein was born in Augsburg. The exact date is not known, but because he was 46 when he died, it is commonly assumed that he was born late in 1497 or early in 1498. He was a member of a family of important artists. His father, Hans Holbein the Elder, and his uncle Sigmund were renowned for their somewhat conservative examples of late Gothic painting in Germany. One of Holbein's brothers, Ambrosius, became a painter as well, but he apparently died about 1519 before reaching maturity as an artist. The Holbein brothers no doubt first studied with their father in Augsburg; they both also be-



Hans Holbein, the Younger, self-portrait miniature on a playing card, 1543. In the Wallace Collection, London. Diameter 3.7 cm.
By courtesy of the trustees of the Wallace Collection, London

gan independent work in about 1515 in Basel, Switzerland. It should be noted that this chronology places Holbein firmly in the second generation of 16th-century German artists. Albrecht Dürer, Matthias Grünewald, and Lucas Cranach all were born around 1470 and were producing their mature masterpieces by the time Holbein was just beginning his career. Holbein is, in fact, the only truly outstanding German artist of his generation.

Early years in Basel

Holbein's work in Basel during the decade of 1515–25 was extremely varied, if also sometimes derivative. Trips to northern Italy (*c.* 1517) and France (1524) certainly affected the development of his religious subjects and portraiture, respectively.

Holbein was associated early with the Basel publishers and their Humanist circle of acquaintances. There he found not only portrait commissions (*e.g.*, "Portrait of Bonifacius Amerbach") but was especially active in designing woodcuts for title pages and book illustrations. The artist's most famous work in this area, a series of 41 scenes illustrating the medieval allegorical concept of the "Dance of Death," was designed by him and cut by another artist as early as *c.* 1523–26 but was not published until 1538. Holbein entered the painters' corporation in 1519, married a tanner's widow, and became a burgher of Basel in 1520; by 1521 he was executing important mural decorations in the Great Council Chamber of the town hall.

Protestantism, introduced into Basel as early as 1522, grew considerably in strength and importance there during the ensuing four years. By 1526 severe iconoclastic riots and strict censorship of the press swept over the city. In the face of what, for the moment at least, amounted to a freezing of the arts, Holbein left Basel late in 1526, with a letter of introduction from the Dutch Humanist scholar Erasmus, to travel by way of the Netherlands to England. Still only about 28 years old, he achieved remarkable success in England. His most impressive work of this time was executed for the statesman and author Sir Thomas More and included a magnificent single portrait of the Humanist as well as a group portrait of the scholar's family (now lost; preserved in copies and preparatory drawing, Basel, Kunstmuseum). The latter was the first example in northern Europe of a large group portrait in which the figures are not shown kneeling—the effect of which is to suggest the individuality of the sitters rather than impiety.

First trip to England

Before Holbein journeyed to England in 1526, he had apparently designed works both pro- and anti-Lutheran in character. On returning to Basel in 1528, he was admitted, after some hesitation, to the new—now official—faith. It would be difficult to interpret this as a very decisive change, for Holbein's most impressive religious works, like his portraits, are brilliant observations of physical reality but seemingly were never inspired by Christian mysticism or spirituality. In any case, Holbein apparently quite voluntarily gave up almost all religious painting after *c.* 1530.

In Basel, from 1528 to 1532, Holbein continued his important work for the Town Council. He painted as well what is perhaps his only psychologically penetrating portrait, that of his wife and two sons (Kunstmuseum, Basel). This picture no doubt conveys some of the unhappiness of the abandoned family. In spite of generous offers from the city of Basel, Holbein left his family there for a second time, in 1532, to spend the last 11 years of his life primarily in England.

By 1533 Holbein was already painting court personalities, and four years later he officially entered the service of Henry VIII. Commissions from the King were varied and demanding but must have been above all enormously satisfying to the artist's professional and personal ambitions. It is estimated that during the last ten years of his life Holbein executed about 150 portraits, life-size and miniature, of royalty and nobility alike. Holbein acted not only as a portraitist but also as a fashion designer for the court. The artist made designs for all the state robes of the King; he left, in addition, more than 250 delicate drawings for everything from buttons and buckles to pageant weapons, horse outfittings, and bookbindings for the royal household. This choice of work indicates Holbein's Mannerist concentration on surface texture and details of design, a concern that probably precluded the incorporation of serious psychological depth. Holbein died in a London plague epidemic in 1543.

The fact that Holbein's portraits do not reveal the character or spiritual inclinations of his sitters is perfectly paralleled by knowledge of the artist's life. His biography is basically a recounting of disparate facts; about his personality practically nothing is known. Not one note or letter from his own hand survives. Other men's opinions of him are often equally inscrutable. Erasmus, one of Holbein's most renowned sitters, praised and recommended him on one occasion but scorned the artist as opportunistic at another time. Indeed, Henry VIII, who sent him to the Continent to help select a bride by providing a dependable portrait for his scrutiny, was perhaps the only person who had absolute confidence in Holbein.

The biographical problem

The artist's detachment and his refusal to submit to an authority that might inhibit his own creative (but very worldly) powers enabled him to produce works whose beauty and brilliance have never been questioned. Had he been a more devout Christian or more subject to the turmoil of his times, his artistic achievement might have been quite different. In recent times, the lack of spiritual involvement in his work has been consistently noted, especially inasmuch as the 16th century was a time when few artists managed to remain above the religious conflict sweeping Europe. Thus, the effect of Holbein's life and work has often been felt to be more artistic and external than expressionistic or emotional. Only in that sense, however, is his achievement finally limited.

MAJOR WORKS

PAINTINGS: "Portrait of Bonifacius Amerbach" (1519; Kunstmuseum, Basel, Switzerland); "The Last Supper" (1519–20; Kunstmuseum); "The Flagellation" (1519–20; Kunstmuseum); "Dead Christ" (1521; Kunstmuseum); "Portrait of Erasmus" (1523–24; Louvre, Paris); "Virgin with the Family of the Burgomaster Jakob Meyer" ("Darmstadt Madonna"; 1526; Schlossmuseum, Darmstadt); "Magdalena Offenburg as Venus" (1526; Kunstmuseum); "Sir Thomas More" (1526; Frick Collection, New York); "Portrait of William Warham, Archbishop of Canterbury" (1527; Louvre, Paris); "Sir Henry Guildford" (1527; Windsor Castle, Berkshire); "Portrait of Lady Guildford" (1527; St. Louis Art Museum, St. Louis); "Portrait of Nicolas Kratzer" (1528; Louvre); "Double Portrait of Sir Thomas Godsalve and His Son Sir John" (1528; Gemäldegalerie, Dresden); "Portrait of the Artist's Wife with Katharina and Philipp" (1528; Kunstmuseum); "Erasmus in the Roundel" (1532; Kunstmuseum); "Portrait of a Member of the Wedigh Family" (1532; Metropolitan Museum of Art, New York); "Hans of Antwerp" (1532; Windsor Castle); "Portrait of the Merchant Georg Gisze" (1532; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Derick Born" (1533; Windsor Castle); "Dirk Tybis of Duisburg" (1533; Kunsthistorisches Museum, Vienna); "Jean de Dinteville and Georges de Selve" ("The Ambassadors"; 1533; National Gallery, London); "Portrait of Charles de Solier, Sieur de Morette" (1535; Gemäldegalerie, Dresden); "Sir Richard Southwell" (1536; Uffizi, Florence); "Henry VIII" (1536; Thyssen-Bornemisza Collection, Castagnola, Switzerland); "Jane Seymour" (1536; Kunsthistorisches Museum, Vienna);

"Christina of Denmark, Duchess of Milan" (1538; National Gallery, London).

WOODCUTS: *Dance of Death* (1538; designed *c.* 1523–26); *Icones historiarum Veteris Testamenti* (94) (1538; designed before 1531); "Erasmus im Gehüs" (1535); title page, Coverdale's Bible (1535; British Museum).

MINIATURES: "Portrait of Mrs. Pemberton" (*c.* 1540; Victoria and Albert Museum, London); "Anne of Cleves" (*c.* 1540; Victoria and Albert Museum); "Charles Brandon" (*c.* 1540; Windsor Castle).

BIBLIOGRAPHY. A.F.G. WOLTMANN, *Holbein und seine Zeit*, 2 vol. (1866–68; Eng. trans., *Holbein and His Time*, 1872), one of the first, most exhaustive, and still basic biographies of the artist; PAUL GANZ (ed.), *Handzeichnungen von Hans Holbein dem Jüngeren* (n.d.), the most elaborate and only facsimile edition of all of Holbein's drawings; A.B. CHAMBERLAIN, *Hans Holbein the Younger*, 2 vol. (1913), an exhaustive, still valuable biography of the artist in English; K.T. PARKER, *The Drawings of Hans Holbein in the Collection of His Majesty the King at Windsor Castle* (1945), the most recent scholarly account of what is probably the single most important collection of Holbein's mature drawings; H.A. SCHMID, *Hans Holbein der Jüngere; sein Aufstieg zur Meisterschaft und sein englischer Stil*, 2 vol. (1945–48), a lengthy and valuable monograph on the artist in German; PAUL GANZ (ed.), *Hans Holbein: Die Gemalde* (1949; Eng. trans., *The Paintings of Hans Holbein*, 1950), the most detailed monograph on Holbein's painted work in English; BASEL, UNIVERSITAT, KUNSTSAMMLUNG, *Die Malerfamilie Holbein in Basel* (1960), a comprehensive catalog of the most important exhibition ever organized concerning Holbein and his family; R.C. STRONG, *Holbein and Henry VIII* (1967), the most recent treatment in English of this important aspect of Holbein's career.

(C.S.Ha.)

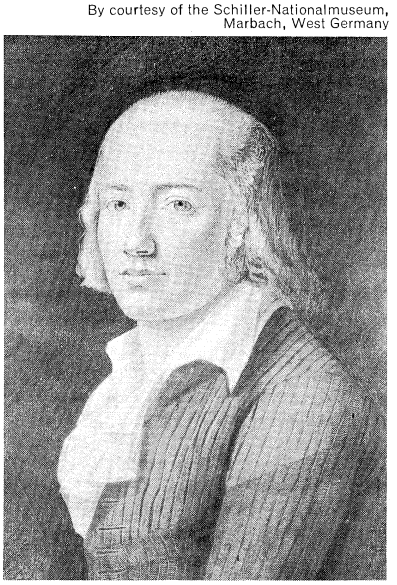
Hölderlin, Friedrich

One of the outstanding lyric poets in the German language, Johann Christian Friedrich Hölderlin gained little recognition during his lifetime and was almost totally forgotten for nearly 100 years. It was not until the early years of the 20th century that he was rediscovered in Germany and that his reputation was established in Europe. Today he is ranked among the greatest of German poets, especially admired for his uniquely expressive style: like no one before or since, he succeeded in naturalizing the forms of classical Greek verse in the German language. With passionate intensity he strove to reconcile the Christian faith with the religious spirit and beliefs of ancient Greece; he was a prophet of spiritual renewal, of "the return of the gods"—utterly dedicated to his art, hypersensitive, and therefore exceptionally vulnerable. In the end his mind gave way under the strains and frustrations of his existence.

Hölderlin was born in the little Swabian town of Lauffen, on the River Neckar, on March 20, 1770. His father died in 1772, and two years afterward his mother married the burgomaster of the town of Nürtingen, where Friedrich attended school. But his mother was again widowed, in 1779, and left alone to bring up her family —which included Friedrich, his sister Heinrike, and his half-brother Karl. His mother, a parson's daughter, and a woman of simple and rather narrow piety, wanted Friedrich to enter the service of the church. Candidates for the ministry received free education—a chance not to be missed by a gifted but impecunious boy—and accordingly he was sent first to the "monastery schools" (so called since pre-Reformation times) at Denkendorf and Maulbronn and subsequently (1788–93) to the theological seminary in the University of Tübingen, where he obtained his master's degree and qualified for ordination.

Training for the ministry

By courtesy of the Schiller-Nationalmuseum, Marbach, West Germany



Hölderlin, pastel by Franz Karl Hiemer, 1792. In the Schiller-Nationalmuseum, Marbach, West Germany.

Hölderlin could not, however, bring himself to enter the ministry for which he had studied. Contemporary Protestant theology, an uneasy compromise between faith and reason, offered him no safe spiritual anchorage, while acceptance of Christian dogma was not wholly compatible with his devotion to Greek mythology, which made him see the gods of Greece as real living forces whose presence manifests itself to men in sun and earth, sea and sky. The strain of divided allegiance remained a permanent condition of his existence. Although he did not feel called to be a Lutheran pastor, Hölderlin did have a strong sense of religious vocation; for him, being a poet meant exercising the priestly function of mediator between gods and men.

Meeting with Schiller

In 1793 Hölderlin was introduced to Friedrich Schiller, and it was through Schiller's recommendation that he obtained the first of several posts as a tutor (in most of which he failed to give satisfaction). Schiller befriended the younger man in other ways too; in his periodical *Neue Thalia*, he published some of the poetry that Hölderlin had begun to write, as well as a fragment of his novel *Hyperion*. This elegiac story of a disillusioned fighter for the liberation of Greece remained unfinished. Hölderlin held Schiller in great reverence; he saw him again when in 1794 he left his tutor's post in order to move to Jena. His early poems clearly reveal Schiller's influence, and several of them acclaim the new world the French Revolution had seemed to promise in its early stages: they include hymns to freedom, to humanity, to harmony, to friendship, and to nature.

In December 1795 poverty forced Hölderlin to take a post in the house of J.F. Gontard, a wealthy Frankfurt banker. Before long, the susceptible young tutor was deeply in love with his employer's wife, Susette, a woman of great beauty and sensibility, and his affection was returned. In a letter to his friend C.L. Neuffer (February 1797), he described their relationship as "an everlasting happy sacred friendship with a being who has really strayed into this miserable century." Susette appears in his poems and in his novel *Hyperion*, the second volume of which appeared in 1799, under the Greek name of "Diotima"—a reincarnation of the spirit of ancient Greece. Their happiness was short-lived; after a painful scene with Susette's husband, Hölderlin had to leave Frankfurt (September 1798).

Onset of mental illness

Though physically and mentally shaken, he finished the second volume of *Hyperion* and began a tragedy, *Der Tod des Empedokles* (*The Death of Empedocles*), the first version of which he nearly completed; fragments of a second and a third version have also survived. Symptoms of great nervous irritability alarmed his family and friends. Nevertheless the years 1798–1801 were a period of intense creativity; in addition to a number of noble odes, they produced the great elegies "Menons Klagen um Diotima" ("Menon's Lament for Diotima") and "Brot und Wein" ("Bread and Wine"). In January 1801, he went to Switzerland as tutor to a family in Hauptwyl. His employer, however, found himself obliged to make other arrangements and Hölderlin returned home in April of the same year.

After vainly attempting to obtain a lectureship in Greek literature at Jena through the influence of Schiller, he once more accepted a post as tutor, this time at Bordeaux,

in France. Susette Gontard died in June 1802; in the same summer, Hölderlin suddenly left Bordeaux and travelled homeward on foot through France, arriving at Nürtingen completely destitute and mentally deranged, in an advanced stage of schizophrenia. He seemed to recover somewhat as a result of the kind and gentle treatment he received at home. The poems of the period 1802–06, including "Friedensfeier" ("Celebration of Peace"), "Der Einzige" ("The Only One"), and "Patmos," products of a mind on the verge of madness, are apocalyptic visions of unique grandeur. He also completed verse translations of Sophocles' *Antigone* and *Oedipus Tyrannus*, published in 1804. In this year a devoted friend, Isaak von Sinclair, obtained for him the sinecure post of librarian to the landgrave Frederick V of Hesse-Homburg. Sinclair himself provided a modest salary, and Hölderlin improved noticeably under his care and companionship. In 1805, Sinclair (who refused to believe that Hölderlin was insane) was falsely accused of subversive activities and held in custody for five months. By the time he was released, Hölderlin had succumbed irretrievably and, after a spell in a clinic in Tübingen, was moved to a carpenter's house in that town, where he died on June 7, 1843, having passed the last thirty-six years of his life under the shadow of insanity.

Two years before he became mentally unbalanced, Hölderlin had summed up his destiny in the concluding lines of his ode "Die Heimat" ("Home"):

> For they who lend us the heavenly fire, the Gods, give us sacred sorrow too. Let it be so. A son of earth I seem; born to love and to suffer.

MAJOR WORKS

Hyperion, oder der Eremit in Griechenland, vol. 1 (1797), vol. 2 (1799), epistolary novel; verse translations of Sophocles' *Antigone* and *Oedipus Tyrannus* (1804). *Hyperion* and the translations were the only works published in book form before the onset of Hölderlin's madness. Several editions of selected poems were published in the 19th century, the first in 1826 (edited by L. Uhland and G. Schwab), but it was not until the 1920s that a reliable and comprehensive edition was available. Among his best known poems are: (written before 1802) "Der Mensch," "Hyperions Schicksaalslied," "An die Parzen," "An die Hoffnung," "Brod und Wein"; (written between 1802 and 1806) "Hälfte des Lebens," "Der blinde Sänger," "Germanien," "Der Rhein," "Der Einzige," and "Patmos."

BIBLIOGRAPHY. For detailed bibliographical information, see F. SEEBASS, *Hölderlin-Bibliographie* (1922); A. KELLETAT and M. KOHLER, *Hölderlin-Bibliographie 1938–1950* (1953); and the periodical surveys in the *Hölderlin-Jahrbuch*.

The major modern edition of Hölderlin's works is *Sämtliche Werke*, ed. by FRIEDRICH BEISSNER (vol. 1–5, 1946–52) and ADOLF BECK (vol. 6–7: the Letters, 1954–). Beissner has also edited a plain text of the complete works in one volume (1964). There is a one-volume collection of Hölderlin's letters, edited by ERNST BERTRAM (1935). The poems of Hölderlin's maturity, together with the second and third fragmentary versions of his drama *Der Tod des Empedokles*, have been translated by MICHAEL HAMBURGER: *Friedrich Hölderlin: Poems and Fragments*, bilingual edition (1966). A translation of *Selected Poems* was published in 1944 (2nd ed., 1954) by J.B. LEISHMAN. *Hyperion* has been translated by W.R. TRASK (1965).

There are several biographical and critical studies in English: RONALD PEACOCK, *Hölderlin* (1938); AGNES STANFIELD, *Hölderlin* (1944); and L.S. SALZBERGER, *Hölderlin* (1952). Important aspects of Hölderlin's art are discussed by E.L. STAHL in *Hölderlin's Symbolism* (1945); and in "Hölderlin's Idea of Poetry," in *The Era of Goethe* (1959). WHILHELM DILTHEY'S essay on Hölderlin in *Das Erlebnis und die Dichtung* (1906) remains a classic. A number of illuminating recent studies by various hands are gathered together in *Über Hölderlin*, ed. by J. SCHMIDT (1970). ALESSANDRO PELLEGRINI has written a detailed history of Hölderlin criticism: *Friedrich Hölderlin: Sein Bild in der Forschung* (1965).

(W.Wi.)

Holiness Churches

At least 20 religious groups in the United States with a combined membership in excess of 1,000,000 comprise the Holiness movement. Most of these churches place great stress on a doctrine of sanctification (from the Latin *sanctus*, "holy"), a postconversion experience that enables a believer to live a sinless, perfect, holy, Christian life. Some of these perfectionist bodies have incorporated the term holiness in their official church names: Holiness Methodist Church (which was absorbed by the Evangelical Church of North America in 1969), Church of Christ Holiness U.S.A., Fire Baptized Holiness Church; others, like the Wesleyan Church, Free Methodist Church of North America, Primitive Methodist Church, Church of God (Anderson, Indiana), Church of the Nazarene, Christian and Missionary Alliance, and the Salvation Army, do not.

NATURE AND SIGNIFICANCE

The beliefs that are characteristic of most Holiness churches include (1) the basic tenets of Fundamentalism —*i.e.*, biblical inerrancy, the virgin birth of Christ, his substitutionary atonement (*i.e.*, the reconciliation of man with God through Christ), physical resurrection, and imminent return of Christ to earth; (2) the principles of Arminian theology, namely the doctrines of free will and free grace; (3) the truths proclaimed by John Wesley, founder of the Methodist Church, and popularized during the Methodist revivals of the 18th and 19th centuries (*i.e.*, that by "praying through" at a mourners' bench every sinner can experience a positive inner assurance of personal salvation followed by a "second blessing"—sanctification); and (4) a doctrine of separatism that advocates the avoidance of "worldly" practices such as attending the movies, dancing, using tobacco or alcohol, or belonging to "secret societies."

Methodist and Pentecostal influences

In many ways, a number of Holiness churches are quasi-Methodist sects. They are Wesleyan in faith and practice, their founders frequently were Methodist ministers, and their original members were former Methodists. One of the largest Holiness bodies, the Church of the Nazarene, clearly illustrates this: its polity is Methodistic; its *Manual* is admittedly an adaptation of the Methodist *Discipline*; and five of its first seven general superintendents had once been Methodist preachers. On the other hand, the doctrinal stance of still other Holiness groups—their Fundamentalism, millenarianism (doctrines concerning the earthly aspects of the consummation of history), revivalism (activities encouraging religious conversions), and asceticism (denial of bodily pleasures or needs)—has allowed them to gravitate toward theologically conservative bodies, such as the National Association of Evangelicals. Currently, the Free Methodist Church and the Wesleyan Church are actively affiliated with that organization. Finally, the emphasis that many Holiness churches place on a second-blessing experience, the result of the direct operation of the Holy Spirit in the heart of a believer, closely resembles the primacy accorded the work of the Spirit by Pentecostal groups, which stress the marvels of the Spirit's first coming to the church (Acts 2). Although it is a fact, however, that the taproots of a number of the older Pentecostal denominations reach back into the Holiness revivals of the 19th century and despite the fact that both groups teach that there is a blessing to be sought and to be experienced subsequent to and distinct from conversion, most Holiness churches are reluctant to be identified with the Pentecostal movement. As a matter of fact, early in its development (the General Assembly in 1919) the Church of the Nazarene deleted the word Pentecostal from its official church name. Many of the larger Holiness denominations have presently abandoned an earlier view that unbridled expressions of emotion—boisterous praying, vigorous bodily movement, or vociferous shouting—are normative outward evidences of the inward sanctifying work of the Holy Spirit. Most Pentecostals, on the other hand, still believe that their second-blessing experience is certified by an outward sign, the ecstatic manifestation of speaking in tongues.

HISTORY OF THE HOLINESS MOVEMENT

The quest for perfection, or personal holiness, is to those who seek it as old as Christianity itself. The pursuers of perfection find sufficient warrant for it in the Gospels

as well as in the apostolic writings, in passages such as "You, therefore must be perfect, as your heavenly Father is perfect" (Matt. 5:48), or "For God has not called us for uncleanness, but in holiness" (I Thess. 4:7). Yet opponents of perfectionism hasten to point out that throughout the history of Christianity those who have espoused such views have, for the most part, been pneumatics (spiritualists), heretics, or fanatics, forming groups such as the Gnostics, Pelagians, Montanists, Fraticelli, Anabaptists, or Shakers. Moreover, they argue that the principal orthodox theologians from St. Augustine to John Calvin have staunchly declared that one can never be entirely free from sin and its temptations and, therefore, can never achieve perfection in this life.

Opponents of perfectionism

Methodist roots. On the whole, this view prevailed until the 18th century when John Wesley, the founder of Methodism, issued his call to Christian perfection. For Wesley, the doctrine of perfection was a principle based upon a scriptural foundation. Perfection was to be the goal of all those who desired to be *altogether* Christian; it implied that the God who is good enough to forgive sin (justify) is obviously great enough to transform the sinner into a saint (sanctify), thus enabling him to be free from outward sin as well as from "evil thoughts and tempers," in short, to attain to a measure of holiness. And indeed, concepts such as Christian perfection, entire sanctification, and personal holiness were central both in Wesley's preaching as well as in the evangelical revival that swept across 18th-century England and eventually reached the American colonies.

From the outset, the motto of American Methodism was "to spread Christian holiness over these lands." The 1787 edition of the *Discipline* admonished those who planned to be itinerant preachers "to save as many souls as you can; to bring as many sinners as you possibly can to Repentance, and with all your power to build them up in that holiness without which they cannot see the Lord." All this notwithstanding, the doctrine of holiness was largely ignored by American Methodists during the early decades of the 19th century. The slackening of interest in perfectionism has been attributed to the rising national prosperity, which did little to encourage deepening piety, and to excesses that were associated with it in the popular mind because of the conduct of certain contemporaneous antinomian (antilegalistic) perfectionist societies, such as the Oneida Community. The neglect of the doctrine of holiness was only temporary, however. In the 1830s perfectionism received new impetus with the publication of a monthly periodical entitled *Guide to Christian Perfection* and the inauguration of the "Tuesday Meeting for the Promotion of Holiness" in the Manhattan home of Phoebe Palmer, a laywoman active in the Allen Street Methodist Church. Moreover, many prominent leaders in Methodist circles, including Bishops Edmund S. Janes and Leonidas L. Hamline, heartily endorsed the revival of interest in holiness and "the higher Christian life."

Growing polarization

By the time of the Civil War, tensions were developing. In some quarters of Methodism, at least, the principles of Christian perfection and holiness had become merely part of official doctrinal statements. In practice these ideals were largely ignored, and some church papers reflected a growing hostility to the subject of entire sanctification and an increasing skepticism in regard to the role of the camp meetings; *i.e.*, revivalist retreats, often attended for several days, centring around services at a large tent. On the other hand, however, there were many Methodists who were convinced that personal sanctification, or holiness, was, at the very least, one of the cardinal ideas of their founder, and they were determined to remain loyal to their Wesleyan heritage. They looked askance at the proposed abandonment of camp meeting revivalism; they bemoaned the expanding autocracy of the bishops and the increasing indulgence of worldly amusements, "broidered hair," gold, pearls, and costly array. When differences between these factions became irreconcilable, withdrawals and schisms resulted.

Schisms and associations. In 1843 about two dozen ministers and 6,000 members withdrew from the Methodist Episcopal Church to found the Wesleyan Methodist Church of America. Then, in 1860, after a decade of controversy over the issues of perfectionism and the power of the episcopacy, another body of Methodists left the Genessee (N.Y.) Conference to establish the Free Methodist Church of North America. Despite these defections, however, it is a fact that most Protestants who favoured holiness or perfectionist views were reluctant to abandon their churches. They preferred to retain their memberships therein and then to seek fellowship with other like-minded Christians in independent organizations such as the National Campmeeting Association for the Promotion of Holiness, which was established in 1867. As a matter of fact, in the years that followed, this organization became an effective means of disseminating Holiness ideas. At diverse encampments, on both coasts—hundreds of conferees testified to receiving the second blessing, the consciousness of sanctifying grace.

This arrangement to retain one's denominational affiliation while simultaneously finding an outlet for one's perfectionistic proclivities in some local, regional, or national society for the promotion of holiness worked temporarily. As time passed, however, it became increasingly difficult for people who were thus inclined to keep up the practice. For one thing, sizable numbers of Protestants from the rural areas of the Midwest and South were joining the Holiness movement. These people had a penchant for Puritan-like codes of dress and behaviour. Most of them had little sympathy for the "superficial, false, and fashionable" Christians whose preoccupation with wealth, social prestige, and religious formalism was repelling. Therefore, they began to advocate a policy of separatism. Besides, a growing cleavage began to develop between some of the denominational leaders and the pro-Holiness sector of their constituency.

Policy of separatism

It was especially difficult for some Methodist bishops, who were used to functioning in a denomination accustomed to tight discipline, to countenance the seeming plethora of "voluntary alliances, leagues, unions, and associations." They were also concerned with the increasing number of Holiness evangelists who were conducting campus revivals, camp meetings, and city-wide crusades without any sort of "official" sanction, as well as with the rapidly expanding Holiness press with its ever growing capacity for promoting second-blessing views. They urged that all such activities be placed under "some official supervision or limitation," but their advice went unheeded. Therefore, in the Bishops' Address for 1894, they denounced all those who proclaimed "holiness as a watchword" and who maintained "holiness meetings, holiness preachers, holiness evangelists, and holiness property." Doubtless such pronouncements only served to hasten the schismatic process that had begun in 1843, when 6,000 dissidents withdrew to found the Wesleyan Methodist Church.

Emergence of Holiness groups. Accordingly, between 1880 and World War I a number of new Holiness groups emerged. Some, such as the Church of God (Anderson, Ind.), were established to protest against bureaucratic denominationalism. Others, such as the Christian and Missionary Alliance and the Church of the Nazarene, were organized to serve the spiritual and social needs of the urban poor, who quite frequently were ignored by the stylish downtown congregations of churches, representing the so-called mainstream of Protestantism. Substantially all of these Holiness bodies arose in order to facilitate the proclamation of a second-blessing experience of sanctification with its concomitants—a life of separation and practical holiness, views that apparently could no longer be endorsed by or tolerated within the larger, more established denominations. On the whole, these newly emergent Holiness groups were destined to have limited spheres of influence.

Holiness associations

The Metropolitan Church Association, an outgrowth of the Metropolitan Methodist Church of Chicago, was organized in 1894 for the purpose of serving the helpless and the outcast in the densely populated environs of that city. A recent report reflects its modest rate of growth:

15 churches with an inclusive membership of 443. In 1896 the Church of God (Apostolic) was established at Danville, Ky. Presently 22 churches, principally located in West Virginia and South Carolina, with a total constituency of 600 make up this branch of the Holiness movement. The Missionary Church Association—known as the Missionary Church since its merger with the United Missionary Church in 1969—was organized in 1898 to promote the cause of Holiness in Pennsylvania, Michigan, Indiana, and Ohio, states wherein most of its 278 churches and 17,700 members are still located. Two years later the Lumber Mission Conference of the Holiness Methodist Church was established "to bring new emphasis to home missions and scriptural holiness" to the rural areas of North Carolina. At last count, it was supervising three churches with a membership of 240.

During the pre-World War I years, still other Holiness churches appeared on the American religious scene: one, the Pillar of Fire, in 1901, was largely the handiwork of a dynamic woman, Alma White, who, despite the opposition of the Methodist Church in which her husband was ministering, organized the Pentecostal Union, which later changed its name to that mentioned above, in order to spread "Scriptural holiness similar to the societies which John Wesley organized." Mrs. White's work survives chiefly in New Jersey, Pennsylvania, and Colorado, where there are 61 churches and 5,100 members. Another group, the Churches of Christ in Christian Union, traces its beginning to 1909, when a controversy over holiness wracked the Council of the Christian Union Church. The pro-Holiness contingent withdrew to form this new denomination, which, over the years, has concentrated its efforts in Ohio, where in 1968 its 7,930 members attended 242 churches. The last group of this type, the Church of God (Holiness), dates back to 1914, when Rev. K.H. Burruss began preaching the gospel of entire sanctification to a congregation of eight people in Atlanta, Ga. Recent statistics show that this small Holiness denomination now comprises 42 churches and 25,600 members, primarily in Georgia and South Carolina.

Although most of the aforesaid churches are relatively small and only locally or regionally important, a number of others that make up the Holiness movement have demonstrated a remarkable capacity for sustained growth. Among these, one might include the "older" denominations—the Wesleyan Church (82,400 members) and the Free Methodist Church of North America (63,000)—as well as the newer ones: the Church of God (Anderson, Ind.; 146,800), the Christian and Missionary Alliance (119,800), the Salvation Army (329,500), and the Church of the Nazarene (364,800). The Church of the Nazarene, whose members constitute nearly a third of the total membership of the Holiness movement, is generally recognized as being its most influential representative.

Church of the Nazarene

The Church of the Nazarene was founded in Los Angeles in 1895 by Phineas Bresee, who had been a pastor and a presiding elder in the Methodist Church. Although he did testify that he experienced "a transformed condition of life and blessing and unction and glory, which I had never known before," the chief cause for leaving the Methodists was apparently criticism of his methods for evangelizing the poor in a Los Angeles mission called Peniel Hall. When his superiors removed him from the supervision of this project, he and 82 people, most of whom were recent converts from the slums, joined together to organize the Church of the Nazarene. From the outset, Bresee announced that this new venture would be "a simple, primitive church, a church of the people and for the people." In the years that followed, Bresee's efforts prospered. By 1903 he was pastoring a congregation of 1,500 and supervising the work of a host of other Nazarene churches in the environs of Los Angeles. Shortly thereafter, in 1907, this group of churches that was headed by Bresee united with the Association of Pentecostal Churches of America, which had been founded in New York in 1895. A year later, 1908, the recently formed Pentecostal Church of the Nazarene voted to join the Holiness Church of Christ, an organization of Southern Holiness churches. This merger added a substantial number of churches to the expanding denomination; more important, perhaps, it climaxed the process of geographical consolidation, which for the first time gave the church a truly national character. When the term Pentecostal began to be used to identify that segment of the Holiness movement that stressed charismatic manifestations such as glossolalia (speaking in tongues), healings, and prophesyings as normative outward evidences of the work of the Holy Spirit in the heart of the believer, the General Assembly of 1919 voted to delete the word Pentecostal from the official church name. Henceforth the denomination would be known simply as the Church of the Nazarene.

BELIEFS AND PRACTICES

Much difficulty is encountered in describing those elements of faith and practice that are held in common by all Holiness groups, principally because so many different individuals representing such a wide variety of religious traditions and revivalistic and pietistic movements had a part in shaping the Holiness movement. For example, there were men such as Charles G. Finney, the Presbyterian-Congregational revivalist and president of Oberlin College, who, during his 19th-century revival campaigns, popularized the notion that "a mighty Baptism of the Holy Spirit" could transform the lethargic spiritual life of a believer. Likewise, Asa Mahan, a Congregationalist and also onetime president of Oberlin College, who, contrary to the generally accepted Wesleyan view that sanctification provided absolute perfection instantaneously, interpreted sanctification simply as an experience that allowed the believer "to grow in grace" toward perfection. Others were William E. Boardman, an American Presbyterian whose volume *The Higher Christian Life* (1859) kindled a widespread interest in perfection among non-Methodists; A.B. Earle, who was described as the most prominent Baptist evangelist of the antebellum period; Society of Friends evangelist David B. Updegraff; and Dougan Clark, professor of Bible at Earlham College. All of the aforesaid people, together with a host of others, either were influenced by the Holiness revival or, in turn, exerted a profound influence upon its development. Doubtless they were chiefly responsible for the differing doctrinal emphases and interpretations as well as the diversities in polity and practice that are manifested by the churches that make up the Holiness movement.

Individuals influencing Holiness beliefs and practices

Theological positions. These diversities notwithstanding, it is possible to provide a general estimate of the theological stance of most contemporary Holiness bodies.

Most of the churches claim loyalty to the historic Wesleyan tradition. Time and again, statements such as these appear in their official declarations: "In doctrine the Church of the Nazarene is essentially in accord with historic Methodism." "The great cardinal doctrines of Christianity as interpreted in the general standards of Methodism are received by this church" (the Wesleyan Church). Accordingly, most Holiness churches, patterning their beliefs on the tenets of Methodism, ignore the stricter doctrines of Calvinism, predestination and reprobation, and accept the milder emphases of Arminianism as regards repentance, faith, and holiness. In addition, most Holiness churches still stand resolutely by John Wesley's views regarding Christian perfection, despite the fact that such principles have become less vitally important to the majority of 20th-century Methodists. Having been affected by 19th-century pietism and revivalism, contemporary Holiness churches tend to stand closer, doctrinally speaking, to Fundamentalism than to their Methodist antecedents. In examining their tenets, one encounters such evidences of conservative evangelical belief as "plenary inspiration (verbal inspiration of the whole Bible)," "Christ's atonement for the entire human race," and "the personal second coming of Christ." In the doctrinal statements of a few churches—Church of the Nazarene and Christian and Missionary Alliance—brief allusions to divine healing and a Pentecostal experience do appear. However, these must not be construed as

Wesleyan theological tenets

constituting sufficient grounds for their being identified with the Pentecostal movement, the so-called left wing of perfectionism, against which, in fact, many right-wing Holiness groups have inveighed.

Social, economic, and political positions. Needless to say, the task of defining the position of Holiness churches on nondoctrinal matters such as interchurch relations, ecumenism, missionary and evangelistic activity, styles of worship, and socioeconomic problems is an equally enigmatic one. As noted above, some Holiness bodies have remained numerically small and have spheres of influence that are merely local or regional. As might be expected, these groups have retained their sectarian character: they shun all forms of worldliness—dancing, theatre attendance, card playing, lotteries, membership in oathbound secret orders, "superfluous ornaments and costly apparel." They deprecate both internal organization as well as extramural fellowship with other national and international bodies of Christians, and they make little attempt to grapple with the intricate socioeconomic problems that have emerged in our highly industrialized, urbanized, secularized society. On the other hand, a number of other Holiness churches have grown in size, wealth, and administrative complexity. These have renounced or have modified many of the so-called sectarian practices and attitudes that, initially, had set them apart from other religious groups. Thus, many of the critical appraisals of Holiness churches that were made concerning the early years of their development, and, moreover, that are still rather widely circulated, are no longer creditable. No longer do critics describe Holiness churches as refuges of the urban poor, the crassly emotional, or the anti-intellectual rural "come-outers" with their narrow-minded antipathy toward worldliness.

Sectarian versus accommodating churches

According to recent studies there may be no existing denomination more loyal to the principle of Christian perfection and yet more free from emotional extremes than the Wesleyan Methodist Church. Furthermore, forceful, personal evangelism and boisterous revival meetings are apparently giving way to make room for greater emphases on Christian education and a trained clergy and laity. This is attested to by the fact that most of the larger Holiness groups presently operate a dozen reputable, regionally accredited colleges and seminaries. Besides this, the process of accommodation is causing conflicts among constituents of the same denomination to be eliminated and tensions between members of different religious groups to be reduced. Thus, whereas early leaders of the Church of God (Anderson, Indiana) believed that all religious organizations were Satanic conceptions and refused to form one of their own or to cooperate with any other, recent administrators have encouraged churches to join local councils of churches and pastors to become members of ministerial associations. Moreover, while the denomination is not a member of the National Council of Churches, each of its general agencies is integrated into the appropriate divisions of that council. Although the primary objective of Holiness churches has been the spiritual regeneration of mankind, one should not gloss over the correlative activity of groups like the Salvation Army in the area of social service. In striving to meet the needs of the errant, the wayward, the orphaned, or the unemployed "down-and-outer" by establishing settlement houses, industrial centres, orphanages, rescue missions, and hospitals, the Salvation Army has challenged the often repeated contention that Holiness churches lack a vital social concern.

Salvation Army

Holiness churches have been experiencing phenomenal growth. (In the 30-year period 1933–63, the Nazarenes multiplied their membership by 80 percent.) And, unquestionably, Holiness churches have experienced dramatic change. In the eyes of most observers, however, they have remained constant in their devotion to a distinctive doctrine—the postconversion experience of sanctification. Moreover, they have not deviated appreciably from their position regarding a life of personal holiness with its concomitant renunciation of "worldly" preoccupations such as tobacco, liquor, movies, and high fashion. Camp meetings, revivals, and the mourners' bench, where sinful men can be restored to holiness of heart and life, still remain as hallmarks of the contemporary Holiness movement. Amid the pervasive intellectualism, formalism, and moral ineffectuality that seemed to characterize the American religious scene during the post-Civil War period, as well as during the early decades of the 20th century, these "Spirit centred" churches have survived. More than that, they have so flourished that they have been lauded as constituting a "third force" (in conjunction with traditional Protestantism and Roman Catholicism) in American Christianity.

BIBLIOGRAPHY

Doctrinal works: WILLIAM ARNETT, "Current Theological Emphases in the American Holiness Tradition," *Mennonite Quarterly Review*, 35:120–129 (1961); CHARLES EWING BROWN, *The Meaning of Sanctification* (1945); D. SHELBY CORLETT, *The Meaning of Holiness* (1944); JOHN L. PETERS, *Christian Perfection and American Methodism* (1957); A.B. SIMPSON, *The Fourfold Gospel* (1925), *Wholly Sanctified* (1925); BENJAMIN B. WARFIELD, *Studies in Christian Perfection*, vol. 1 (1931).

Historical studies: General works include EMORY S. BUCKE (ed.), *The History of American Methodism*, vol. 3, pp. 606–627 (1964); ELMER T. CLARK, *Small Sects in America*, rev. ed., pp. 51–84 (1949); FREDERICK E. MAYER, *Religious Bodies of America*, 4th rev. ed. (1961); and TIMOTHY L. SMITH, *Revivalism and Social Reform in Mid-Nineteenth-Century America* (1957). For information on the Church of God, see VAL B. CLEAR, "The Urbanization of a Holiness Body," in ROBERT LEE (ed.), *Cities and Churches: Readings on the Urban Church* (1962); JOHN W. SMITH, *Heralds of a Brighter Day: Biographical Sketches of Early Leaders in the Church of God Reformation Movement* (1955); and CHARLES EWING BROWN, "The Church of God," in VERGILIUS FERM (ed.), *The American Church of the Protestant Heritage*, pp. 435–454 (1953). On the Church of the Nazarene, see M.E. REDFORD, *Rise of the Church of the Nazarene* (1951); and TIMOTHY L. SMITH, *Called Unto Holiness: The Story of the Nazarenes* (1962). On the Salvation Army, see HARRY EDWARD NEAL, *The Hallelujah Army* (1961).

(J.T.N.)

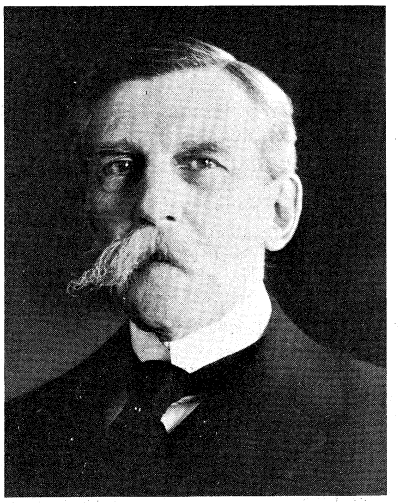
Holmes, Oliver Wendell

A distinguished historian and philosopher of the law and for 30 years a justice of the United States Supreme Court, Oliver Wendell Holmes, Jr., earned renown as a persuasive advocate of judicial restraint. He was born in Boston, Massachusetts, on March 8, 1841, the first child of the celebrated writer and physician Oliver Wendell Holmes. The family background on both sides represented the New England "aristocracy" of character and accomplishment. The Doctor was descended from the Puritan poetess Anne Bradstreet; he married Amelia Lee Jackson, whose father, Charles, was a justice of the Supreme Judicial Court of the State of Massachusetts, a bench on which Oliver Wendell Holmes, Jr., was to sit for 20 years. He was proud of this heritage and spoke of it often. It helped shape his mind and character.

Young Holmes went to a private school and then to Harvard College. He was graduated in the class of 1861 and like his father before him was class poet. At the outbreak of the Civil War he enlisted as a private in the 4th Battalion of Infantry and began training at Boston's Fort Independence, not expecting to finish the academic year or take his degree. The battalion was not called up, and after graduation the young man applied for and received, in July, a commission as first lieutenant in the 20th Massachusetts Regiment of Volunteers. He was 20 years old at that time.

Service in Civil War

His letters and diary give vivid pictures of his war experiences. He was seriously wounded three times, at the battles of Ball's Bluff, Antietam, and Chancellorsville. He left the army after three years, having been commissioned lieutenant colonel although mustered out with the rank of captain. Holmes described war as "an organized bore." He said, "I trust I did my duty as a soldier respectably, but I was not born for it and did nothing remarkable in that way." In a Memorial Day address to fellow veterans, in 1884, he attributed a certain value to



Holmes.
By courtesy of the Library of Congress, Washington, D.C.

the war experience: "Through our great good fortune, in our youth our hearts were touched with fire. It was given to us to learn at the outset that life is a profound and passionate thing." This is an aspect of his conviction that ". . . it is required of a man that he should share the passion and action of his time at peril of being judged never to have lived."

Legal studies

In the autumn of 1864 he entered Harvard Law School, ironically without any clear sense of vocation. He had even contemplated medicine, to which his father objected. On different occasions, he said that his "Governor" "put on the screws to have me go to the Law School" or "kicked" him into it. There is a story that when young Holmes announced to his father the decision to enter the law school, the Doctor said, "What's the use of that, Wendell? A lawyer can't be a great man." There was not a deep affinity between father and son. The little Doctor's puns and quips, his easy display of emotion, and a somewhat patronizing attitude chafed the tall, less talkative, inherently shy law student. The philosopher William James, perhaps the closest friend of Wendell in the immediate postwar years, once remarked that "no love is lost" between father and son.

Holmes experienced a certain restlessness in law school, finding the tradition of the law as presented in an uninspired curriculum to be stagnant and narrowly precedent centred. The science, philosophy, or history of law were slighted, and these, rather than what he later called "the small change of legal thought," were what captured Holmes's mind and drew him into the depths of a profession toward which at first he had not felt a powerful incentive.

After finishing law school in 1866 he made the conventional "pilgrimage" abroad, visiting England, France, and Switzerland and meeting a variety of distinguished men. He was admitted to the bar in 1867 and for 15 years practiced law as a member of several firms. From 1870 to 1873 he was an editor of the *American Law Review*. He edited the 12th edition of the classic survey of early American law, Chancellor James Kent's (1763–1847) *Commentaries on American Law* (1873). He also lectured at Harvard on law.

During this busy time he was engaged in courtship. Always something of a ladies' man, he had maintained a long friendship with Fanny Bowditch Dixwell, daughter of his onetime schoolmaster. She had waited patiently through wartime, his law studies, travel, and apprenticeship. Holmes and Fanny were married at last on June 17, 1872. The marriage, happy and long lasting, was childless.

The Common Law

In 1880–81 Holmes was invited to lecture on the common law at the Lowell Institute in Boston, and from these addresses developed his book *The Common Law* (1881). Here the genius of Holmes was first clearly revealed and the consistent direction of his thought made evident. A fresh voice was speaking in his words:

> The life of the law has not been logic: it has been experience. The felt necessities of the time, the prevalent moral and political theories, intuitions of public policy, avowed or unconscious, even the prejudices which judges share with their fellow-men, have had a good deal more to do than the syllogism in determining the rules by which men should be governed. The law embodies the story of a nation's development through many centuries, and it cannot be dealt with as if it contained only the axioms and corollaries of a book of mathematics. In order to know what it is, we must know what it has been, and what it tends to become. We must alternately consult history and existing theories of legislation. But the most difficult labor will be to understand the combination of the two into new products at every stage. The substance of the law at any given time pretty nearly corresponds, so far as it goes, with what is then understood to be convenient; but its form and machinery, and the degree to which it is able to work out desired results, depend very much upon its past.

In January 1882 Holmes was made Weld Professor of Law, a chair established for him at Harvard Law School. In December of the same year he accepted appointment to the Supreme Judicial Court of the State of Massachusetts, knowing the judgeship was his destiny and the function through which he could most influence the development of law. He was to sit on that bench for 20 years, becoming its chief justice in 1899. In 1902 President Theodore Roosevelt appointed him associate justice of the United States Supreme Court. He sat on that court to a more advanced age than did any other man, retiring on January 12, 1932, soon before his 91st birthday.

Fanny Holmes, devoted, witty, wise, tactful, perceptive, died on April 30, 1929. Holmes wrote to his intimate friend, the English jurist Sir Frederick Pollock, "For sixty years she made life poetry for me and at 88 one must be ready for the end. I shall keep at work and interested while it lasts—though not caring very much for how long." He died two days before his 94th birthday on March 6, 1935.

Justice of the Supreme Court

In that long span of years on the Supreme Court he became acknowledged as one of the most notable jurists of the age—in the opinion of many the foremost. Often he has been called "The Great Dissenter" because of the brilliance of his dissenting opinions, but the phrase gives a falsely negative emphasis, and his penetration and originality are seen as fully in the opinions in which he expressed or concurred in the majority view of the court as in those in which he was in dissent.

Holmes believed that the making of laws is the business of legislative bodies, not of courts, and that within constitutional bounds the people have a right to whatever laws they choose to make, good or bad, through their elected representatives. He stated the concept of "clear and present danger" as the only basis for curtailing the right of freedom of speech, illustrating it with the homely example: "The most stringent protection of free speech would not protect a man in falsely shouting fire in a theatre and causing a panic."

He wrote that "the best test of truth is the power of the thought to get itself accepted in the competition of the market. . . . That at any rate is the theory of our Constitution." Again: "If there is any principle of the Constitution that more imperatively calls for attachment than any other it is the principle of free thought—not free thought for those who agree with us but freedom for the thought that we hate."

A man austerely dedicated to his work, he also enjoyed the earthy and the droll. He loved Rabelais. Sometimes in Washington he attended burlesque shows and was said to have remarked, "I thank God I am a man of low tastes." The newly inaugurated President Franklin D. Roosevelt called upon the retired justice and found him reading Plato. "Why do you read Plato, Mr. Justice?" "To improve my mind, Mr. President," replied the 92-year-old man.

Holmes won the love and admiration of generations of lawyers and judges in his long career. When he resigned

from the Supreme Court, his "brethren," as he always addressed his fellow justices, wrote him a letter signed by all, saying in part:

> Your profound learning and philosophic outlook have found expression in opinions which have become classic, enriching the literature of the law as well as its substance. . . . While we are losing the privilege of daily companionship, the most precious memories of your unfailing kindliness and generous nature abide with us, and these memories will ever be one of the choicest traditions of the Court.

MAJOR WORKS

Commentaries on American Law by James Kent, 12th ed. edited by Oliver Wendell Holmes, Jr. (1873); *The Common Law* (1881, 1938); *Speeches* (1891, 1913, and 1938); *Collected Legal Papers*, ed. by Harold Laski (1920); *The Dissenting Opinions of Mr. Justice Holmes* (1929); *Representative Opinions of Mr. Justice Holmes*, ed. by A. Lief (1931); *Justice Oliver Wendell Holmes: His Book Notices and Uncollected Letters and Papers*, ed. by Harry C. Shriver (1936); *Holmes-Pollock Letters* (1941), *Touched with Fire: Civil War Letters and Diary of Oliver Wendell Holmes, Jr., 1861–1864* (1946), *Holmes-Laski Letters* (1953), and *The Occasional Speeches of Justice Oliver Wendell Holmes* (1962), all ed. by Mark de Wolfe Howe.

BIBLIOGRAPHY. The principal biography, not yet concluded, is MARK DEWOLFE HOWE, *Justice Oliver Wendell Holmes*, vol. 1, *The Shaping Years, 1841–1870* (1957), vol. 2, *The Proving Years, 1870–1882* (1963). Others are: CATHERINE DRINKER BOWEN, *Yankee from Olympus: Justice Holmes and His Family* (1944); FRANCIS BIDDLE, *Mr. Justice Holmes* (1942); and SILAS BENT, *Justice Oliver Wendell Holmes* (1932). A good selection from Holmes's writings is *The Mind and Faith of Justice Holmes*, edited with Introduction and Commentary by MAX LERNER (1943). Both appraisal and biographical material may be found in FELIX FRANKFURTER (ed.), *Mr. Justice Holmes: A Collection of Essays* (1931), and *Mr. Justice Holmes and the Supreme Court* (1938).

(E.Fu.)

Holocene Epoch

The Holocene Epoch, sometimes referred to as the Recent, is the latest interval of geological time, covering approximately the last 10,000 years of the earth's history. The sediments of the Holocene, both continental and marine, cover the largest area of the globe of any epoch in the geological record, but the Holocene is unique because it is coincident with the late and post-Stone Age history of man. The influence of man is of world extent and is so profound that it is appropriate to have a special geologic name for this time. In 1833 Sir Charles Lyell proposed the name Recent for the period that has elapsed since "the earth has been tenanted by man." It is now known that man has been in existence a great deal longer. The term Holocene was proposed in 1867 and was formally submitted to the International Geological Congress at Bologna in 1885. It was officially endorsed by the U.S. Commission on Stratigraphic Nomenclature in 1969.

The Holocene is also the latest division of the Quaternary Period, about the last 2,000,000 years of geologic time and often referred to as the Great Ice Age. A characteristic of ice ages is the permanent ice cover of some areas, like Antarctica today, while other regions in more temperate latitudes are invaded by continental glaciers repeatedly in an oscillatory manner. Such oscillatory cycles, called glacial and interglacial stages, are further modulated by minor oscillations, called stadials and interstadials (or stades and interstades). The Holocene Epoch is the latest interglacial interval of the Quaternary Period. The preceding and much longer sequence of alternating glacial and interglacial ages is referred to as the Pleistocene Epoch. Because there is nothing to suggest that the Pleistocene, or Great Ice Age, is now ended, some authorities prefer to extend the Pleistocene up to the present time; this tends to ignore man and his geologic role, however.

The Holocene forms the chronological framework for human history. Archaeologists use it as the time standard against which they trace the evolution of man, his economy, and his art and architecture. At the beginning of the Holocene, there were faunas such as the woolly mammoth, the giant elk, the cave bear, and many other species of large mammals, all of which are now extinct. Some scientists think that these animals were exterminated by the sudden change in conditions due to a shift of climatic belts, but others consider that man was responsible. The role of man in the disappearance of these animals is not proven, though sometimes his artifacts occur in close association with their skeletal remains.

This article treats Holocene stratigraphy, chronology, and climate, and the diverse manifestations of the climatic change from Pleistocene to Holocene time in continental and marine areas. For additional information on the place of the Holocene Epoch in early history, see EARTH, GEOLOGICAL HISTORY OF; CENOZOIC ERA; and STRATIGRAPHIC BOUNDARIES; for details of the glaciation that occurred prior to Holocene time and discussion of the climatic factors involved, see PLEISTOCENE EPOCH; and CLIMATIC CHANGE. See also articles on specific landforms or natural features that reflect Pleistocene and Holocene environmental conditions (*e.g.*, RIVER DELTAS; SOILS; CONTINENTAL SHELF AND SLOPE; LAKES AND LAKE SYSTEMS; DESERTS).

HOLOCENE STRATIGRAPHY

Chronology and correlation. The Holocene is unique among geological epochs because varied means of correlating deposits and establishing chronologies are available. One of the most important means is radiocarbon dating. A radiocarbon date is usually expressed in years BP (before the present), the reference year being AD 1950, and the date depends upon the radioactive decay of carbon originally present in organic matter (see DATING, RELATIVE AND ABSOLUTE). Since 1950, all vegetation, bones, shells, and other organic matter containing carbon have been contaminated to some extent by radiocarbon created by nuclear explosions. Studies of wood from ancient tree rings show, however, that there were only small variations in the radiocarbon values (amounting to several percent) during the past. Because the date or age that is provided by the geochemical procedure may be appreciably different from the true age, it is customary to refer to those dates in "radiocarbon years." These dates, obtained from a variety of deposits, form an important framework for Holocene stratigraphy and chronology.

Radiocarbon dating

The limitations of geochemical accuracy are expressed as ± a few tens or hundreds of years, but in addition to this calculated error, there also is question of error due to contamination of the material measured. For instance, an ancient peat may contain some younger roots and thus given a falsely "young" age unless it is carefully collected and treated to remove contaminants. Marine shells consist of calcium carbonate ($CaCO_3$), and in certain coastal regions there is upwelling of deep oceanic water that can be 500 to over 1,000 years old. An "age" from living shells in such an area can suggest that they are already hundreds of years old.

Table 1 shows the comparative dates of radiocarbon years and those obtained by other means. Two sets of radiocarbon years are given because the half-life (time required for one-half of a unit amount of an element to decay due to radioactivity) of carbon-14 was reassigned a value of 5,730 years by agreement of scientists. Many dates available in the literature, however, are based on the originally established half-life of 5,570 years.

In certain areas a varve chronology can be established. This involves counting and measuring thicknesses in annual paired layers of lake sediments (*varv* in Swedish) deposited in lakes that suffer an annual freeze-up (see VARVED DEPOSITS). Because each year's sediment accumulation varies in thickness according to the climatic conditions of the melt season, any long sequence of varve measurements provides a distinctive "signature" and can be correlated for moderate distances from lake basin to lake basin. The pioneer in this work was the Swedish investigator Baron Gerard de Geer (1858–1943), who developed a long chronology on which that shown in Table 1 is partly based.

Table1: Comparative Dating Systems for the Holocene Epoch* (and latest Pleistocene)

uncorrected radiocarbon BP dates (radiocarbon years)		U.S. tree-ring dates in absolute years BP (before present, AD 1950, in sidereal years, adjusted to radiocarbon according to Damon)‡	AD/BC dates (sidereal years)	conventional BP varve years (estimated to be 350 ± 200 years too young)§
(T_2=5570)†	(T_2=5730)†			
1000	1000	900	AD 1050	...
2000	2050	1950	AD 0	(1600)
3000	3100	3250	1300 BC	(2900)
4000	4150	4650	2700	(4300)
5000	5180	5920	3970	(5630)
6000	6200	6900	4950	(6550)
7000	7220	(7450)	5550	7100
8000	8240	(8350)	6400	8000
9000	9270	(9200)	7250	8850
10,000	10,300	(10,550)	7600	10,200
11,000	11,330	(11,550)	9600	11,200
12,000	12,350	(12,550)	10,600	12,200

*Wood from the tomb of the pharaoh Djoser at Saqqārah, Egypt, is dated historically at 4650 ± 75 BP in sidereal years, but according to multiple analyses by many laboratories it is about 4100 BP in radiocarbon years, or 550 years too young. The anomaly is most probably explained in terms of solar radiation, residence time of CO_2, and paleomagnetics. Tree-ring dating combined with ^{14}C measurements has confirmed this trend and provides a general curve for correcting Holocene ^{14}C dates. Almost all radiocarbon dates given are uncorrected. U.S. tree-ring chronology most closely resembles the sidereal dates; the Scandinavian varve chronology also is close to the astronomical chronology, subject to a 350 ± 200 year correction. †T_2 signifies half-life. ‡Dates older than 7450 BP are based on varve years, corrected by 350 ± 200 years. §Dates for varve years less than 6550 BP are extrapolated.

In some relatively recent continental deposits, obsidian, a black, glassy rock of volcanic origin, can be used for dating. Obsidian weathers slowly at a uniform rate, and the thickness of the weathered layer is measured microscopically and gauged against known standards to give a date in years. This has been particularly useful where arrowheads of obsidian are included in deposits.

Paleomagnetism is another phenomenon used in chronology. The earth's magnetic field undergoes a secular shift that is fairly well known for the last 2,000 years. The magnetized material to be studied can be natural, such as a lava flow; or it may be man-made, for example an ancient brick kiln or smeltery that has cooled and thus fixed its magnetic orientation of the bricks to correspond to the geomagnetic field of that time. See ROCK MAGNETISM for the methods involved in this type of dating.

Another form of dating is tephrochronology, so called because it employs the tephra (ash layers) generated by volcanic eruptions. The wind may blow the ash 1,500–3,000 kilometres (1,000–2,000 miles), and, because the minerals or volcanic glass from any one eruptive cycle tend to be distinctive from those of any other cycle, even from the same volcano, these can be dated from the associated lavas by stratigraphic methods (with or without absolute dating). The ash layer then can be traced as a "time horizon" wherever it has been preserved. When the Mt. Mazama volcano in Oregon exploded at about 6600 BP (radiocarbon dated by burned wood), 70 cubic kilometres (16 cubic miles) of debris were thrown into the air, forming the basin now occupied by Crater Lake. The tephra were distributed over ten states, thereby providing a chronologic marker horizon. A comparable eruption of Thera on Santorin in the Aegean Sea about 3,400 years ago left tephra in the deep-sea sediments and on adjacent land areas. Periodic eruptions of Mt. Hekla in Iceland have been of use in Scandinavia, which lies downwind.

Biological dating

The most important biological means of establishing Holocene chronology involves palynology, the study of pollen, spores, and other microscopic organic particles. Pollen from trees, shrubs, or grasses is generated annually in large quantities and often is well preserved in fine-grained lake, swamp, or marine sediments. Statistical correlations of modern and fossil assemblages provide a basis for estimating the approximate makeup of the local or regional vegetation through time. Even a crude subdivision into tree pollen (AP) and nonarboreal pollen (NAP) reflects the former types of climate (see POLLEN STRATIGRAPHY). The tundra vegetation of the last glacial epoch, for example, provides predominantly NAP, and the transition to forest vegetation shows the climatic amelioration that heralded the beginning of the Holocene.

The first standard palynological stratigraphy was developed in Scandinavia by Axel Blytt (1843–98), Johan Rutger Sernander (1866–1944), and E.J. Lennart von Post (1884–1951), in combination with a theory of Holocene climate changes. The so-called Blytt–Sernander system was soon tied to the archaeology and to the varve chronology of Gerard de Geer. It has been closely checked by radiocarbon dating, establishing a very useful standard. Every region has its own standard pollen stratigraphy, but these are now correlated approximately with the Blytt–Sernander framework. To some extent this is even true for remote areas such as Patagonia and East Africa. Indeed, it was by palynology that the climatic events of Scandinavia were demonstrated to match changes in equatorial mountain regions and in the Southern Hemisphere.

The Blytt–Sernander classification, together with standard pollen zone codes, is indicated in Table 2, together with an estimation of mean annual temperature departures for midlatitude countries. Particularly important is the fact that the middle Holocene was appreciably warmer than today. In Europe this phase has been called the Climatic Optimum (zones Boreal-1 to Atlantic-2), and in North America (zones B-2 to C-2) it has been called the hypsithermal (also altithermal and xerothermic).

Finally, the measurement and analysis of tree rings, or dendrochronology, must be mentioned. The age of a tree that has grown in any region with a seasonal contrast in climate can be established by counting its growth rings. Work in this field by the University of Arizona's Laboratory of Tree-Ring Research, by selection of both living trees and deadwood, has carried the year-by-year chronology back over 7,500 years. As with palynology, certain pitfalls have been discovered in tree-ring analysis. Sometimes, as in a very severe season, a growth ring may not form. In certain latitudes the tree's ring growth correlates with moisture, but in others it may be correlated with temperature. From the climatic viewpoint these two parameters are often inversely related in different regions. Nevertheless, in experienced hands, just as with varve counting from adjacent lakes, ring measurements from trees with overlapping ages can extend chronologies back for many thousands of years. The bristle-cone pine of the White Mountains in California has proved to be singularly long-lived and suitable for this chronology; some individuals still living are over 4,000 years old, certainly the oldest living organisms. Wood from old buildings and even old paving blocks in western Europe and in the U.S.S.R. have contributed to the chronology. This technique not only offers an additional means of dating but also contains a built-in documentation of climatic characteristics. In certain favourable situations, particularly in the drier, low latitudes, tree-ring records sometimes document 11- and 22-year sunspot cycles.

The Pleistocene–Holocene boundary. Arguments can be presented for the selection of the lower boundary of the Holocene at several different times in the past. Some Soviet specialists have proposed a boundary at the beginning of the Allerød, a warm interstadial age that began about 12,000 BP. Others, in Alaska, proposed a Holocene section beginning at 6,000 BP. Marine geologists have recognized a worldwide change in the character of deep-sea sedimentation about 10,000–11,000 BP. In warm tropical waters, the clays show a sharp change at this time from chlorite-rich particles often associated with fresh feldspar grains (cold, dry climate indicators) to kaolinite and gibbsite (warm, wet climate indicators).

Some of the best preserved traces of the boundary are found in southern Scandinavia, where the transition from the latest glacial stage of the Pleistocene to the Holocene was accompanied by a marine transgression. These beds, south of Göteborg, have been uplifted and are exposed at

Table 2: Generalized Subdivisions of the Holocene in Southern Scandinavia, with Corresponding Marine Stages and Other Correlations

Blytt–Sernander divisions	pollen zones* T. Nilsson	pollen zones* K. Jessen	C^{14} "ages"	climate*	forest vegetation	marine episodes Baltic Sea	marine episodes Kattegat	marine stratigraphic stages† in southern Scandinavia		marine stratigraphic stages† in southern North Sea‡	stratigraphic code§
Sub-Atlantic	SA-2 SA-1	IX	BP 1,250	cool and humid	beech (arable land and heath)	Mya Sea	Mya Sea		T	III	
									R	II	
			2,200				Scrobicularia-Mactra Sea		T		
									R	Dunkerquian	
Subboreal	SB-2	VIII		warm and dry	oak–ash (oak–mixed forest) forest clearance	Limnaea Sea		Vendsysselian	T	I	Qhu
	SB-1		3,600						R		
									T	0	
			5,100						R		
									T	IV	
	AT-2								R		
Atlantic		VII		warm and humid	oak–elm (oak–mixed forest)	Litorina Sea	Tapes Sea		T	III	
	AT-1		6,350					Kattegattian	R	Calaisian	
									T	II	Qhm
			7,750						R		
									T	I	
	BO-2	VI			hazel oak				R		
Boreal			8,500	warm and dry		Ancylus Lake					
	BO-1	V			hazel–pine		(*Ostrea*)	Varbergian	T	Ostendian	
			9,500								Qhl
									R		
Preboreal	PB	IV		warmer and moister from cold and dry	birch–pine	Yoldia Sea	Cardium-Mytilus Sea	Gothenburgian	T	—	
			10,100						R		

*Both climatic and pollen zones and archaeological limits are partly diachronous. †Under "marine stratigraphic stages" transgressive and regressive eustatic trends are indicated by T or R. ‡Marine stages in the North Sea basin (mainly northern France, The Netherlands, etc.) are widely used as a world standard. §The stratigraphic code (used for general mapping) employs the initials Q (Quaternary), h (Holocene), l,m,u (lower, middle, upper). ‖Absolute chronology corresponds to sidereal years, corrected from tree-ring, varve, and radiocarbon measurements.
Source: In part based on S. Hansen, 1965; B.P. Hageman, 1969; N.A. Mörner, 1969.

Pleistocene–Holocene boundary dates

the surface. The boundary is dated around 10,300 ± 200 years BP (in radiocarbon years). This boundary marks the very beginning of warmer climates that occurred after the latest minor glacial advance in Scandinavia. This built the last Salpausselkä moraine, which corresponds in part to the Valders Substage in North America. The subsequent warming trend was marked by the Finiglacial retreat in northern Scandinavia, the Ostendian (early Flandrian) marine transgression in northwestern Europe.

NATURE OF THE HOLOCENE RECORD

The very youthfulness of the Holocene stratigraphic sequence makes subdivision difficult. The relative slowness of the earth's crustal movements (usually rates of less than one millimetre per year) means that most areas that contain a complete marine stratigraphic sequence are still submerged. Fortunately, in areas that were depressed by the load of glacial ice there has been progressive postglacial uplift (crustal rebound) that has led to the exposure of the nearshore deposits.

In southern Scandinavia postglacial lakes and bogs lay immediately landward of the shoreline of each stage, so that the pollen classification can be applied to the corresponding marine facies. Some of these old beachlines are richly fossiliferous, and numbers of them have been named after the most distinctive fossil of the assemblage (*e.g.*, *Litorina*, *Limnaea*, *Mya*) and thus are technically "biozones." In contrast to the central and northern glaciated areas, the outer peripheries have been subsiding in recent millennia. In the southern Baltic, for example, the nearshore Holocene beds are found only by boring in vertical successions, and similar facies are traceable in southern Denmark, northwest Germany, The Netherlands, and Belgium. The situation is somewhat similar in the southern New England area of the U.S. and in eastern Canada.

Deep oceanic deposits. The marine realm, apart from covering about 70 percent of the earth's surface, offers far better opportunities than coastal environments for undisturbed preservation of sediments. In deep-sea cores, the boundary usually can be seen at a depth of about 10–30 centimetres, where the Holocene sediments pass downward into material belonging to the late glacial stage of the Pleistocene. The boundary often is marked by a slight change in colour. For example, globigerina ooze, common in the ocean at intermediate depths, is frequently slightly pinkish when it is of Holocene age because of a trace of iron oxides that are characteristic of tropical soils. At greater depth in the section, the globigerina ooze may be grayish because of greater quantities of clay, chlorite, and feldspar that have been introduced from the erosion of semi-arid hinterlands during glacial time.

Character and fossil content of globigerina ooze

During each of the glacial epochs the cooling of the ocean waters led to reduced evaporation and thus fewer clouds, then to lower rainfall, then to reduction of vegetation, and so eventually to the production of relatively more clastic sediments (due to reduced chemical weathering). Furthermore, the worldwide eustatic (glacially related) lowering of sea level caused an acceleration of erosion along the lower courses of all rivers and on exposed continental shelves, so that clastic sedimentation rates in the oceans were higher during glacial stages than during the Holocene. Turbidity currents, generated on a large scale during the low sea-level periods, became much less frequent following the rise of sea level in the Holocene.

Studies of the fossils in the globigerina oozes show that at a depth in the cores that has been radiocarbon dated

Adapted from G. Wollin, D. Ericson, and M. Ewing, "Late Pleistocene Climates Recorded in Atlantic and Pacific Deep-Sea Sediments," in *Late Cenozoic Glacial Ages*, (1971); Yale University Press, New Haven and London

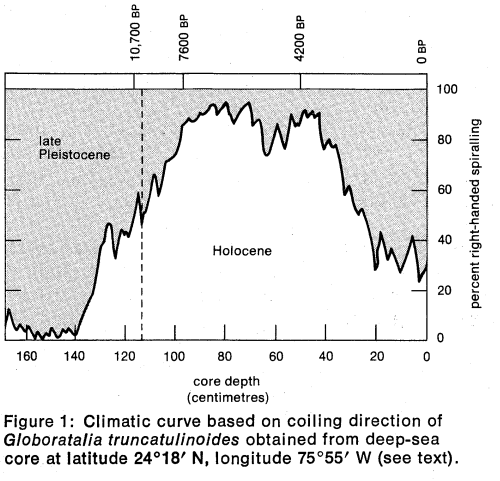


Figure 1: Climatic curve based on coiling direction of *Globoratalia truncatulinoides* obtained from deep-sea core at latitude 24°18′ N, longitude 75°55′ W (see text).

Blytt and Sernander divisions	archaeological periods in Scandinavia		absolute chronology‖ years, sidereal	radiocarbon years, uncorrected BP (1950)
Sub-Atlantic	Historical time		AD 1950	0
	Viking time		AD 1000	1,050
	Iron Age		0	2,000
			600 BC	2,600
Subboreal	Bronze Age			
			1900 BC	3,600
	Neolithic Stone Age			
			4100 BC	5,100
Atlantic	Mesolithic Stone Age	Ertebølle Culture		
			6150 BC	7,750
Boreal		Maglemose Culture		
			7100 BC	9,500
Preboreal		Klosterlund Culture		
			7700 BC	10,100

at about 10,000–11,000 BP the relative number of warm-water planktonic foraminifera increases markedly. In addition, certain foraminiferal species tend to change their coiling direction from a left-handed spiral to a right-handed spiral at this time. This is attributed to the change from cool water to warm water, an extraordinary (and still not understood) physiological reaction to environmental stress. Many of the foraminifera, however, responded to the warming water of the Holocene by migration poleward, by distances of 1,000–3,000 kilometres, in order to remain within their optimal temperature habitats.

It appears also that there was some evolution of new subspecies during the Holocene. Within the Holocene itself, the foraminifera also show a fluctuation in the ratio of warm and cool species, which suggests a sensitivity to relatively minor temperature changes.

Nannoplankton and oceanic circulation

In addition to foraminiferans in the globigerina oozes there are nannoplankton, minute fauna and flora which are mainly coccolithophorids. Research on the present coccolith distribution shows that there is maximum productivity in zones of oceanic upwelling, notably at the subpolar convergence and the equatorial divergence. During the latest glacial stage the subpolar zone was displaced toward the Equator, but with the subsequent warming of waters it shifted back to the borders of the polar regions.

The distribution of the carbonate plankton bears on the problem of rates of oceanic circulation. Is the Holocene rate higher or lower than during the last glacial stage? It has been argued that because of the higher mean temperature gradient in the lower atmosphere from Equator to poles during the last glacial period, there would have been higher wind velocities and, because of the atmosphere–ocean coupling, higher oceanic current velocities. There were, however, two retarding factors for glacial age currents. First, the eustatic withdrawal of oceanic waters from the continental shelves reduced the effective area of the oceans by 8 percent. Second, the greater extent of floating sea ice would have further reduced the available air–ocean coupling surface, especially in the critical zone of the westerly circulation. According to climatic studies by Hubert H. Lamb, British meteorologist, the presence of large continental ice sheets in North America and Eurasia would have introduced a strong blocking action to the normal zonal circulation of the atmosphere, which then would be replaced by more meridional circulation. This in turn would have been appreciably less effective in driving major oceanic current gyres.

Continental shelf and coastal regions. It was recognized as early as 1842 that a logical consequence of a glacial age would be a large-scale withdrawal of ocean water. Consequently, deglaciation would produce a post-glacial "glacioeustatic" transgression of the seas across the continental shelf. The trace of this Holocene rise of sea level was first discerned along the New England coast and along the coast of Belgium, where it was named the Flandrian Transgression by Georges Dubois in 1924. Some writers prefer to call the entire Holocene the "Flandrian Epoch." The transgression itself, however, began at least 5,000 years before the Holocene proper.

Whereas the deep-sea Holocene sediments usually follow without interruption upon those of the Upper Pleistocene, on the continental shelf there is almost invariably a break in the sequence upon the continental formations there. As sea level rose it paused or fluctuated at various stages, leaving erosional terraces, beach deposits, and other indicators of the stillstand. Brief regressions (withdrawal of the seas) in particular permitted the growth of peat deposits that are of significance in the Holocene record because they can be dated by radiocarbon analysis. Dredging in certain places on the shelf, such as off eastern North America, also is useful because terrestrial fossils from the latest glacial period or early Holocene have been found; these range from mammoth and mastodon bones and tusks to human artifacts. On about 70 percent of the world's continental shelves today the amount of sedimentary accumulation since the beginning of the Holocene is minimal, so that dredging or coring operations often disclose hard rock, with older formations at or very close to the surface. In other places, especially near the former continental ice fronts, the shelf is covered by periglacial fluvial sands (meltwater deposits), which, because of their unconsolidated nature, became extensively reworked into beaches and bars during the Holocene Transgression.

Sedimentary deposits and coral reefs

In places of strong shelf currents today many of these deposits still are being dynamically shifted around, disclosing giant sand ripples with amplitudes as great as ten metres.

In warm coral seas the major pauses in the Holocene eustatic rise were long enough for fringing reefs to become established; and when the eustatic rise resumed, the reefs grew upward, either in ribbonlike barriers or from former headlands as patch reefs or shelf atolls. Since coral generally does not colonize a sediment-covered shelf floor at depths of more than about ten metres, those reefs now rising from greater depths must have been emplaced during the early Holocene or must have grown on foundations of ancient reefs (see also CORAL ISLANDS, CORAL REEFS, AND ATOLLS).

Strandlines and the Holocene marine limits

The great ice-covered areas of the Quaternary Period included Antarctica, North America, Greenland, and Eurasia. Of these, Antarctica and Greenland have relatively high latitude situations and do not easily become deglaciated. Some melting occurs, but there is a very great melt-retardation factor in high latitude ice sheets (high albedo or reflectivity, short melt season, etc.). In the case of midlatitude ice sheets, however, once melting starts, the ice disappears at a tremendous rate. The melt rate reached a maximum around 8,000 BP, liberating 18,000,000,000,000 (18×10^{12}) metric tons of meltwater annually. This corresponds to a five-centimetre-per-year rise of sea level. Hand in hand with melting, the sea level responded so that, as the ice began to retreat from its former terminal moraines, the sea began to invade the former coastlands (now the outer continental shelf). Radiocarbon-dated shell beds, for example, from the continental shelf off the eastern United States clearly show the stages of the transgression according to American marine geologist Kenneth Orris Emery and his colleagues at the Woods Hole (Mass.) Oceanographic Institution.

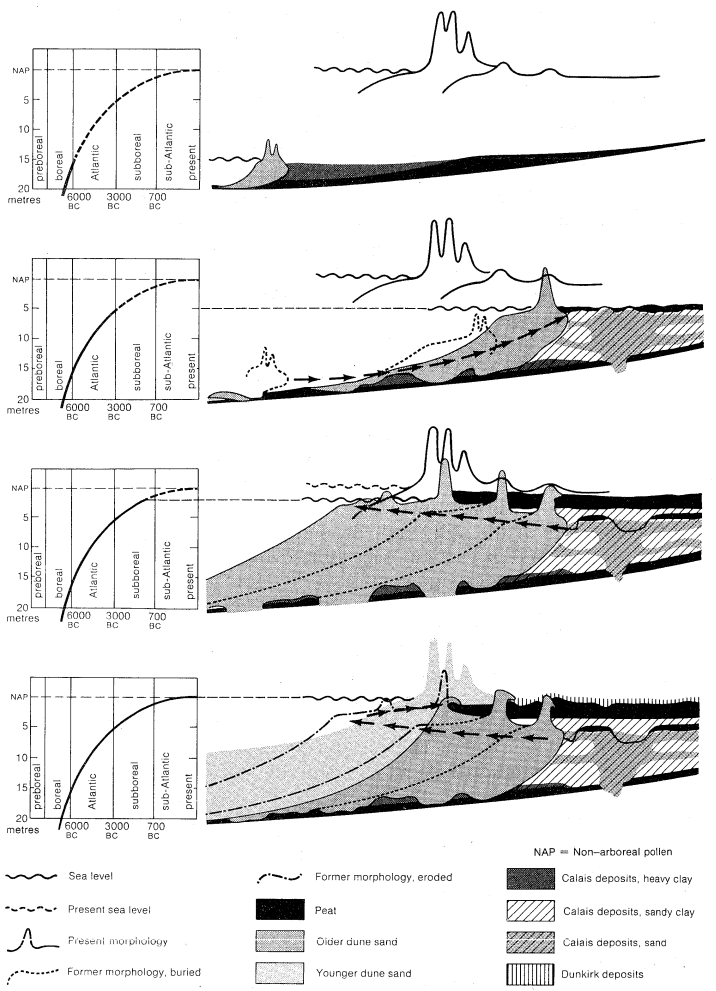


Figure 2: Deltaic accumulation in the western part of The Netherlands during the Holocene.
From B. Hageman, *Geologie en Menjnbouw*, vol. 48 (1969)

As the sea level rose, the earth's crust responded buoyantly to the removal of the load of ice, and at critical times the rate of rise of the water level was outstripped by the rate of rise of the land. In these places the highest ancient shoreline that is now preserved is known as the marine limit. The nearer the former centre of the ice sheet, the higher is the marine limit. In northern Scandinavia, Ontario and northwestern Quebec, around Hudson Bay, and in Baffin Island it reaches more than a 300-metre elevation. In central Maine and Spitsbergen it may exceed 100 metres, whereas in coastal Scotland and Northern Ireland it is rarely above 10–15 metres. A most typical development of the marine limit has been carefully mapped across central Sweden. The age of the beach there is about 7000 BP and is known as Litorina I. This maximum marine limit does not equal the highest eustatic transgression of the Holocene but merely represents the position of the shoreline when the rate of crustal uplift exceeded the eustatic rise. The marine limit varies in height and age in different places.

In addition to the marine-limit strandlines there are row upon row of lower beach levels stretched out across Scandinavia, around Hudson Bay, and on other Arctic coasts. These strandlines are dated and distinctive and do not grade into each other. Each represents a specific period of time when the rising crust and rising sea level remained in place long enough to permit the formation of beaches, spits and bars, and sometimes the erosion of headlands ("fossil cliffs"). The end of beach building at any one strandline occurred when the rise of sea level was reversed. The shoreline rapidly withdrew across the rising landmass and then became re-established at a lower site for another few decades or perhaps centuries.

According to Scandinavian workers, the negative eustatic fluctuations amounted to up to three–five metres. The crust seems to have been rising steadily, and it is principally the eustatic level that fluctuated, because the raised strandlines are roughly parallel and only locally disclose variations due to irregularities in crustal response. In the area of the Kattegat of southwestern Sweden there was a long interplay between the ice front and the transgressing sea in late glacial and early Holocene time. This was closely studied by the Swedish scientist N.A. Mörner, who concluded in 1969 that there was "a striking agreement between the changes in ice recession, climate as recorded by pollen and mollusks, shore-level displacement, isostasy and eustasy."

The old strandlines are associated with well-dated shell beds and swamp deposits. Each high strandline corresponds to a warm phase in the climatic (palynological) history, and every negative fluctuation, namely the retreat stage, during which no beach was preserved, to a cool phase. It is believed that each oscillation was controlled by advance of glaciers situated mainly in the intermediate latitudes (*e.g.*, in the Alps, Alaska, or Iceland) where glacial response to climate variations is rapid and without retardation, that is, on a year to year basis. Rapid response of the great ice sheets to short-lived climatic oscillations seems to be ruled out because of melt-retardation.

A complicating factor near the periphery of former ice sheets is the so-called marginal bulge. Reginald Aldworth Daly, American geologist, postulated that if the ice load pressed down the middle of the glaciated area, then the earth's crust in the marginal area tended to rise up slightly, producing a so-called marginal bulge. With deglaciation the marginal bulge should slowly collapse. A fulcrum should develop between postglacial uplift and peripheral subsidence. In North America that fulcrum seems to run across Illinois to central New Jersey and

Adapted from Magnusson, N., Lundquist, G., and Regnell, A., *Sveriges geologi* Svenska Bokforlaget, Stockholm, 1963

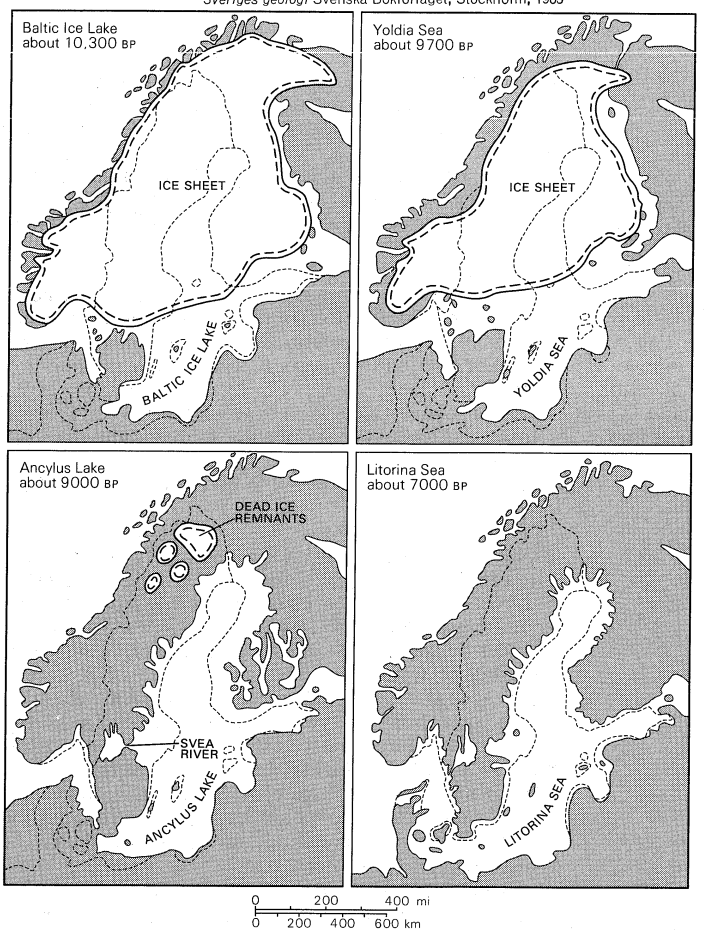


Figure 3: Evolution of the Baltic Sea during Holocene ice retreat.

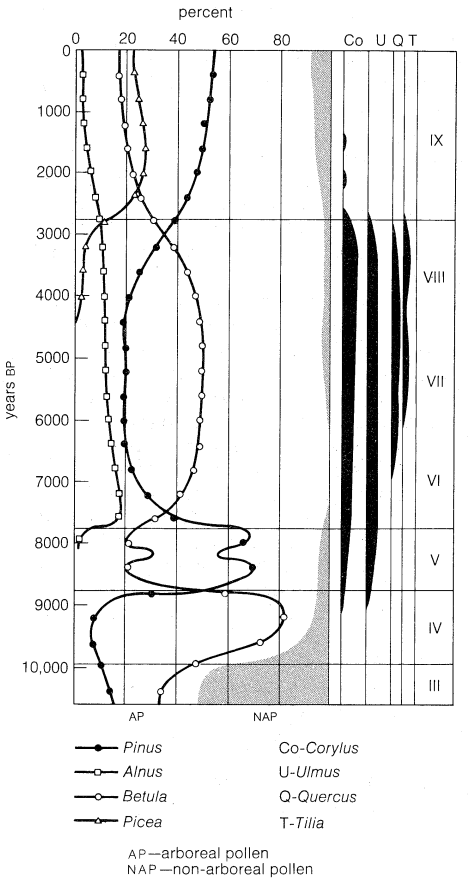


Figure 4: Holocene pollen stratigraphy for southern Finland indicating the sensitivity of vegetation to climatic change.

Adapted from J. Donner, in *The Quaternary* (1965); Interscience Publishers, John Wiley & Sons

swing northeastward, paralleling the coast and turning seaward north of Boston. In the Scandinavian region the fulcrum crosses central Denmark to swing around the Baltic Sea and then trends northeastward across the Gulf of Finland, north of Leningrad, so that the southeastern Baltic and northwestern Germany are subsiding. The Netherlands area is subsiding also, but here the pattern is complicated by the long-term negative tectonic trend of the North Sea Basin and the Rhine Delta.

It seems likely that this fulcrum shifted inward toward the former glacial centre during the early part of the Holocene. Passing inland, the lines of equal uplift (isobases) are positive, whereas seaward they are negative. The coastal area of southern New England is still slowly subsiding at the present time (1–3 millimetres per year). The eastern United States and Gulf Coast areas are complicated somewhat because they both are long-lived geosynclinal features (areas of major subsidence). Tide gage data show that these coasts were submerging faster than the world sea level rise of the first half of the 20th century would justify in terms of a eustatic explanation.

The great deltas of the world, those of the Mississippi, Rhine, Rhône, Danube, Nile, Amazon, Niger, Tigris-Euphrates, Ganges, and Indus, all coincide with regions of tectonic subsidence. Because water-saturated sediment has a tendency to compact under further sediment loading, there is an additional built-in mechanism that adds to the subsidence in such areas (see also RIVER DELTAS).

The deltaic environment and sedimentary facies

In this deltaic setting Holocene sequences are found that are quite different from those in the postglacial uplifted regions. Whereas the Holocene beaches in the uplift areas extend horizontally across the country in concentric belts, the Holocene sequence in the deltaic regions is, for the most part, vertical in nature and can be studied only from well data.

In both the Mississippi and Rhine deltas sediments that represent the earliest marine Holocene are missing. The sediments must lie seaward on the shelf margin, and the oldest marine layers are found to rest directly upon the late Pleistocene river silts and gravels. In a delta settling at around 0.5 to 3 millimetres per year, the rising sea of the Flandrian Transgression extended quickly across the river deposits to the inner margin (where there is a fulcrum comparable to that of the glaciated regions), marking the boundary between areas of downwarp and those of relative stability or gentle upwarp. The marine beds alternate with continental deposits that represent river or swamp environments. Six major fluctuations are recognizable in both the Mississippi and Rhine deltas. By radiocarbon dating the transgressive and regressive phases (Table 3) have been shown to be correlative in time.

On a subsiding coast there tends to be an alternation in importance between two types of associated sedimentary facies. During a regression of the sea the river distributaries are rejuvenated and there is an increase in the supply of sand and silt; beaches are widened and beach ridge dunes or cheniers may be formed. During a transgressive stage the saltwater wedge at river mouths causes a backup, and the estuary becomes much more sluggish (thalassostatic).

In the case of the Rhine Delta, the coast is bounded by extensive dunes, in contrast to the cheniers of Louisiana, and peat is much thicker than in the Mississippi Delta.

Beaches, dunes, and peat deposits

In The Netherlands the basal Holocene is buried in the fluvial deposits of the Lower Rhine. The postglacial eustatic rise had to traverse the North Sea Plain and advance up the English Channel several hundred kilometres before it reached the Netherlands area. At about 9000–8500 BP (Ancylus Stage in the Baltic) the coastal beaches still lay seaward from the present shore. Subsequently they became stabilized by a brief eustatic regression, while the high water table permitted the growth of the Lower Peat. This is contemporaneous with the late Boreal Peat that is widespread in northern Europe, as well as Peat #5 of the Mississippi Delta (Table 3).

A further eustatic rise (of about 10–12 metres) ensued around 7750 BP, corresponding to a warming of the climate marked by the growth of oak forests in western Europe (the BAT, or "Boreal–Atlantic Transition"). In The Netherlands the barrier beaches reformed close to the present coastline, and widespread tidal flats developed to the interior. These are known as the Calais Beds (or Calaisian) from the definition in Flanders by Dubois.

In the protected inner margins, the peat continued to accumulate during and after the "Atlantic" time. Because the peat has compacted much more than the muds and silts, it has contributed to the continued maintenance of low elevations in The Netherlands today with its lakes and swamps or man-made polders behind artificial dikes.

From evidence outside the areas of subsidence, it seems likely that the worldwide eustatic sea level rise reached its maximum sometime between 5500 and 2500 BP (many workers consider the date to be about 2000 BP). In The Netherlands, in spite of subsidence, the western coastline became more or less stabilized around 4000 BP with the beginning of the formation of the Older Dunes alternating with interdune soils. At the same time in the tide flat

Table 3: Peat Building, Sea-Level Fluctuation, and the Holocene Record of the Mississippi Delta

Mississippi stage	years BP	international correlation (regression)	probable eustatic range
Balize Delta	300	Late Medieval	0 to −0.5 m
Plaquemines Delta	700	Paria	+0.5 to −1.5 m
Lafourche Delta	1500	pre-Dunkerquian III	0 to −1 m
Peat #1 (St. Bernard or La Loutre Delta)	1700–2100	Florida (Roman)	+1.5 to −2 m
Teche Delta	2800–3300	Pelham Bay	+2 to −2 m
Peat #2 (Cocodrie Delta)	4000–4300	Bahama	? +3 to −3 m
Peat #3 (Marinquin, Sale, or Cypremort Delta)	4700–5000	early Subboreal	? +3 to −3 m
Peat #4	6000–6500	Rhine Delta	−4 to −9 m
Peat #5	7000–7500	late Boreal	−12 to −22 m

Source: Partly from Kolb and Van Lopik, 1966, and Fairbridge, 1968.

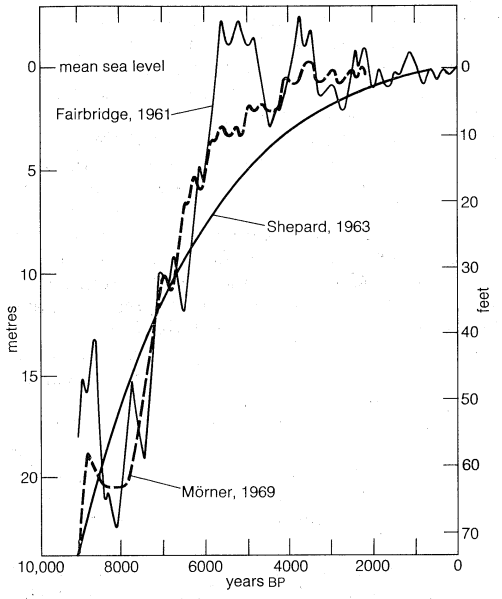


Figure 5: *Holocene sea level curves.*
The smooth curve of Shepard represents an average of many radiocarbon-dated shoreline indicators from both stable and subsiding areas; the oscillating curve of Mörner is based upon the isostatically emerged coast of western Sweden, corrected for uplift; and the oscillating curve of Fairbridge is based upon the geomorphology and radiocarbon dates of coastal features of the world, adjusted for crustal movements.
Adapted from N. Morner, *Sveriges Geologiska Undersokning*, Series C, no. 640 (1969)

areas the Calaisian was followed by the Dunkirk Stage or Dunkerquian. There are three important stages of accelerated tide flat sedimentation (transgressions) corresponding to three stages of Older Dune soils, around 3700 BP, 2500 BP, and 1700 BP (see Figure 2).

The Younger Dune sequence of The Netherlands began with a dry climatic phase in the 12th century AD. With several fluctuations of cold continental climates, dune building continued until the 16th century. Only brief positive oscillations of sea level occurred until the 17th century, when the "modern" warming and eustatic rise started, accompanied also by dune stabilization.

Broadly comparable patterns occur in other areas, from France and Britain to Texas, Oregon, and Brazil. There is normally a threefold or fourfold subdivision in all the Holocene coastal dune belts, each extensively vegetated and consolidated before the successively younger dune belt was added. In a number of cases there is evidence from buried beach deposits that the foundations of the inner dunes are older strandlines that were instituted when the sea was somewhat higher than today. An important regressive phase seems to have initiated each new dune belt.

Other coastal regions. Besides regions of glacioisostatic crustal adjustment, both positive and negative, and the deltaic or geosynclinally subsiding areas, there are many tens of thousands of kilometres of coastlines that are relatively stable and a smaller fraction that are tectonically active.

Holocene terraces and the eustatic curve

Most striking scenically are the coasts with Holocene terraces undergoing tectonic uplift. Terraces of this sort, backed in successive steps by Pleistocene terraces, are well developed in South America, the East Indies, New Guinea, and Japan. By careful surveys every few years the Japanese geodesists have been able to establish mean rates of crustal uplift (or subsidence) for many parts of the country and have been able to construct a residual eustatic curve that is comparable with those obtained elsewhere.

Besides uplifted coasts outside of glaciated areas there are also certain highly indented coasts that show clear evidence of Holocene "drowning." These coasts typically are characterized by the rias, or drowned estuaries, sculptured by fluvial action, but many of the valleys were cut 10,000,000 to 20,000,000 years ago, and the Holocene history has been purely one of eustatic rise.

On the basis of the known climatic history of the Holocene, from the strandline record of Scandinavia and from the sedimentologic evolution of the Mississippi and Rhine deltas, an approximate chronology of Holocene eustasy can be worked out. The amplitudes of the fluctuations and the finite curve are less easily established. A first approximation of the oscillations was published in 1959 and in a more detailed way in 1961 (the so-called Fairbridge curve). Smoothed versions have been offered by several other workers.

Early Holocene Preboreal and Boreal stages (10,000–7750 BP)

Continental regions. In the formerly glaciated regions the Holocene has been a time for the reinstitution of ordinary processes of subaerial erosion and progressive reoccupation by a flora and fauna. The latter expanded rapidly into what was an ecologic vacuum, although with a very restricted range of organisms, because the climates were initially cold and the soil was still immature.

Wind erosion in the deglaciated alluvial valleys and on piedmont fans often was quite severe in the early Holocene because of the limited vegetation. In the driest and coldest areas, loess (a periglacial dust, characteristic of the very cold and arid phases of glaciation) was deposited, but mostly the eolian deposits were limited to sand dunes, downwind of the meltwater river valleys. In the high Arctic, such as in parts of Alaska, loess accumulation has continued until the present.

In the midlatitudes and tropics the end of the last glacial period was marked by a tremendous increase in rainfall, which some believe corresponds to a peak in solar radiation that occurred about 10,000 BP at 65° N (which is critical because of the high land to water area ratio in this belt). The increased radiation would have led to increased westerly and trade wind velocities, greater zonal circulation, increased evaporation from the sea surface, and thus increased rainfall.

The increased precipitation toward the end of the Pleistocene was marked by a vast proliferation of pluvial lakes (those extant and expanded during Pleistocene time) in the Rocky Mountain region, notably Lake Bonneville and Lake Lahontan (giant ancestors of the present Great Salt Lake and Pyramid Lake). Two peaks of lake levels were reached at about 12,000 ± 500 BP (the beginning of the Allerød Warm Stage) and about 9000 ± 500 BP (the early Boreal Warm Stage). At Lake Balaton (in Hungary) high terrace levels also mark the Allerød and early Boreal Warm stages. Lake Victoria (in East Africa) shows the identical twin oscillation in its terrace levels (see further CLIMATIC CHANGE).

In equatorial regions the same evidence of high solar radiation and high rainfall at the end of the Pleistocene and during the early Holocene is apparent in the record of the Nile sediments. The Nile, like the other great rivers of Africa (notably, the Congo, Niger, and Senegal), became very reduced, if not totally blocked, by silt and desert sand during the low-precipitation, arid phases of the Pleistocene. An erroneous correlation between glacial phases and pluvial phases in the tropics has been widely accepted in the past, although cold ocean water means less precipitation, not more. The pluvial phases correspond to the high solar radiation states, the last maximum being about 10,000 BP. Thus tremendous increases of Nile discharge are determined, by radiocarbon dating, to have occurred around 12,000 and 9000 BP, separated and followed by alluviation, indicating reduced runoff in the headwaters. These impressive fluctuations coincide with overflows into the White Nile from Lake Victoria, corresponding to the high lake levels noted above. The lake's outlet was closed during the dry stages prior to 12,000 BP and again, briefly, around 10,000 BP. Erosion of the bar at Ripon Falls has kept the Nile running since then, but drying trends in the headwaters have greatly decreased its discharge.

The expansion of monsoonal rains during early Holocene in the tropical latitudes permitted an extensive spread of moist savanna-type vegetation over the Sahara in North Africa and the Kalahari in South Africa and in

broad areas of Brazil, India, and Australia. Most of these areas had been dry savanna or arid during the last glacial period. Signs of late Paleolithic and Late Stone Age men can be seen throughout the Sahara today, and art is representative of the life and hunting scenes of the time. Lake deposits have been dated as young as 5000–6000 BP. Lake Chad covered a vast area in the very late Pleistocene and up to 5000 BP. The Dead Sea throughout the early Holocene shows a record of sedimentation from humid headwaters; there was a Neolithic settlement at Jericho around 9000–10,000 BP.

Atlantic Stage (7750–5100 BP)

In the high to midlatitudes after the early Holocene, with its remnants of ice-age conditions (tundra passing to birch forests), there was a transition to the mid-Holocene, marked by a progressive change to pine forest and then oak, beech, or mixed forest. The mean annual temperature reached 2.5° C above that of today. Neolithic man pressed forward across Europe and Asia. In North America the paleo-Indians reached into the region of the Great Lakes and the St. Lawrence Valley. In the Canadian Arctic and in Manitoba the mean temperature passed 4° C above present averages. It was a "milk-and-honey" period for primitive men over much of the world, and in Europe it paved the way for the cultured races of the Bronze Age. Navigators started using the seaways to trade between the eastern Mediterranean, the British Isles, and the Baltic.

In the Atlantic Stage in the midlatitudes there was widespread development of chocolate-brown soils and *terra rossa*. The vegetational cover became almost worldwide, except in the high mountains. As a result, the fluvial regimes passed from braided streams to meandering streams adjusted to a high degree of equilibrium.

Subboreal Stage (5100–2200 BP)

In the midlatitude continental interiors there was still evidence of hot summers, but the winters were becoming colder and partly drier. There was an expansion of steppe or prairie conditions and their associated fauna and flora. Many lake levels showed a fall. The Dead Sea first showed a rise, followed by desiccation and deposition of evaporites (*q.v.*) at about 4300 BP. In the later part of the stage the first signs of the late Holocene climatic deterioration can be detected.

In Europe there was also the beginning of widespread deforestation as Bronze Age man started to use charcoal for smelting and extended agriculture to tilling and planting. In consequence, soil erosion began almost immediately, hillsides developed lynchets (terracettes), and "anthropogenic sediments" began to accumulate on the lower floodplains. In western Europe, at about 3500 BP, there was a crescendo of megalithic cultures, with giant statues, menhirs, burial mounds, and the mathematical ingenuity of Stonehenge.

At this time, in the eastern Mediterranean, the Minoan culture (of Crete) rose, and then catastrophically declined, some believe because of climate change, others because of the ash showers and tidal waves from the Santorin eruption (*c.* 3450 BP). It was replaced by the Mycenaean civilization (3400–3200 BP). It has been suggested that drought was the basic reason for the Cretan decline in the period 3200–2700 BP. Northern Europe at that time was cool and dry; the Alps were also in a dry phase and world sea level was low. High rainfall did not return to the Mediterranean until the mild period that was marked by the historical date 776 BC (2736 BP), which marked the First Olympiad in Greece.

At the corresponding latitudes in the Southern Hemisphere (approximately 30°–35° S), pollen analysis indicates increasing desiccation during the Subboreal Stage, with a maximum dryness about 3200 BP.

In the subtropical regions of Mesopotamia and the Nile Valley, people had learned to harness water. The stationary settlements, advanced agriculture, and mild climates favoured a great flowering of human culture. It is surmised that when the normal floods began to fail, man's ingenuity rose to the occasion, and he devised irrigation canals and machinery. In the reign of the pharaoh Ramses II (about 1230 BC, or 3180 BP) a freshwater canal was dug from Cairo to Suez, initially for irrigation, but for

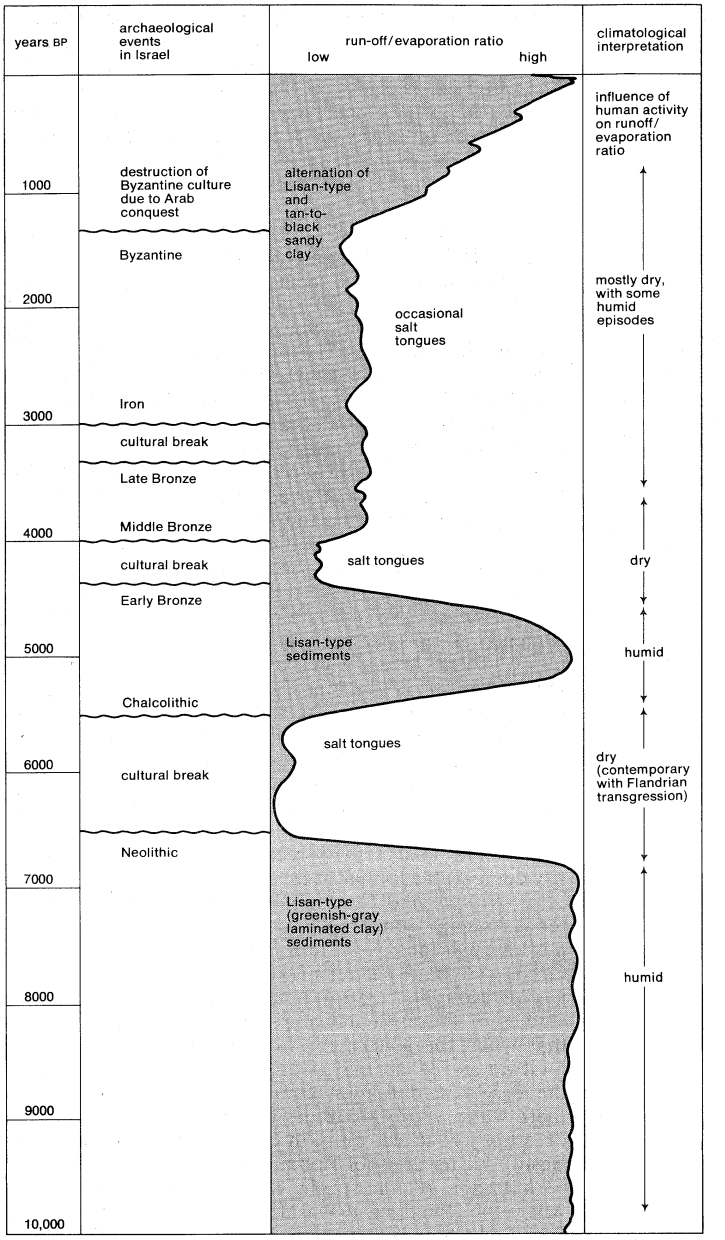


Figure 6: Holocene history of Dead Sea and Jordan Valley.
From D. Neev and K. Emery, *Israel Geological Survey*, bulletin 41 (1967)

the first time small ships were able to go from the eastern Mediterranean (and thus from Britain and the Baltic) to the Red Sea, East Africa, India, and China.

Sub-Atlantic Stage (2200–0 BP)

This is the last major physical division of the geological record. Historically its beginning coincides with the rise of the Roman Empire in Europe, the flowering of the classical dynasties of China, the Ptolemies in Egypt, the Olmecs of central Mexico and Guatemala, and the pre-Incan Chavin cultures of Peru.

It is important to recognize that as the present day is approached, the minor cycles and modulations become relatively more and more prominent; indeed, it is all too easy to confuse minor fluctuations with the long-term trends. Numerous additional sources of information have become available for the study of these last two millennia. Most fundamental probably is the record of solar activity, as disclosed by documentation of aurora in ancient Chinese court records and later by sunspot numbers. Both phenomena reflect solar activity in general, but correlation with weather records in the higher latitudes is complicated. Other indicators of climate, such as tree-ring analysis and palynology, were previously mentioned, but many documentary indications are also useful: the time of the cherry blossom festival in Kyoto,

Japan, the freezing of lakes, the incidence of floods, blizzards, or droughts, the economics of harvests, salt evaporation production, some disease statistics, and so on. The water levels of closed basins such as the Caspian Sea and particularly the evaporite basin of the Kara Bogaz Gulf reflect runoff to the Volga. The Dead Sea reflects eastern Mediterranean precipitation.

The main trends of the Sub-Atlantic are identifiable as follows:

Classical Roman Period. This time interval is marked by the Florida or Roman emergence in the eustatic record about the BC–AD boundary and succeeded by a transgression.

The solar record is not complete, but indications are for low activity. Records of rainfall kept by the astronomer Ptolemy (AD 127–151) in Alexandria noted thunderstorm activity in every summer month in comparison with the totally dry summer today, which suggests a slightly wetter overall pattern in this latitude.

In northern Europe and in other high latitudes, in contrast, the cool stage at the beginning of the 1st century AD may have been drier and more continental, as shown by dune building. It was immediately followed in the 2nd century by much stronger westerly activity, as indicated by a 6-metre rise in the Kara Bogaz Gulf.

Late Roman Period. After the 1st century AD there is evidence of a progressive rise in sea level. Roman buildings and peat layers were covered by the marine transgression in The Netherlands, southern England, and parts of the Mediterranean. At the same time, drying and warming trends were associated with alluviation of streams and general desiccation in southern Europe and North Africa. Similar alluviation occurred in the American Southwest. This warming and desiccation trend is evident also in the subtropics of the Southern Hemisphere. The solar activity record indicates a mean intensity comparable to that of the mid-20th century.

Post-Roman and Carolingian Period. This period extends roughly from AD 400 to 1000 (Dunkirk II). The important invasions of western Europe by the Huns and the Goths may have been generated by deteriorating climatic conditions in Central Asia. Radiocarbon dating and studies of the ancient Chinese literature have disclosed that when the glaciers of Central Asia were large, the meltwaters fed springs, rivers, and lakes on the edge of the desert, and human communities flourished. When there was a warm phase the water supply failed and the deserts encroached. Thus, in Central Asia (and the Tarim Basin) during the cool Roman Period, the Old Silk Road permitted a regular trade between Rome and China, where the Han dynasty was flourishing. During the Ch'in, Wei, and Chou dynasties this trade declined. In the T'ang dynasty (AD 618–907) there was a reopening of the trade routes, and likewise during the Yüan dynasty (AD 1279–1368). Marco Polo passed this way in AD 1271. Radiocarbon dates of the 8.6-metre-high lake level at Sogo Nuur showed overflow conditions from AD 1300–1450, after which gradual, fluctuating, but progressive desiccation followed, and today the area is almost total desert.

In North America the Post-Roman–Carolingian Period was marked by warm temperatures in the northern parts, with mean paleotemperatures in central Canada about 1° C above the present. In the semi-arid Southwest of the United States, the arroyos, washes, and ephemeral river valleys were filling slowly with alluvium (younger "Tsegi alluvium"), an indication that stream energy was generated by the summer flash floods. The Sierra Nevada experienced a warm phase, and there were marginal retreats in almost all the mountain glacier regions of the world from the Alps to Patagonia.

The radiation fluctuations of the 7th to 9th centuries is reflected by flood levels on the Nile. On a ruined temple built by Ramses II north of Wādī Halfā (now drowned by Lake Nasser) the early Christians had built a superstructure that was destroyed by floods around AD 750–800; a second church suffered the same fate around AD 940–980 when there was another high solar activity peak.

In the tropical region of Central America there was the unexplained decline of the coastal Mayan people (Mexico and Guatemala) about the 10th century AD. The mountain Mayas continued to flourish, however, and it is possible that the high precipitation of this warming period introduced critical ecological limits to continued occupation of the (now) swampy coastal jungles.

Decline of the Mayas in 10th century

The Viking–Norman Period. Approximately AD 1000–1250 the worldwide warm-up that culminated in the 10th century and has been called the early Medieval Warm Period or the "Little Climatic Optimum," continued for two more centuries, although there was a brief drop in mean solar activity in the period around 1030–70. In the 8th to 10th centuries the Vikings had extended as far afield as the Crimea and had established coastal salt pans, the existence of which speak for seasonally high evaporation conditions and eustatic stability.

The levels of the Caspian Sea and Kara Bogaz Gulf were lowered at this time, reflecting high temperatures and low rainfall there and in the Volga Basin. Corresponding to the increased solar activity there were heavier monsoonal rains in the equatorial regions, and lake levels were high in East Africa.

In the Arctic regions during the 10th, 11th, and 12th centuries there was widespread navigation by the Vikings. Partly in response to reduced sea-ice conditions and milder climates they were able to establish settlements in Iceland, southern Greenland (Eric the Red, *c.* 985), and in eastern North America (Vinland; Leif Eriksson, *c.* 1000). In Alaska, from tree-ring evidence, the mean temperature was 2–3° C warmer in the 11th century than today. Eskimos had settled in Ellesmere Island about AD 900. Records of sea-ice off Iceland show negligible severity from 865–1200. Often the westerly storm tracks must have passed north of Europe altogether.

After a brief interval of cold winters in Japan, the cherry blossoms returned to early blooming in the 12th century. In the semi-arid Southwest of the United States there appears to have been increased precipitation, leading to a spread of vegetation and agriculture. Pueblo campsites dated AD 1100–1200 are found on top of the youngest Tsegi Alluvium. The snowline in the Rocky Mountains was about 300 metres (1,000 feet) higher than today.

Similar trends are recorded in the Southern Hemisphere, notably in Australia and Chile. The first immigration of Maori peoples into New Zealand probably occurred at this mild time.

Early Medieval Cool Period (AD 1250–1500). This interval corresponds to the Paria Emergence in the eustatic record and has been called one of the "little ice ages" by certain authors. Solar activity records show a decline from 1250 to 1350, a brief rise from 1350 to 1380, and then a phenomenal low that lasted until 1500. Pollen records in northern Europe reveal rather consistently cool conditions, and smoothed mean temperature curves show a cumulative drop during this period. Stalactite studies in a karst cave in France showed a travertine growth peak (indicating cool, moist conditions) in 1450. In North America cool, moist conditions were widespread at first, becoming dry later. The arroyos and washes became filled with the Naha Alluvium, and the human population decreased markedly. There is pollen evidence of a temperature drop of about 1° C. This is the period of the "Great Drought." In the upper Mississippi Valley the Indian cultures began a general decline, accompanied by a transfer from agriculture to hunting. It was similar in the western prairies, and it was this hunting culture that the first Spanish explorers encountered.

In the Canadian north the mean temperatures had dropped about two degrees below the previous high. In the Sierra Nevada, the Rockies, and in Alaska there were glacial readvances, with evidence of a 2° C temperature drop. In the Arctic regions, the Eskimo economy underwent a marked change to adjust to these more extreme conditions, which amounted to about 5 or 6° C below the mean of the climatic optimum.

The Norse settlements in Greenland were abandoned altogether as the permafrost advanced. Pollen studies at

Godthåb indicate a shift from a maritime climate to a cold, dry continental regime. The sea ice off Iceland reveals an extraordinary growth in severity, from zero coverage before the year 1200 to eight-week average cover in the 13th century, rising to 40 weeks in the 19th century, and dropping again to eight weeks in the 20th century. In Japan there were glacial readvances and a mean winter temperature drop of 3.5° C. Summers were marked by excessive rains and bad harvests.

The equatorial regions now began a marked desiccation, with a drop in level of all of the great African lakes. The Nile suffered a decreased flow and alluviation, although brief fluctuations also are recorded. (A detailed documentation is available because taxes in Egypt were based on the effectiveness of the Nile flood, and careful readings were taken.) Many early Christian settlements along the middle Nile were abandoned at this time.

South of the Equator in the temperate belts there was a general return to cooler and wetter conditions that have continued (with oscillations) until the present time in southern Chile, Patagonia, southernmost Africa, southwest Australia, and New Zealand.

Medieval Little Ice Age (AD 1500–1850). Throughout most of this period the mean solar activity was quite low, but positive fluctuations occurred around 1540–90 and 1770–1800. The main westerly storm belts shifted about 500 kilometres to the south, and for much of the time the northern latitudes came under cool continental conditions. Observed temperature series in Europe from Paris to Leningrad show large fluctuations until 1850.

Glacier advances are recorded in the Alps, in the Sierra Nevada, and in Alaska. Corresponding low sea levels are recorded by early tide gauge records in The Netherlands and Germany. Even in equatorial latitudes there are traces of mountain glacier advances (as in the Andes of Colombia).

The Industrial Age (AD 1850–1950). The year 1850 started a brief warming trend that persisted for 100 years. It also approximates a critical turning point in climatic, sea level, glacial, and sedimentologic records. In many regions of central and southern Europe "anthropogenic" sediments (or cultural layers) started to appear in Neolithic times (early to mid-Holocene). But elsewhere in the world (*e.g.*, in North America, Australia, South Africa) this type of sedimentation began around the middle of the 19th century, depending upon soil erosion stimulated by mechanized (disk) plowing, large-scale deforestation, and engineering activity. Thus, independently of natural climatic change, the century 1850–1950 is marked outstandingly by anthropogenic aridification, the time of man-made deserts.

The earth now is on a long-term cooling trend of the glacial-interglacial cycles and is likely to continue so for several thousand years, but there are numerous modulating influences, meteorological, geological, and man-made. These may accelerate or reverse the general trend, and from the point of view of human history they can well play a critical role.

BIBLIOGRAPHY. There is a vast array of specialized research papers that treat the many aspects of the Holocene Epoch. This results from the fact that evidence of Holocene environments is based on data so diverse as Paleolithic stone axes and sunspot magnitude and frequency. The bibliography presented here is restricted to readily available books or parts of books that cover the Holocene record in general and will serve as a guide to the more specialized literature.

W.W. BISHOP and J.D. CLARK (eds.), *Background to Evolution in Africa* (1967); C.E.P. BROOKS, *Climate Through the Ages*, 2nd rev. ed. (1949, reprinted 1970); J.K. CHARLESWORTH, *The Quaternary Era*, 2 vol. (1957); R. CLAIBORNE, *Climate, Man, and History* (1970); E.J. CUSHING and H.E. WRIGHT, JR. (eds.), *Quaternary Paleoecology* (1967); R.W. FAIRBRIDGE, "Eustatic Changes in Sea Level," in L.H. AHRENS *et al.* (eds.), *Physics and Chemistry of the Earth*, vol. 4, pp. 99–185 (1961), and "African Ice-Age Aridity," in A.E.M. NAIRN (ed.), *Problems in Palaeoclimatology*, pp. 356–363 (1964); J.L. HOUGH, *Geology of the Great Lakes* (1958); H.H LAMB, *The Changing Climate* (1966); P.S. MARTIN and H.E. WRIGHT, JR. (eds.), *Pleistocene Extinctions* (1967); J.M. MITCHELL, JR., "Theoretical Paleoclimatology," in H.E. WRIGHT, JR., and D.G. FREY, (eds.), *The Quaternary of the United States*, pp. 881–901 (1965); R.B. MORRISON and H.E. WRIGHT, JR. (eds.), *Means of Correlation of Quaternary Successions* (1968); K. RANKAMA (ed.), *The Quaternary*, 2 vol. (1965–67); H.G. RICHARDS and R.W. FAIRBRIDGE, *Annotated Bibliography of Quaternary Shorelines 1945–1964* (1965); J.S. SAWYER (ed.), *World Climate from 8000 to 0 B.C.* (1966); M. SEARS (ed.), *The Quaternary History of the Ocean Basins*, vol. 4, *Progress in Oceanography* (1967); C. VITAFINZI, *The Mediterranean Valleys* (1969).

(R.W.F.)

Holography

Holography is, in principle, a means of creating a unique photographic image without the use of a lens. The photographic recording of the image is called a hologram, which appears to be an unrecognizable pattern of stripes and whorls but which—when illuminated by coherent light, as by a laser beam—organizes the light into a three-dimensional representation of the original object.

An ordinary photographic image records the variations in intensity of light reflected from an object, producing dark areas where less light is reflected and light areas where more light is reflected. Holography, however, records not only the intensity of the light but also its phase, or the degree to which the wave fronts making up the reflected light are in step with each other, or coherent. Ordinary light is incoherent; that is, the phase relationships between the multitude of waves in a beam are completely random; wave fronts of ordinary light waves are not in step.

Dennis Gabor, Hungarian-born scientist, invented holography in 1948, and received the Nobel Prize for Physics in 1971. Gabor considered the possibility of improving the resolving power of the electron microscope, first by utilizing the electron beam to make a hologram of the object and then by examining this hologram with a beam of coherent light. A degree of coherence can be obtained by focussing light through a very small pinhole, but this technique reduces the light intensity too much for it to serve in holography; therefore, Gabor's proposal was for several years of only theoretical interest. The development of lasers in the early 1960s (see LASER AND MASER) suddenly changed the situation. A laser beam not only has a high degree of coherence but high intensity as well.

Of the many kinds of laser beam, two have especial interest in holography, the continuous-wave (CW) laser and the pulsed laser. The CW laser emits a bright continuous beam of a single, nearly pure colour. The pulsed laser emits an extremely intense, short flash of light that lasts only about $\frac{1}{100,000,000}$ of a second. Two scientists in the United States, Emmett N. Leith and Juris Upatnieks of the University of Michigan, applied the CW laser to holography and achieved great success, opening the way to many research applications.

PRINCIPLES AND BASIC TYPES OF HOLOGRAPHY

Gabor coined the name holography from the Greek *holos*, or "whole," and *gram*, "message," because the image-forming mechanism he conceived recorded all the optical information in a wave front of light. Basically the problem Gabor conceived in his attempt to improve the electron microscope was the same as the one photographers have confronted in their search for three-dimensional realism in photography. To achieve it, the light streaming from the source must itself be photographed. If the waves of this light, with their multitude of rapidly moving crests and troughs, can be frozen for an instant and photographed, the wave pattern can then be reconstructed and will exhibit the same three-dimensional character as the object from which the light is reflected. Holography accomplishes such a reconstruction by recording the phase content as well as the amplitude content of the reflected light waves of a laser beam. How this works is shown in Figure 1 at the top.

Gabor's search for three-dimensional realism

Continuous-wave laser holography. *Producing a hologram.* In a darkened room, a beam of coherent laser light is directed onto object O from source B. The beam is reflected, scattered, and diffracted by the physical features

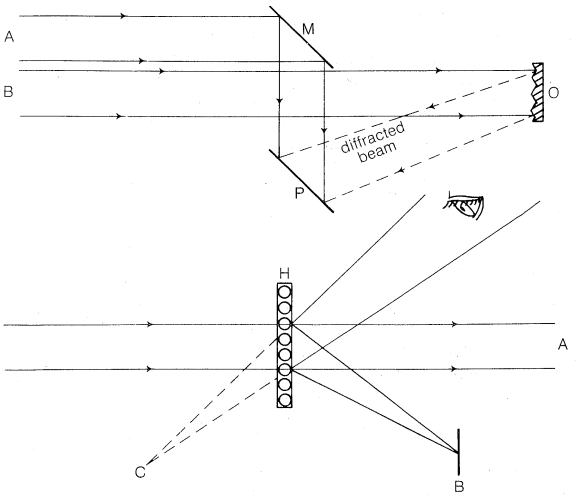


Figure 1: Arrangements for (top) creating a hologram and (bottom) reconstructing an image from a hologram (see text).

of the object, and arrives on a photographic plate at P. Simultaneously, part of the laser beam is split off as an incident, or reference, beam A and is reflected by mirror M also onto plate P. The two beams interfere with each other; that is, their respective amplitudes of waves combine, creating on the photographic plate a complex pattern of stripes and whorls called interference fringes. These fringes consist of alternate light and dark areas. The light areas result when the two beams striking the plate are in step—when crest meets crest and trough meets trough in the waves from the two beams; the beams are then in phase, and so reinforce each other. When the two waves are of equal amplitude but opposite phase—trough meeting crest and crest meeting trough—they cancel each other and a dark area results (see also LIGHT).

The plate, when developed, is called a hologram. The image on the plate bears no resemblance to the object photographed but contains a record of all the phase and amplitude information present in the beam reflected from the object. The two parts of the laser beam—the direct and the reflected beams—meet on the plate at a wide angle and are recorded as very fine and close-packed interference fringes on the hologram. This pattern of fringes contains all the optical information of the object being photographed.

Reconstructing the image. By reversing the procedure, as shown at the bottom in Figure 1, an image of the original object can be reconstructed. The coherent light of a laser beam illuminates the hologram negative H. Most of the light from the laser passes through the film as a central beam A and is not used. The close-packed, fine-detailed fringes on the hologram negative act as a diffraction grating, bending or diffracting the remaining light to exactly reverse the original condition of the coherent light waves that created the hologram. The diffracted light is transmitted at a wide angle from that of the laser's reference beam.

Diffraction grating in hologram negatives

On the light source side of the hologram, at C, a virtual image visible to the eye is formed. On the other side, at B, a real image that can be photographed is formed. Both these reconstituted images have a three-dimensional character because in addition to amplitude information, which is all that an ordinary photographic process stores, phase information also has been stored. This phase information is what provides the three-dimensional characteristics of the image, as it contains within it exact information on the depths and heights of the various contours of the object. It is possible to photograph the reconstituted image, at B, by ordinary photographic means, at a selected depth, in exact focus.

The difference in appearance between a hologram and the reconstituted image is remarkable, as illustrated in Figure 2. The object in this instance was a model of a royal crown, decorated with fur trim. The hologram, an enlarged part of which is shown at the left, is a complex pattern of fringes. On the right is the astonishingly crisp, reconstructed image obtained from this hologram.

The real image from a hologram—that is, the one that can be photographed—appears pseudoscopic, or with a reversed curvature. This reversal can be eliminated by making a double hologram, first by preparing the single hologram and then by using it as an object in the creation of a second hologram. With a double reversal the image becomes normal again, as when a mirror image of writing is made legible by viewing it in a second mirror. The real image of a hologram has valuable properties. A viewing camera or microscope can be positioned and focussed on various selected positions in depth. The original object also can be brought into the position in space.

The hologram not only offers images at different depths (different cross-sections of the object) but also images seen along different directions if the viewer moves off the axis on which the principal image is viewed. Direct images can be seen under these conditions. In holography it is also possible to record on the same plate a succession of numerous multiple images that can be reconstructed as one image, leading to the possibility of holography in colour. Three holograms could be superimposed on the same plate, using three lasers of different colours. Reconstruction with the three different lasers would produce an image in its natural colour, even though the hologram plate itself is black-and-white.

Colour holography from three black-and-white holograms

Pulsed-laser holography. A moving object can be made to appear to be at rest when a hologram is produced with the extremely rapid and high-intensity flash of a pulsed ruby laser. The duration of such a pulse can be less than $\frac{1}{10,000,000}$ part of a second; and, as long as the object does not move more than $\frac{1}{10}$ of a wavelength of light during this short time interval, a usable

By courtesy of the National Physical Laboratory, Teddington, Middlesex, Crown copyright reserved

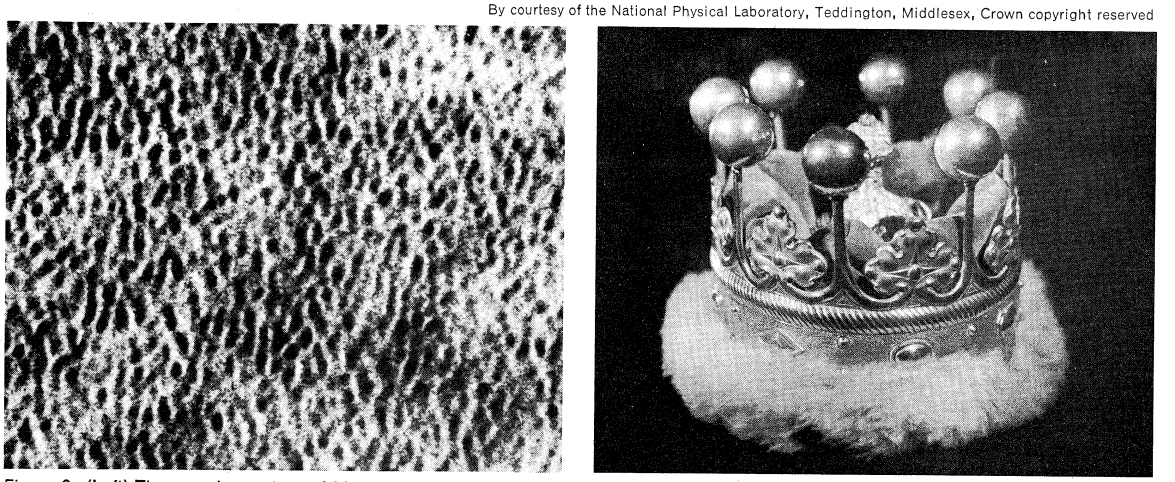


Figure 2: (Left) The complex pattern of fringes in a hologram. (Right) The reconstructed image from the hologram.

hologram can be obtained. A continuous-wave laser produces a much less intense beam, requiring long exposures; thus it is not suitable when even the slightest motion is present.

With the rapidly flashing light source provided by the pulsed laser, exceedingly fast-moving objects can be examined. Chemical reactions often change optical properties of solutions; by means of holography, such reactions can be studied. Holograms created with pulsed lasers have the same three-dimensional characteristics as those made with CW sources.

Pulsed-laser holography has been used in wind-tunnel experiments. Usually high-speed air flow around aerodynamic objects is studied with an optical interferometer (a device for detecting small changes in interference fringes, in this instance caused by variations in air density). Such an instrument is difficult to adjust and hard to keep stable. Furthermore, all of its optical components (mirrors, plates, and the like) in the optical path must be of high quality and sturdy enough to minimize distortion under high gas flow velocities. The holographic system, however, avoids the stringent requirements of optical interferometry. It records interferometrically refractive-index changes in the air flow created by pressure changes as the gas deflects around the aerodynamic object.

The arrangement is illustrated in Figure 3. Coherent

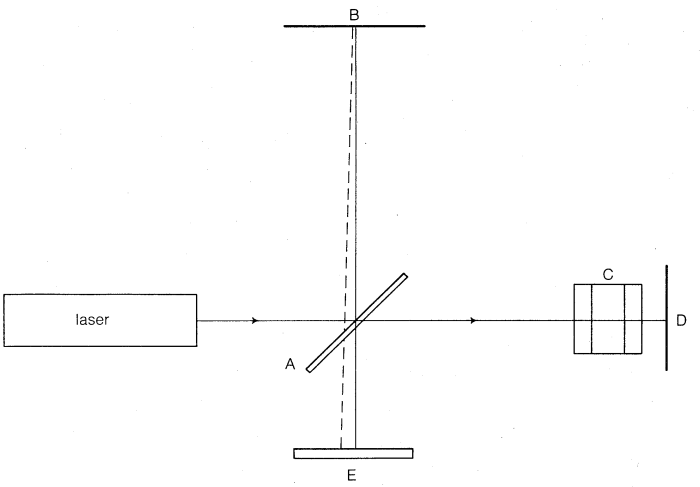


Figure 3: Arrangement for detecting changes in the refractive index of gas (see text).

light from a laser is divided by beam splitter A. One-half reflects to reference mirror B. The other half passes through gas chamber C, in which the refractive-index change takes place, and is then reflected from D to photographic plate E, on which the first holograph image is formed. A hologram is made with the gas in chamber C undisturbed. The refractive-index change is then induced in chamber C. Here, the change is induced in a flow chamber, but it also could be in the heat flow in a flame, the chemical change in a solution, or the alteration of a liquid with the passage of high-energy particles. From the induced refractive change a second hologram is doubly exposed, superimposed on the first. A later reconstruction shows a fringe pattern the same as that given by an optical interferometer.

At first, holography applied to standard interferometry may seem to be elaborately complex, but it is not. The instrument just described is in reality a type of Michelson interferometer. The process is easier to handle and less expensive than previous arrangements; it is referred to as the Michelson holograph system. The standard Michelson interferometer is not used in the classical optical method in wind-tunnel studies because of the awkward localization of fringes that appear on plane B and not on the object. It is not possible to photograph the object and fringes simultaneously in focus. The defect does not exist in the Michelson holograph system, because the fringes in holography can be seen in coincidence with the object producing the fringes.

Of equal or greater importance is the fact that in holography there is no need to be restricted to highly precise optical mirrors, windows, and lenses. With imperfect optical components, all of the optical imperfections appearing in one holographic image are faithfully repeated in the second image and therefore do not contribute to interference fringes when the two images are superimposed. The fringes seen are created only by the deliberately imposed refractive-index changes. Furthermore, in this system, since the second hologram is being matched against that from a reasonably plane surface, the fringe pattern is of the type familiar in optical interferometry and thus quite easy to evaluate.

Imperfections permitted in optical components

Nonphotographic holography. Holograph images are recorded also on materials other than photographic plates. These nonphotographic materials can be classed as either fixed or erasable.

Fixed holograms. Although some fixed holograms are similar to photographic plates, their primary purpose is to raise the resolving power, which is limited in photographic plates by emulsion grain. Holograms can be recorded by the use of a glass plate coated with gelatin containing dichromate. When illuminated, the gelatin hardens in those areas receiving light and after immersion in water produces differential swelling. An etched "photographic" record is created when the water is later removed from the gelatin. This type of hologram records extremely fine detail, of the order of several thousand lines per millimetre, far exceeding the performance of a good photographic emulsion. As yet, however, such holograms are difficult to prepare, as they are subject to defacement and distortion because of water absorption from the atmosphere.

In another approach, a photopolymer, a polymerized material sensitive to light and more stable than gelatin, is employed. A typical photopolymer is a metallic-acrylate monomer that contains a dye catalyst. The monomer can be polymerized; that is, transformed from a soft gel to a hard solid by using ultraviolet light. The process is not reversible and so produces a permanent record. With laser-scattered illumination, small, solid particles faithfully following the hologram patterning are formed on the hologram plate. These particles are fixed by the ultraviolet light as soon as the hologram is formed. Concurrently, the ultraviolet light bleaches out the catalyst dye. Since there is a different refractivity between gel and solid, a hologram in transparent material can be created.

Erasable holograms. Erasable holograms offer intriguing technical possibilities, and several types have been proposed. An erasable hologram could store information in computer systems. Furthermore, it could be used as a temporary hologram for reasons of economics. In photochromic (colour-sensitive) materials, the absorption of light varies with the wavelength of light that illuminates them. One wavelength can increase light absorption, while another can return the material to normal transmission (*i.e.*, it can remove the absorption formerly created). Thus a hologram can be created and viewed with one wavelength and wiped off with another because of the photoelectric effect in the molecule of the dye, which is called a photochrome. Phase-erasable holograms have been developed using thin films of special crystals, such as lithium niobate. When holographic light falls on the crystal, electrons are excited in the area stimulated by the light because of the crystal's photoelectric properties. The excited electrons cause a local change in refractivity, causing a phase shift in the incident light. The phase hologram can be used later to reconstitute an image. Ultraviolet-light illumination or heat can be applied to erase the image.

Phase-erasable holograms on thin crystal films

Another promising material consists of thermoplastic, photoconductor, and transparent conductor layers. The incident light pattern creates an electric charge pattern on the material. If the material is then heat softened, the thermoplastic deforms in areas that contain residual electric charges, producing a wrinkled surface and a hologram. When the charge disappears, a reheat results in an erasure.

High resolution is possible with thin, magnetizable

films. When illuminated, local changes in magnetization are created in the material because it exhibits a large magneto-optical effect; that is, its magnetic characteristics are altered by exposure to light. The changes in magnetization cause physical changes in the pattern of magnetic domains (regions in which all element dipoles are aligned in the same direction), creating the hologram. The hologram is erased by subjecting it to a suitable magnetic field.

EXPERIMENTAL DIFFICULTIES IN HOLOGRAPHY

Most of the nonphotographic devices are in an experimental stage, and the photographic production of holograms remains the only widely used process. Yet it continues to encounter practical problems. As a rule, theoretically predictable perfection in the appearance of the reconstituted image is not attainable for a variety of reasons. There is always some loss in the recorded information and subsequently in the final image. As the field of view is necessarily restricted to the dimensions of the photographic plate that records the hologram, the ability of the system to resolve fine detail is also limited. To reconstitute the image, an adequately large diffracting angle must be provided so that the images will appear far off the axis of the laser beam illuminating the hologram. To produce a wide angle, the hologram's fringe lines must be close together to provide a very fine line grating. Making the hologram behave in this manner depends both on the structural detail of the object and on the angle formed by the directions of the diffracted wave front from the object and the auxiliary interfering wave front deflected to mix with it. The need for such very fine fringe detail thus imposes a condition of high resolution on the photographic plate that records the hologram. No matter how good the plate, invariably some detail is lost.

Speckle patterns

Another frequent source of trouble is the speckle pattern. When coherent light strikes a nonspecular (nonmirror-like) body, its high coherence creates interference effects because its waves are diffused by the uneven surface of the body. As a consequence, a rough object appears covered with a peculiar random pattern of small, bright speckles, or light dots. Such speckling can seriously affect the resolution of detail in the reconstituted image.

Since a hologram is essentially a record of optical interference fringes, slight alterations in the optical distances can cause bad blur and reduce definition. Thus distance between the object and the holographic plate must be maintained uniform to within close limits during exposure. The holograph pattern will be distorted by any changes in the optical properties of the space between the object and the recording plate. There must be good temperature control and careful exclusion of drafts. Movements and optical length changes greater than $\frac{1}{10}$ of a light wave lead to blurring and a reduction in definition of the final reconstituted image. As will be shown later, this extreme sensitivity can be exploited holographically to detect small movements and small changes in the optical properties of materials.

Some small random imperfections in laser light occur; occasionally these imperfections can introduce defects into the holograph image. Furthermore, laser coherence may vary during exposure time; fine-grain, high-resolution plates are low in light sensitivity and thus require ten or more minutes of exposure. A wavelength drift can also occur. The greater the depths within the object (depths can exceed 50 centimetres), the more stringent are coherence conditions. For high-performance holography, high laser stability is required, increasing considerably the cost of the equipment.

APPLICATIONS OF HOLOGRAPHY

It is possible to obtain high-resolution holographic reconstructions that compare favourably with ordinary, good-quality microscope images, but such an application is hardly justified. Resolution tests with holograms indicate that lines 0.01 millimetre (0.0004 inch) apart can be resolved, but this does not remotely approach what can be accomplished with high-resolution optical microscopy. Speckle effects often reduce resolution. Some success has been achieved in averaging out and reducing speckle by moving the hologram plate during exposure, but this is an awkward procedure.

Use of cameras and microscopes in hologram studies

Many applications are made possible because the real image from the hologram can be viewed by a camera or microscope, making it possible to examine difficult and even inaccessible regions of the original object. A deep, narrow depression on a plane, for example, cannot often be reached by a microscope objective because of working distance limitations. If the detail can be reached by coherent light, however, a hologram can be taken and its image reconstructed. Since this image is aerial, the microscope can be positioned in such a way that it can focus on the required region. In the same way, a camera also can be focussed at the required depth and can photograph objects inside a deep transparent chamber.

Many holographic applications exploit the fact that composite repeat holograms of a surface tilted slightly after each exposure can be treated as composite, repeat wave patterns. If two such patterns are matched, a condition arises that is effectively the same as that which exists in ordinary classical two-beam interferometry, in which a single light source is split into two beams and the beams recombined to form interference patterns. Such an arrangement can be set up in several ways; in one, a holographic exposure is made of a surface, then before the hologram is removed or developed, the surface is slightly tilted and a repeat hologram is made, superimposed on the first hologram. When this double hologram is reconstructed, the object as well as the surface covered by the interference fringes caused by surface irregularities can be seen. These fringes reveal microtopographic information about the object.

It is much easier to interpret surface microtopography from fringe patterns that arise when the surface being studied is matched against a perfectly flat surface, rather than from the fringes that arise when the surface is matched against a slight displacement of itself—a condition known in classical interference optics as differential interferometry. One important advantage of holographic interferometry is that it makes possible the examination of rough, nonspecular surfaces as well as smooth, polished, or naturally specular surfaces.

Holographic interferometry can be applied successfully to any situation in which the wave front is modified slightly, no matter how complex the surface may be. Elastic deformation effects can be studied by superimposing the two wave fronts on the hologram, reflected before and after the elastic distortion effect has been introduced. When reconstructed, the hologram provides a clear picture of the object, crossed by interference fringes. Even highly complex shapes respond to this approach in a manner that would be impossible in classical interferometry. There is also great flexibility in the choice of methods used to apply distortions, and even these conditions alone often completely exclude optical interferometry. Not only static distortion but also slow dynamic variations can be studied in this manner. And with pulsed ruby lasers, very fast, short-time variations can be studied.

Use in the study of elastic deformation effects

Time variations in the shape of an object are not usually studied with a single, double-exposure hologram but by an alternative method. First, a hologram is made of the object in its free, unstressed condition. Then the object is stressed and a new hologram made. The stressed hologram is viewed through the original unstressed hologram, and the superposition provides the interference fringe pattern that would have been produced by a double exposure. By such means, time variations can be studied. Valuable studies have been made of mechanically vibrating systems, such as diaphragms, musical instruments (*e.g.*, the belly of a violin), vibrating steam-turbine blades, and the like. The examination of large engineering components, measuring as much as one metre in length, imposes special problems. The distance between the hologram plate and the object must be great enough to ensure that all of the object can be seen at once. In turn, laser power must be increased, high demands on the

coherence of light are imposed, and mechanical stability of the whole setup must be exceptionally good.

When hologram interferometry is applied to the examination of vibrations set up in a rapidly rotating turbine blade, stroboscopic techniques aid the analysis. The laser light is stroboscopically interrupted at the same frequency as the rotation of the turbine blade, and, with the blade thus apparently at rest, a hologram is produced. Consequently, a holographic interferometric pattern is created for the blade whose motion is stopped by stroboscopic action. By slightly altering the frequency of the stroboscope arrangement, a slow scan can be made over the complete vibrational stress pattern to which the blade is subjected. Much information about stresses in turbine blades and other rotating or vibrating objects can be obtained from such holograms.

Although holography can solve many problems, it still is a relatively expensive procedure. It has been used—or misused—in applications more amenable to simpler and cheaper methods. The laser system by itself is a fairly complex and costly piece of equipment, and costs are aggravated further by the additional equipment and the long exposure times required to produce holograms and reconstruct images. Holography is therefore applied only when other methods have failed or are not precise enough.

BIBLIOGRAPHY. D. GABOR, "Microscopy by Reconstructed Wave-Fronts," *Proc. R. Soc.*, Series A, 197:454–487 (1949), original paper first proposing the then new concept of reconstruction of wave fronts; E.N. LEITH and J. UPATNIEKS, "Reconstructed Wavefronts and Communication Theory," *J. Opt. Soc. Am.*, 52:1123–1130 (1962) and "Wavefront Reconstruction with Diffused Illumination and Three Dimensional Objects," *ibid.*, 54:1295–1301 (1964), papers in which the newly developed laser light source was first successfully applied to the creation of holograms; R.L. POWELL and K.A. STETSON, "Interferometric Vibration Analysis by Wavefront Reconstruction," *J. Opt. Soc. Am.*, 55:1593–1598 (1965); M.H. HORMAN, "An Application of Wavefront Reconstruction to Interferometry," *Appl. Optics*, 4:333–336 (1965); and K.A. HAINES and B.P. HILDEBRAND, "Surface-Deformation Measurement Using the Wavefront Reconstruction Technique," *ibid.*, 5:595–602 (1966), three references on a group of early combinations of holography and interferometry for interferometric study of nonspecular objects; G.B. BRANDT, "Techniques and Applications of Holography," *Electro Technol.*, 81:53–72 (1968), on the further application of holographic interferometry to technological objects; J.W.C. GATES, "Holography with Scatter Plates," *J. Scient. Instrum.*, Series 2, 1:989–994 (1968), on interferometric holography for engineering materials; J.W.C. GATES, R.G.N. HALL, and I.N. ROSS, "Holographic Recording Using Frequently-Doubled Radiation at 530 nm," *J. Phys.*, Series E, 3:89–94 (1970), a description of the scatter-plate method of producing matching beams in interference holography; L.H. LIN, "Hologram Formation in Hardened Dichromated Gelatin Films," *Appl. Optics,* 8:963–966 (1969), a description of the replacement of photoplate by gelatin containing dichromate to be used for the production of holograms; D.H. CLOSE *et al.*, "Hologram Recording on Photopolymer Materials," *Appl. Phys. Letter*, 14:159–160 (1969), on the use of a photopolymer material instead of photoplate for making a hologram; Z.J. KISS, "Photochromics," *Physics To-day*, 23:42–52 (1970) and F.S. CHEN, J.T. LA MACCHIA, and D.B. FRASER, "Holographic Storage in Lithium Niobate," *Appl Phys. Letter*, 13:223–225 (1968), descriptions of erasable nonphotographic holograms.

(S.To.)

Holostei

The Holostei are one of the three major groups of ray-finned fish (Actinopterygii). Most paleontologists divide the Holostei into the Holosteans proper, which include living and extinct forms, and the Halecostomi, an extinct group.

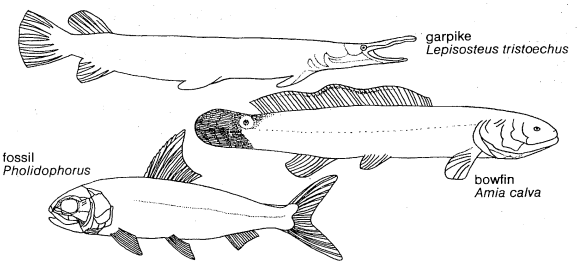
The origin of the Holosteans and the Halecostomi is not fully understood, but it is believed that they arose from some advanced chondrostean fishes (a group that includes the sturgeon). The Holostei were particularly abundant and diversified during the Mesozoic Era (65,000,000–225,000,000 years ago). Today they are represented by only two living genera, *Amia* (bowfin) and *Lepisosteus* (gar). One species of bowfin has been recognized, and about eight species of gar have been described so far.

Size range, distribution, and abundance

The gar occurs only in North America and Central America from southeastern Canada to Panama; it is not found west of the Rocky Mountains. The longnose gar (*L. osseus*) is the most widely distributed species. The gar is primarily a freshwater fish but sometimes ventures into saltwater or brackish water. The so-called alligator gar (*L. spatula*), one of the largest of freshwater fishes, is particularly abundant in the Everglades region of southern Florida, where it is caught locally as a food fish; it sometimes grows to a length of nearly three metres (ten feet) and may attain a weight of 136 kilograms (300 pounds). The names gar, garfish, and garpike are sometimes applied, especially in Europe, to the needlefishes (Belonidae), which are coastal fishes of warm seas and have very long and slender jaws. These fishes, however, are not closely related to the Holostei.

The bowfin, also known as grindle, mudfish, and dogfish, is found in sluggish waters from the Great Lakes to the Gulf of Mexico. It was once common throughout Europe but is now extinct there. Females grow to about 75 centimetres (30 inches) and weigh up to 3.5 kilograms (eight pounds); males are smaller. Bowfins eat all kinds of fish and invertebrates and are sometimes destructive to game-fish populations. Bowfins are seldom caught as food fish.

Drawing by J. Helmer, from (top, centre) N.B. Marshall, *The Life of Fishes* (1965), Weidenfeld & Nicolson, London, (bottom) A.S. Romer, *Vertebrate Paleontology* (1966), University of Chicago Press



Living and fossil forms of Holostei fishes.

Natural history. *Reproduction.* The bowfin spawns in weedy areas along the edges of streams and lakes. The male constructs the nest and guards the eggs as well as the newly hatched young. The young bowfin has an adhesive organ at the tip of its snout that enables it to cling to weeds. The fish grows rapidly and at the end of its first year may be as long as 23 centimetres (nine inches).

The female gar lays its large, yolk-filled eggs in shallow water in the spring. The gar hatchlings grow rapidly, feeding on minnows. The long rows of needle-sharp teeth are effective in capturing fast-swimming prey.

Ecology. Gars and bowfins are voracious predators, feeding on invertebrates and other fishes. All the amiiform fishes (the bowfin order) of former times were probably predaceous. For the most part they were a marine group; the modern bowfin, however, is confined to freshwater. Because of its highly developed air bladder, which can also function as a lung, the bowfin is able to live out of water for as long as 24 hours.

Gars occasionally venture into saltwater, but apparently they do not feed there. They often float quietly at the surface of sluggish waters, breathing atmospheric air.

Form and function. *General features.* The Holostei are characterized by having the dermal bone of the upper jaw (maxilla) freed from the cheek elements and attached to the skull only in the ethmoid region or near the nasal chambers. The palate is separated from the cheek elements, and the adductor mandibulae muscle, which closes the jaws, is larger and more subdivided than it is in the chondrosteans. Primitively, the centrum (*i.e.*, the lower, heavy part of a vertebra) surrounding the notochord (a flexible rod that passes through the vertebral column) was absent, but this structure apparently developed independently in most of the holostean orders. The scales, too, were primitively rhomboidal, or diamond shaped, with a reduced or absent layer of dentine (*i.e.*,

the substance of which teeth are largely made). The scales, however, became thin and cycloidal (*i.e.*, rounded and overlapping) in several groups. The fin rays of the unpaired fins are always equal in number to their basal supports, and the fins themselves may or may not be bordered at the anterior end by fulcra (*i.e.*, modified scales or spines). The caudal, or tail, fin is typically hemiheterocercal (*i.e.*, the upper lobe larger than the lower) and externally symmetrical. The braincase is always composed of separate ossifications (centres of bone formation) that resemble, in number and placement, those found in the teleosts.

The division Holosteans includes the orders Semionotiformes, Pycnodontiformes, Amiiformes, and perhaps Pachycormiformes. In these orders the preoperculum (an L-shaped bone anterior to the operculum, or gill cover) is tied to the palatal elements and provides part of the originating area for the adductor mandibulae muscle.

Extant groups. The order Semionotiformes includes two families. The oldest known holostean, *Acentrophorus* (Upper Permian—about 225,000,000,000–250,000,000 years ago), belongs to the Semionotidae. Members of this family have small mouths and strong teeth; heavily ossified (*i.e.*, composed of true bone rather than cartilage) dermal bones; and hemiheterocercal tails. The body may be fusiform (*i.e.*, tapered at both ends), as in *Semionotus*, or flat and disk-shaped, as in *Dapedium*.

Body structure of living gars

The other semionotiform family, the Lepisosteidae, includes the living gar. The snout of the gar is greatly lengthened by multiplication of the small tooth-bearing bones anterior to the eye. The premaxilla bone is situated at the anterior end of the series; the maxilla bone seems to be reduced to a small, bony sliver at the angle of the mouth.

The body of the gar is encased in an armour of thick, diamond-shaped, enamelled (ganoid) scales. The jaw ends in a beak that, in the alligator gar, is broad and relatively short; in the longnose gar the beak is long and forceps-like. The dorsal and anal fins, both located far back on the body, are without spines and have fewer than 12 rays each.

The Amiiformes, represented today by one species of bowfin, include about six families that show considerable diversity in the length of the jaw, the development of the teeth and fins, and details of the dermal skull pattern. In general, the earlier amiiforms had well-developed rhombic scales and a persistent notochord. In later forms the scales usually became thinner and cycloidal. Ossified centra developed around the notochord, either restricting it or eliminating it. The caudal fin was either forked or lobed. The amiiform body was generally fusiform, similar to that of the living bowfin.

The bowfin has a long, spineless dorsal fin with about 58 rays. This extends over most of the back to near the tail. The males have an orange- or yellow-encircled dark spot near the tail. In females either the outer circle or the entire marking is absent. Bony plates cover the head; the rest of the body has cycloid (*i.e.*, fan-shaped) scales.

Extinct groups. In some ways the Pachycormiformes superficially resemble certain living teleosts, such as mackerels and swordfishes. Their bodies are generally fusiform, with a widely forked caudal fin, a fairly wide gape, and moderately well-developed teeth. The pectoral fins may be elongated and the dorsal and anal fins somewhat enlarged.

The Pycnodontiformes, which may be related to the Semionotiformes, are unique among the holosteans in having their upper and lower dentitions modified to form an open pavement of crushing teeth. In many cases, however, the anterior teeth of the premaxilla and the dentary are incisiform and thus must have been used for grasping (as such teeth are in the living porgies and sparids). In addition to skull modifications related to feeding, the pycnodonts are characterized by deep, almost disk-shaped bodies, elongated anal and dorsal fins, and an externally symmetrical caudal fin. In a number of genera scales are absent on the posterior part of the body, a condition that apparently increased flexibility. Scales were usually present but modified on the anterior half. The body and fin form of pycnodonts suggest that they were fairly fast and powerful swimmers. The affinities of this order remain problematical, as the ossification pattern of the braincase and the caudal-fin skeleton do not closely resemble those of other Holosteans or Halecostomes.

The second major division of the subclass Holostei is the Halecostomes; all are relatively small, fusiform fishes. The group presently includes only one order, the Pholidophoriformes. Some authorities include a second order, Leptolepiformes.

Historical origin of gars and bowfins

Evolution. The gars probably arose in the Cretaceous Period (136,000,000 to 65,000,000 years ago) from some semionotid stock. They are known from freshwater Tertiary deposits in India, Africa, North America, and Europe. The bowfins also made their first appearance in Cretaceous times. Pycnodont fossils range from the Upper Triassic to the Eocene (from about 190,000,000 to 50,000,000 years ago). Pachycormiforms are known only from marine rocks of Jurassic and Cretaceous age. The pholidophorids are known from the Triassic to the Cretaceous; the other pholidophoriform families all ranged from the Jurassic to the Cretaceous.

Among the seven families presently assigned to the Pholidophoriformes, the pholidophorids probably show the closest resemblance to the early teleosts. Trends toward thinning of the scales and the loss of ganoin (an enamel-like material) on the fin rays, along with the dermal-bone pattern and the development of intermuscular bones, point toward the teleosts. The major difference between the pholidophorids and the teleosts is in the structure of the caudal-fin skeleton. In pholidophorids of the early Jurassic, the caudal fin was still structurally heterocercal, with a fairly stiff axial lobe. Modification toward the teleost condition involved changes that brought about equal flexibility of the upper and lower lobes. The other six families currently assigned to the order Pholidophoriformes are specialized in various ways, but none can be regarded as involved in the ancestry of the teleosts.

Classification. Groups marked with a dagger (†) are extinct and known only from fossils.

Distinguishing taxonomic features. The principal features on which classification of the Holostei is based include general body shape, scale structure, and the number and placement of head bones.

Annotated classification.

SUBCLASS HOLOSTEI

Tail hemiheterocercal; maxilla free of preopercle; rays of median fins approximately equal in number to basal elements; trend toward thinning of scales and loss of ganoid (enamel) layer.

Division Holosteans

Preopercle intimately bound to and supporting the posterior border of the palate.

Order Amiiformes (bowfin and fossil relatives)

Lower Triassic (about 210,000,000–225,000,000 years ago) to Recent; body generally fusiform; early forms with well-developed rhombic (diamond-shaped) scales and persistent notochord; scales thinner and cycloidal (fan-shaped) in later forms; 6 families; 1 living species.

†*Order Pachycormiformes*

Lower Jurassic (about 160,000,000–190,000,000 years ago) to Upper Cretaceous (about 65,000,000–100,000,000 years ago); body fusiform, caudal fin widely forked, long snout; 2 families; Europe and North America.

Order Semionotiformes

Upper Permian (about 225,000,000–250,000,000 years ago) to Recent; 2 families of widely divergent fishes; fossil Lepidotidae with normal holostean fusiform bodies, which became relatively deep and slab sided in some members; marine and freshwater, widely distributed; gars (Lepisosteidae) are elongated, long snouted, primarily freshwater predators, extant in North America.

†*Order Pycnodontiformes*

Upper Triassic (about 190,000,000–210,000,000 years ago) to Eocene (about 38,000,000–54,000,000 years ago); upper and lower teeth modified to form crushing pavement; body nearly disk-shaped; anal and dorsal fins elongated; caudal fin externally symmetrical.

Division Halecostomi (or Halecostomes)

Relatively small; body fusiform; preopercle not buttressing the bones of the palate.

†*Order Pholidophoriformes*

Middle Triassic (about 200,000,000 years ago) to Lower Cretaceous (100,000,000–136,000,000 years ago); holosteans with some trends toward teleosts, notably: loss of ganoin from fin rays, scales, and dermal bones; loss of peg and socket joints between scales; about 7 families; marine and freshwater, of wide distribution.

Critical appraisal. According to some authorities, the Leptolepiformes, a teleost group, should be included among the Halecostomes. This opinion indicates that the boundary between the Halecostomes and the Teleostei is difficult to define. The family Pholidophoridae, in particular, has a skull pattern almost identical with that of the leptolepids; the feeding mechanisms are also quite similar.

BIBLIOGRAPHY. S.M. ANDREWS *et al.*, "Pisces," in W.B. HARLAND *et al.* (eds.), *The Fossil Record: A Symposium with Documentation*, ch. 26 (1967), a recent classification of fish, with first and last occurrences for each family; B.G. GARDINER, "A Revision of Certain Actinopterygian and Coelachanth Fishes, Chiefly from the Lower Lias," *Bull. Br. Mus. Nat. Hist. (Geol.)*, 4:239–384 (1960), important revised descriptions of early Jurassic fish from Great Britain; E.S. GOODRICH, "*Vertebrata craniata*," fasc. 1, "Cyclostomes and Fishes," in E.R. LANKESTER (ed.), *A Treatise on Zoology* (1909), a classic work on the anatomy of fish that is still useful; J.P. LEHMAN, "Actinopterygii," in J. PIVETEAU (ed.), *Traité de paléontologie*, vol. 4 (1966), a recent summary of important characteristics of the higher bony fishes, along with their geologic and geographic distributions; D.V. OBRUCHEV (ed.), *Fundamentals of Paleontology*, vol. 11, *Agnatha, Pisces* (1967), a summary treatment of all fishes, living and fossil; D.H. RAYNER, "The Structure and Evolution of the Holostean Fishes," *Biol. Rev.*, 16:218–237 (1941), an attempt to relate the various families of holostean fishes mainly on the basis of braincase design; A.C. WEED, *The Alligator Gar* (1923).

(Ed.)

Holstein, Friedrich von

For the greater part of his career a subordinate official in the German Foreign Office, Friedrich von Holstein nevertheless played a dominant role in German politics and diplomacy. As the "Grey Eminence of the Wilhelmstrasse" (the street in which the foreign office was located), he has long been regarded as a man of mystery who exercised an uncanny influence on German foreign policy.

Born on April 24, 1837, in Schwedt on the Oder, Holstein was raised on his family's estate in Pomerania and their town house in Berlin. Throughout his youth, his family spent a great deal of time travelling abroad, and Holstein became fluent in several foreign languages. A sickly boy, he was educated mostly by private tutors, and, after studying at the University of Berlin, he joined the legal section of the Prussian government.

Bismarck's protégé

Always a proud and self-witted man, Holstein rarely deferred to his superiors and took a cavalier attitude toward his official duties. He could afford to do so, for his family was wealthy, and he himself enjoyed the patronage of a neighbour of his father, Otto von Bismarck, who was already in the 1850s a power in Prussian politics.

With the support of Bismarck, Holstein entered the Prussian diplomatic service in 1860, serving his apprenticeship under Bismarck at the Prussian legation in St. Petersburg. After being posted to Rio de Janeiro, he returned to Germany at the time of Prussia's war with Denmark in 1864, acting as one of Bismarck's diplomatic representatives at army headquarters and taking part in the international conference in London in 1864–65 to settle the Danish question.

From 1865 to 1867 he was stationed in the United States and there had the opportunity to observe the operation of a democratic government at first hand and to travel in what was then still the "wild" West. In America he became interested in a project for the development of a mechanical device for towing barges and invested in this venture the greater part of his fortune, most of which he appears to have lost. Holstein was recalled by Bismarck but not, as has been alleged, as a result of a love affair with the wife of Senator Charles Sumner.

Just before the outbreak of war with France in the summer of 1870, Bismarck, alarmed by the possibility that the Italian monarchy might side with France, instructed Holstein to enter into secret negotiations with Italian republicans. After the war broke out, Holstein grew bored in Berlin and appeared at Bismarck's headquarters in France. He was attached to Bismarck's staff, though officially he did little more than serve as a translator during the armistice negotiations with France. But Bismarck, who liked to be as fully informed as possible, allowed him a more independent role in maintaining unofficial contact with leaders of the Paris Commune, the city's left-wing government that refused the Prussian peace terms and opposed France's regular government.

After the conclusion of peace with France, Holstein served under the German ambassador to Paris, Count Harry von Arnim. An opponent of Bismarck's support of republican France, Arnim was also believed by the Chancellor to be planning to supplant him. When papers were found to be missing from the embassy, Arnim was disgraced. The story spread by Bismarck's enemies that Holstein had served as Bismarck's spy in bringing about Arnim's eventual ruin was proved false in the course of Arnim's trial for the removal of official documents.

Domination of German foreign policy

In April 1876 Holstein was recalled to the German Foreign Office, where his thorough study of every problem and his network of connections soon enabled him to exert a predominant influence not only on foreign policy but also on domestic policy. To remain at the centre of affairs in Berlin, he declined several offers of advancement to diplomatic posts. In 1900 he even refused Chancellor Bernhard, Fürst von Bülow's offer to make him head (state secretary) of the Foreign Office, reportedly because he did not wish to operate in the limelight, a refusal he himself later acknowledged to have been the greatest mistake of his career.

One of the greatest puzzles in Holstein's life was his metamorphosis from an ardent partisan of Bismarck, whom he served during the 1870s as close collaborator and political confidant, into a bitter critic and opponent of the older statesman. This change in attitude, which took place gradually over many years, was in fact largely prompted by Bismarck's alignment with Russia. Holstein advocated instead a firm alliance with Austria and Britain, and after Bismarck's dismissal in 1890, he joined with other counsellors of the new chancellor, Leo von Caprivi, in advising against the renewal of the Russian treaty.

Under Caprivi, Holstein assumed a more important role in the formulation of German foreign policy, for the new chancellor had no experience in this field. Although Holstein had played an important part in formulating the new anti-Russian and pro-British direction of German policy, he also warmly supported Caprivi's reciprocal trade treaties, including the one with Russia, which reduced tariffs on agricultural imports, thus lowering German food costs and stimulating Germany's export trade.

Holstein's influence increased further under Caprivi's successor, Chlodwig Karl Viktor, Fürst zu Hohenlohe-Schillingsfürst, who became chancellor in 1894, and he retained his influential role as the confidential adviser of Hohenlohe's successor, Bernhard von Bülow, who became head of the German Foreign Office in 1897 and chancellor in 1900. Yet Holstein found himself stymied in the most extended and crucial fight of his career. He was powerless to oppose the policies of his unpredictable sovereign, Emperor William II, nor could he persuade his superiors to do so.

Checkmated by the Kaiser

For the most important German policies in the years after Bismarck's dismissal—the feverish quest for colonies, the construction of a German battle fleet and the ensuing Anglo-German naval rivalry, the tortuous negotiations for an Anglo-German alliance that not only failed but actually heightened the tension between the two countries—were in large part inspired by the Kaiser, often without consulting the members of his government.

Holstein saw the folly of many of these policies. Had he restricted himself to warning against them, he might have gone down in history as the Cassandra of the Wilhelmin-

ian era. But the policies he himself formulated and tried to carry out were hardly more beneficial to his country than those of his emperor. After helping to sever the German alliance with Russia, he failed to secure the alliance he desired with Britain; and when he once again sought an understanding with Russia, that country had formed an alliance with France. Meanwhile, Germany was left with only one reliable ally, the moribund Habsburg Empire, which presented ever greater demands in return for its friendship. Holstein's most notable diplomatic campaign, his attempt to break up the newly formed Anglo-French entente of 1904 by fomenting a crisis over Morocco, only served to expose Germany's global isolation. At the height of the crisis, in April 1906, the Emperor dismissed him. Holstein died in Berlin on May 8, 1909.

Assessment

Holstein was a conservative Prussian aristocrat, an individualist, proud, anxious to make his mark in the world, but with very independent ideas as to how to attain his goal. He was not a German nationalist but rather a proponent of the status quo for Germany, for although he had loved and sought adventure in his youth, he feared and disliked adventures in foreign policy. He himself believed he would go down in history, if he were remembered at all, as an intriguer, although in his opinion he had only tried to be of service to his country. Holstein's greatest weakness was his excessive confidence in his own judgment and his immense grasp of political facts, for the wisdom of his decisions was often debatable. He also overemphasized the personal element in any political situation. He liked to think himself the equal of any man, no matter how exalted, and would concede others superiority only in having greater means at their disposal. He did not know the meaning of political fear, and the greater the power of a potential opponent, the more heedlessly he plunged into the fray, whether it was against Bismarck's son-in-law, while Bismarck was still at the height of his power, or against a favourite of the Kaiser. He was equally unconcerned about his economic status. Although he enjoyed spending money while he still had it, in his old age he lived in almost penurious modesty in his bachelor quarters. The last half of his life was completely taken up with politics, which to the day of his death remained his overriding obsession.

BIBLIOGRAPHY. HELMUTH ROGGE, *Friedrich von Holstein: Lebensbekenntnis in Briefen an eine Frau* (1932), *Holstein und Hohenlohe* (1957), and *Holstein und Harden* (1959); and NORMAN RICH, *Friedrich von Holstein: Politics and Diplomacy in the Era of Bismarck and Wilhelm II*, 2 vol. (1965), are authoritative biographies.

(He.R.)

Homeostasis

Homeostasis is the self-regulating process by which biological systems tend to maintain stability while adjusting to conditions that are optimal for survival. If homeostasis is successful, life continues; if unsuccessful, disaster or death ensues. The stability attained is actually a dynamic equilibrium, in which continuous change occurs yet relatively uniform conditions prevail—very much as occurs in a pool below a waterfall. Such systems are called open, as contrasted with the closed systems of, for example, solutions in test tubes. Biological systems are all of the open-system type, and the dynamic equilibrium, or steady state, achieved through homeostatic control implies a controlled environment.

THE NATURE OF HOMEOSTATIC SYSTEMS

The conception of homeostasis. Homeostatic mechanisms are essential to survival in a ceaselessly changing world. Any system in dynamic equilibrium—whether a clock, a cat, a community, or a computer—tends to reach a steady state, a balance that resists outside forces of change. When such a system is disturbed (a clock is wound, a cat becomes hungry, a forest is burned, or a computer is activated), built-in regulatory devices respond to the departures (or feedback) to establish a new balance (the clock unwinds, the cat eats, plants restore the forest, the computer "reads out" a solution). All processes of integration and coordination of function, whether mediated by mechanical circuits or by nervous and hormonal systems, are examples of homeostatic regulation.

Claude Bernard, a 19th-century French physiologist, first expressed the steady-state concept in terms of the constancy of the internal environment of an organism. The idea of self-regulating mechanical devices was already firmly established (James Watt's "governor" and Andrew Ure's "thermostat" were well-known by the late 1800s) when W.B. Cannon, an American physiologist, formalized the biological concept and coined the term homeostasis in his book *The Wisdom of the Body* (1939). Cannon, himself, limited the concept to physiology but suggested it was applicable in many other areas of biology as well. His observation was well founded: today homeostasis is considered one of the most important concepts in biology. Homeostatic responses can be observed in operation at all levels of life—atom, molecule, cell, organ, organism, population, and community—and in a wide range of time intervals—from a fraction of a second for enzyme-catalyzed reactions to hundreds of years for community changes. The concept also can be applied to analogous nonliving systems, such as W. Ross Ashby's automaton "homeostat," guidance controls of aircraft, and digital computers. The application of self-regulation to complex machine operations resulted in devices called servomechanisms. Crystallization of the mathematics of servomechanisms spawned a new science, cybernetics, the strictly biological aspects of which are the province of bionics.

Ludwig von Bertalanffy, a Canadian biologist, expressed the constant change but apparent persistence of living systems as follows: "Living forms are not *in being*, they are *happening;* they are the expression of a perpetual stream of matter and energy which passes the organism [or cell or population] and at the same time constitutes it." This stream of energy and matter appears to be governed by the same laws of thermodynamics that hold for reactions in test tubes and combustion engines. The fact is often obscured by the extraordinary complexity of living systems and the extremely elusive and tenuous energy relationships between life, nonlife, and the ultimate source of energy, the sun.

Technological versus biological systems

Comparison of homeostatic systems. The stability in the self-regulating devices commonly used in engineering and industry is achieved primarily through the use of such regulators as gyroscopes, mechanical governors, and inducers. Biological systems, of greater complexity, have regulators only very roughly comparable to such mechanical devices. The two types of systems are alike, however, in their goals—to sustain activity within prescribed ranges, whether to control the thickness of rolled steel or the pressure within the circulatory system. The basic difference between the two systems is, theoretically at least, that mechanical regulating devices are characterized by rigidity of operation, whereas biological ones are more flexible and adaptable. This becomes a moot distinction, however, in the light of technological advances that have created sophisticated computer-controlled servomechanisms that seem almost to have "minds" of their own.

Through homeostasis, biological systems perform efficiently enough to survive. A tremendous variety of responses is possible to any set of stimuli; however, some sort of filtering or guiding mechanism is necessary if effort is not to be wasted. Homeostasis permits the storing of blueprints, so to speak—instead of spare parts—which are used only if needed. The pattern of response is economical because it is called into play only when predetermined organismic limits are exceeded and then only until the normal range of action has been obtained.

Types of homeostatic regulation. There are two basic types of homeostatic regulators: off–on switches, in which an action does or does not occur, and feedback controls, in which the system is under continuous adjustment. Off–on switches normally react within well-defined limits and range, whereas feedback controls react proportionally to the governing effect of some product of the system. Neither type of adjustment produces a con-

stant response but fluctuates, or oscillates, about some mean value that is preset for off–on controls and self-determining for feedback controls.

Off–on control. A familiar example of off–on regulation is the action of a room-temperature regulator, or thermostat. The heart of the thermostat is a bimetallic strip that responds to temperature changes by completing or disrupting an electrical circuit. When the room cools, the circuit is completed. The furnace operates, and the temperature rises. At a preset level the circuit breaks, the furnace stops, and the temperature drops.

Typical closed-loop system

Feedback control. In a typical feedback or closed-loop system, input (information) is connected to output (effector) by way of a control modulator. Feedback from the output is routed into the system, where it has some effect on the level of activity of the control modulator, negative when it reduces activity, and positive when it accelerates activity. In technological systems the control modulator is a computer, flywheel, or some other driving mechanism. In biological systems, this modulator—and frequently the input as well—is inaccessible within the system and consists of such things as concentrations of substances in the blood. Response of the system to the input information may be either linear, directly proportional to the input, or nonlinear, variable to the input. Biological systems almost always show a nonlinear response.

The homeostatic pattern. *The homeostatic plateau.* The pattern that identifies a homeostatic system represents equally well the level of white blood cells in the bloodstream, the number of mice in a wheat field, or the concentration of phosphorus in bone. The range between high and low levels constitutes the homeostatic plateau—the "normal" range that sustains life. As either of the two extremes is approached, corrective action (through negative feedback) returns the system to the normal range. Oscillations inherent in the system represent the compensatory adjustments to any deviations from the normal. The adjustments are not perfect, and a lag in response is always present. If the lag becomes too great, the compensating reaction may occur out of phase, resulting in overshoots of increasing magnitude. If such overshoots continue, a condition termed "runaway" may occur, resulting in collapse of the system. Under certain circumstances, the homeostatic plateau may be adjusted upward or downward; oscillations then must be maintained within newly established limits if the system is to survive.

"Runaway" resulting in calamity or death

The limits of homeostatic regulation. The homeostatic control of body temperature in humans will clarify some of the subtle aspects of homeostasis as well as some of the limits to its action. Temperature regulation appears to be an off–on type of control, working, superficially at least, like a thermostat that regulates heat in the home. In humans, the normal body temperature fluctuates around the value of 98.6° F. A large number of factors, however, affect this normal value—*e.g.*, hormones, exposure, metabolic rate, and disease. While a high temperature level is advantageous to man, since it permits enzymes to function at an optimal rate, it also creates a continuous demand for a large supply of energy and, at times, confounds the problem of getting rid of excess heat produced by metabolic activities.

Temperature regulation is thought to be controlled by a region in the brain known as the hypothalamus. Feedback information, carried through the bloodstream, results in compensatory adjustments in the breathing rate, the level of blood sugar, and the metabolic rate. Heat loss in mammals is aided by reduction of activity, by sweating and panting, and by heat-exchange mechanisms that permit large amounts of blood to circulate near the surface (*e.g.*, beaver tails or rabbit ears). Heat loss is reduced by insulation (fur or fat), decreased circulation to the skin, and behavioral and cultural modifications such as the use of clothing, shelter, and fire.

Temperature extremes can be survived through hibernation—with the accompanying reduction in metabolic activity—or by temporarily shifting the homeostatic plateau, as happens during a fever.

A SURVEY OF HOMEOSTATIC PROCESSES

Physiological homeostasis. Human physiology, the first great proving ground for homeostasis, continues to be the realm of greatest interest. Communication within cells and between cells helps maintain the proper functioning of the vital enzymatic activities. Internal communication above the cellular level occurs primarily through and between the nervous and endocrine systems. The nervous system is a complex information network the activities of which function through an extensive feedback circuitry. Many of the responses in the nervous system are of the off–on type. Nerve endings, or receptors, provide input information that is routed to the spinal cord and the brain. The output occurs largely through the action of muscles.

Responses of the nervous system

Most of the endocrine glands are under the influence of the hypothalamus and the anterior lobe of the hypophysis (pituitary gland); for example, the hypophysis produces thyroid-stimulating hormone (TSH), which prompts the thyroid gland to produce thyroxine. Thyroxine, in turn, acts in negative feedback manner on the hypophysis to diminish production of TSH—resulting in a balance, or steady state, between the two hormones. Any illness that restricts the production of TSH will automatically promote the uncontrolled production of thyroxine.

In a similar manner, hormones from the hypophysis affect the activities of the ovary. This relationship is complex and involves concurrent interactions of several feedback systems. Follicle-stimulating hormone (FSH), produced by the hypophysis, stimulates the ovary to produce estrogen, which in turn stimulates the hypophysis to produce luteinizing hormone (LH). Luteinizing hormone stimulates the ovary to produce progesterone, which in turn stimulates the hypophysis to produce FSH, and so on. Feedback from each of the hormones helps control this cyclic steady state. The balance is further complicated by the presence of other hormones. Contraceptive "pills" are able to function because of negative feedback; they consist primarily of progesterone, which inhibits LH production and, hence, prevents ovulation and possible conception.

A class of plant hormones called auxin works in homeostatic fashion to regulate plant growth. Acting primarily through regulation of water intake, auxin stimulates or inhibits cell size and growth. The outer coat of some seeds also contains substances that inhibit germination; rain removes the inhibiting influence, permitting the seeds to grow. Dormancy in trees also appears to be controlled by a balance of inhibiting and stimulating substances.

In man, the blood-forming elements are produced in the bone marrow, from embryonic cells called cytoblasts. Cytoblasts may develop into such diverse cell types as bone cells, white blood cells, red blood cells, or platelets—depending upon the feedback demands of the body. If this dynamic equilibrium is disrupted—as in the case of leukemia—an increased number of white blood cells may be produced, to the corresponding detriment of the red blood cells.

Genetical homeostasis. Each of the body cells of an organism contains identical genetic information. But only those genes that control traits or functions that each part of the body is noted for are activated. The capabilities of the organism are limited only by the variety of input information that can be expressed through its complement of genes.

Enzymatic homeostasis

Genes control metabolism through the production of enzymes. Metabolic pathways frequently involve a series of steps, each of which is under the control of a different enzyme. Inhibition of an enzyme may lead to blocked pathways and accumulations of intermediate products. The accumulation of these products—or of the end product for that matter—may in turn stimulate or inhibit other activities.

Genes appear to be of three types: regulator, operator, and structural. Regulator genes produce a substance that represses operator genes. If operator genes are repressed, structural genes cannot function. The presence of certain metabolites, however, can inhibit the regulator genes

and thus free the operator genes to "switch on" the activity of the structural genes.

Several genetic deficiencies (and resulting enzyme deficiencies) block metabolic pathways in similar fashion; for example, if an enzyme that participates in cortisone synthesis is lacking, cortisone inhibition on the control of adrenal activity is lost, and a metabolic disease called the adrenogenital syndrome results.

Genetic homeostasis also occurs at the population level. New gene combinations provide better opportunities for adaptation to changing environmental situations. In population or community dynamics, the death of individuals is a necessity since only with new individuals may new genetic combinations occur. Populations composed of individuals most of whom are pure breeding for a given trait show less adaptability to changing conditions than do populations consisting of hybrid individuals. The greater the variation within the population, the greater its survival value.

It may be that the gradual breakdown of homeostatic stability is inherent in the genetic makeup of organisms —ensuring that death will eventually occur and clear the way for new gene combinations.

Developmental homeostasis. The control of growth and development is a biological question the answer to which—when found—may permit the slowing down or even reversal of the aging process. Aging is a testimonial to the fallibility of homeostasis. Humans grow older in an increasingly uncomfortable and degenerative manner. Eventually death occurs due to unstable homeostatic oscillations beyond the body's compensatory ability. Proposals involving homeostatic control have been advanced to explain the aging process; see AGING.

Long-term versus specific control

Developmental processes are under two kinds of regulation: long-term control, which regulates maturation from fertilized egg to senility, and specific control, which regulates the ultimate size or mass obtained by organs and tissues and stabilizes and maintains their integrity when once formed.

Maturation patterns are highly stable, although alternative pathways of development exist within the genetic makeup of individuals that permit development to continue under a variety of stressful conditions.

Long-term control is somehow exercised over mitosis—the process of cell division—a fundamental aspect of growth the controlling forces of which are not well understood. It is especially perplexing to realize that during the series of mitotic divisions an organism gradually passes through a series of states the appearance, function, and structure of which may be radically different from those of previous generations of cells.

Specific control of the size reached by an organ such as the liver may be based upon an antimitotic substance produced as a consequence of cell division, the gradual accumulation of which ultimately inhibits further cell division through negative feedback. The liver's normal size would be maintained unless injury or disease—with the accompanying reduction in level of antimitotic substance—once again stimulated growth.

The controls that regulate the growth of different tissues are usually independent of one another. Broken bones, for instance, knit without influencing the growth of the kidney. The removal of one kidney, however—and reduction of antimitotic inhibitor—results in the increase in size of the remaining kidney.

Phototropism—the turning of plants toward the light—occurs because auxin is inactivated by light. Auxin on the shaded side of plants functions normally and results in uneven growth that eventually turns plants toward the light. Similar homeostatic controls, through stimulation and inhibition by auxin concentrations, cause roots to grow downward (positive geotropism) and plant shoots to grow upright (negative geotropism).

Ecological homeostasis. Most of the homeostatic regulators in biotic communities are based on feedback controls. Predator–prey relationships are an example. The population level of the predator regulates the population density of the prey. The predator population builds up until competition for the food supply becomes intense and increased numbers of prey are consumed. With the reduced food supply, the predator population level falls—through negative feedback. Starvation and migration of the predators release the inhibiting effect, and the prey become abundant once more. Predator–prey relationships are less well defined in temperate and tropical regions because alternate food sources may be available.

A similar relationship is found in host–parasite arrangements such as fish–lamprey, dog–tapeworm, or corn plant–smut fungus. In each case the abundance of the parasite population is always directly dependent upon the abundance of the host species.

Oscillations occur routinely in populations of land animals such as rodents, rabbits, and insects. Violent oscillations normally occur only in relatively simple ecosystems (as in the Arctic) that contain a limited number of species. Periodic population crashes and die-offs frequently occur as a result of disease or crowding.

Population control

Crowding also limits the growth of many populations. Competition for a limited food supply may be a partial answer to this condition, but there also seems to be some environmental product that reduces the growth rate of many organisms in negative feedback fashion. Suppression of growth through crowding has been noted in populations of protozoans, diatoms, bacteria, water fleas, flatworms, tadpoles, fishes, and muskrats. The inhibiting substances, which in some cases are excretory products, appear to be species specific; *i.e.*, they do not affect the growth rates of other species that may be present. Spacing in terrestrial plants is under homeostatic control in much the same manner. The wide spacing of desert shrubs and trees results in part from inhibiting substances released from the roots and branches of the plants.

Behavioral patterns contribute to homeostasis. The weaving of a snake's head as it searches for a prey may be plotted as an oscillation. Burrowing, staying in the shade, or having periods of quiescence provide negative feedback to help stabilize temperature environments for animals such as insects, lizards, and snakes. Much of learning, in fact, is accomplished through feedback and adjustment, as when a bird learns its song, a horse its gaits, or a child its language.

Medical homeostasis. Hippocrates viewed health as a harmony between man, the forces within his body, and his external environment. This dynamic viewpoint of health still holds a great deal of merit today. Some investigators have suggested that many minor ailments—headaches, upset stomach, diarrhea—may be normal by-products of homeostatic adjustment taking place within the body as the organism seeks to regain its balance or harmony.

The body's response to disease is definitely a homeostatic response. Disease-causing organisms produce severe stresses, which result in oscillations. White blood cells, antibodies, lysosomes, and other defense mechanisms provide negative feedback to steer conditions back to the homeostatic plateau. Physicians may prescribe medicines or other treatment to help preserve the dynamic equilibrium of the body when a person is ill.

Homeostatic mechanisms occasionally overreact, however. Positive feedback and overshoots, resulting in overproduction of a normal body chemical, may create such conditions as rheumatoid arthritis, allergies, epilepsy, multiple sclerosis, cirrhosis of the liver, and nephritis. Cancer is a classical case of homeostasis gone astray. Normally, cell division stabilizes through negative feedback when a certain concentration of cells is reached. In cancer the inhibitors of cell division do not function adequately, and cells continue to divide.

Hunger and thirst, among many bodily needs, are under feedback control. A satiety centre in the hypothalamus influences the amount of food consumed and helps regulate the level of the sugar in the blood. If eating stops when hunger is satisfied, body weight is stabilized. The satiety centre is not in complete control, however, but is modified by other factors, since some people continue to eat long after the hunger signals have ceased.

Mental health is also under homeostatic regulation. The contributing factors include the external world, the

physiological processes of the body, and the workings of the mind and intellect. Psychology and psychiatry make use of homeostatic concepts in explaining malfunctions, overreactions, and exaggerated forms of behaviour. Feedback pressures are in operation here in the same way as in all other homeostatic situations.

HOMEOSTATIC CONTROL HIERARCHIES

The importance of circulation

In organisms. Homeostasis is attained in the organism as a whole through several ordered levels of control and subsystems. In man, circulation maintains the internal environment. The circulation system works in tandem with the renal system (kidneys) for the regulation of such things as water levels in the body. In addition, the kidneys regularly filter all of the contents from the bloodstream except the red blood cells. Most of the materials are subsequently reabsorbed into the blood, but waste materials remain in the urine and are eliminated. Hormones and other chemical substances are transported in the circulatory system and provide the feedback essential to homeostatic control.

Many muscular activities are complexly controlled. Eye movement, in particular, shows a complicated feedback mechanism of subsystems at work. Focussing involves the coordination of six major muscles for each eye, through input from the brain and semicircular canals in the inner ear.

In natural communities. Social insects such as ants and bees show a division of labour that is similar in pattern to the division of labour found in a group of multicellular organisms. In fact, a population or society of individuals may, for illustrative purposes, be considered to be a "super organism." Such an entity functions with specialized subsystems for maintenance, defense, movement, and nutrition integrated homeostatically into a whole. Similarly, several populations of animals and plants, combined into a more highly integrated system, constitute a community. Both negative and positive feedback, in the form of individual and species behavioral patterns, are involved in maintaining the overall dynamic equilibrium of the community. This large steady state is often referred to as a "web of life."

Many activities are keyed to the external environment through biological rhythms. Familiar rhythms include the heartbeat, the daily tides, seasonal changes in vegetation, and the annual migrations of birds. Less noticeable but equally important rhythms are involved in almost every facet of life.

Plant succession—the replacement of one population of plants by another—functions under homeostatic control. Succession is influenced by factors such as soil texture, moisture, temperature, and nutrients. Populations vary in their degrees of adaptation to these factors and therefore show varying degrees of success in surviving. Conditions under which a plant population may flourish change in time through the activities of the population itself until the environmental conditions become more optimal for a population other than the original one; for example, in one type of succession, shrubs replace grass and trees replace shrubs—as a result of negative feedback from self-condemning activities.

Cybernetics uses an approach similar to community analysis for the planning and management of social, governmental, and political problems. Plotting, directing, and maintaining flight in space, for example, is accomplished through these same system analysis techniques (see OPTIMIZATION, THEORY OF).

ORIGIN AND EVOLUTION OF HOMEOSTASIS

Life seems to defy the laws of physics by creating order (organisms) out of chaos (inert matter)—active, self-regulating systems out of passive, acted-upon units. The appearance of such systems can be explained by known scientific laws; it is not necessary to evoke predetermined or preordained designs.

Self-regulation is of primary evolutionary significance, since it operates for the species and the organism as well as for lesser subsystems such as enzyme concentrations. Life is thought to have first developed through chemical evolution in the oceans. The uniform physical environment provided by the seas made the regulatory systems of less importance initially. As life moved to terrestrial environments, however, the need for stable and self-regulating internal environments became fundamental for the continuation of organisms.

Even under homeostatic control, evolution shows a great deal of latitude. It is incidental whether homeostasis functions well, since efficiency of operation is not a criterion—survival is all that nature demands. Some controls are obviously more finely developed than others. The subtle interplay involved in the hormone balance of the estrus cycle is a far cry from the harsh oscillations exhibited in a predator–prey relationship. Yet the results are the same—survival of the system.

What are the limitations of homeostatic regulation, especially for man? The ultimate success or failure of the human race depends upon the adaptive abilities of homeostatic mechanisms the limits of which may already be under challenge by environmental and emotional stresses: contaminants, pollutants, food additives, and crowding. The fact that man can survive in the face of these assaults does not guarantee that the long-term effects of such stresses are harmless. And, although man himself may in time become extinct, life, in some new self-regulatory capacity, will continue.

BIBLIOGRAPHY. BROOKHAVEN SYMPOSIA IN BIOLOGY, *Homeostatic Mechanisms* (1958), an excellent assortment of advanced papers on physiological, mathematical, and philosophical topics; W.B. CANNON, *The Wisdom of the Body*, rev. ed. (1939), the classical and most readable book on homeostasis by a pioneer in the subject; L.J. HENDERSON, *The Fitness of the Environment* (1927), another early work anticipating the concept of homeostasis; G.M. HUGHES (ed.), *Homeostasis and Feedback Mechanisms* (1964), advanced papers presented at the 18th symposium of the Society for Experimental Biology; L.L. LANGLEY, *Homeostasis* (1965), a popular short book of college-level difficulty; T.G. OVERMIRE, *Homeostatic Regulation* (1963), a pamphlet written for high school students; JOHN M. REINER, *The Organism As an Adaptive Control System* (1968), a graduate-level book involving mathematics and engineering concepts of homeostasis; "Automatic Control," *Scient. Am.*, vol. 187 (1952), an entire issue devoted to servomechanisms and other types of feedback control devices as used in industry; H. SELYE, *The Stress of Life* (1956), a readable book dealing with medicine, health, and homeostasis; G.E.W. WOLSTENHOLME and J. KNIGHT (eds.), *Symposium on Homeostatic Regulators* (1969), research reports given at a Ciba Foundation symposium concerned with chemical and physiological aspects of homeostasis.

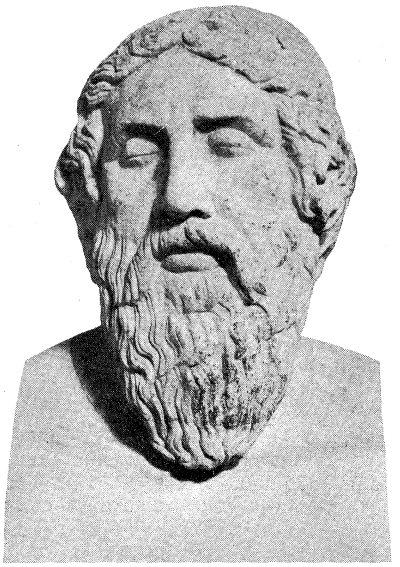
(T.G.O.)

Homer

By Homer is meant the poet or poets primarily responsible for the *Iliad* and *Odyssey*, the two great epic poems of ancient Greece. Very little is known of him beyond the fact that his was the name attached by the Greeks themselves to the two great poems. The dearth of objective information about him has led to such eccentric theories as that of a German critic and philologist that the epics were in origin a kind of spontaneous emanation from the whole people, or Samuel Butler (1835–1902) that the *Odyssey* was composed by a woman. That there was an epic poet called Homer and that he played the primary part in shaping the *Iliad* and *Odyssey*—so much may be said to be probable. If this assumption is accepted, then Homer must assuredly be one of the very greatest of the world's literary artists.

Homer's influence

He is also one of the most influential in the widest sense, for the two epics provided the basis of Greek education and culture throughout the classical age and, indeed, formed the backbone of humane education down to the time of the Roman Empire and the spread of Christianity. Indirectly, through the medium of Virgil's *Aeneid* (which was loosely molded after the pattern of the *Iliad* and *Odyssey*), directly through their revival under Byzantine culture from the late 8th century AD onward, and subsequently through their passage into Italy with the Greek scholars who fled westward from the Ottomans, the Homeric epics had a profound impact on the Renais-



Homer, restored bust copied from a Greek original, *c.* 450 BC. In the Sala delle Muse, the Vatican.
By courtesy of the Vatican Museum

sance culture of Italy. Since then the proliferation of translations has helped to make them the most important poems of the classical European tradition, valued even above the works of Virgil and Dante.

It was probably through their impact on classical Greek culture itself that the *Iliad* and *Odyssey* most subtly affected Western standards and ideas. The Greeks regarded the great epics as something more than works of literature; they knew much of them by heart and valued them not only as a symbol of Hellenic unity and heroism but also as an ancient source of moral and even practical instruction. The roots of the epic lay in the popular oral tradition, and Homer was, thus, a stumbling block for Plato when Plato tried to substitute a more rational model of behaviour for what he conceived to be the artificial histrionics of poetry.

Early references. Implicit references to Homer and quotations from the poems date back to the middle of the 7th century BC. Archilochus, Alcman, Tyrtaeus, and Callinus in the 7th century, Sappho and others in the early 6th, adapted Homeric phraseology and metre to their own purposes and rhythms. At the same time, scenes from the epics became popular in works of art. The pseudo-Homeric Hymn to Delian Apollo, probably of late-7th-century composition, claimed to be the work of "a blind man who dwells in rugged Chios," a reference to a tradition about Homer himself; and, like many others, the historian Thucydides accepted the claim. The idea that Homer had descendants known as "Homeridae," and that they had taken over the preservation and propagation of his poetry, goes back at least to the early 6th century BC. Indeed, it was not long before a kind of Homeric scholarship began: Theagenes of Rhegium in southern Italy toward the end of the same century wrote the first of many allegorizing interpretations. By the 5th century biographical fictions were well under way; the Pre-Socratic philosopher Heracleitus of Ephesus made use of a trivial legend of Homer's death—that it was caused by chagrin at not being able to solve some boys' riddle about catching lice—and the concept of a contest of quotations between Homer and Hesiod (after Homer, the most ancient of Greek poets) may have been initiated in the Sophistic tradition. The historian Herodotus, usually an inspired purveyor of exotic tales, found little to say about Homer but was commendably cautious over his date; after assigning the formulation of Greek theology to Homer and Hesiod he claimed that they could have lived no more than 400 years before his own time, the 5th century BC. This should be contrasted with the superficial idea, popular in many circles throughout antiquity, that Homer must have lived not much later than the Trojan War about which he sang.

Personal history

It is plain from these and other references that factual information about the poet (who gives little away in the epics themselves) is sadly lacking. Even his home is uncertain. The general belief that it was in Ionia (the central part of the western seaboard of Asia Minor) was a reasonable conjecture, supported by the predominantly Ionic dialect of the poems themselves. Cities like Smyrna and Chios early began competing for the honour (the poet Pindar, early in the 5th century BC, associated Homer with both), and others joined in; it is extraordinary but true that nowhere did any authenticated local memory survive of someone who, oral poet or not, must have been remarkable in his time. The absence of hard facts puzzled but did not deter the Greeks; the fictions that had begun even before the 5th century BC were developed in the Alexandrian era in the 3rd and 2nd centuries BC (when false scholarship as well as true abounded) into fantastic pseudobiographies, and these were further refined by derivative scholars under the Roman Empire. The longest to have survived purports to be by Herodotus himself; it does not scruple to assign to Homer several generations of ancestors and a detailed catalog of travels; in spite of a faintly rational tone and the elimination of Orpheus and the Nymphs from the poet's genealogy, however, this work is quite devoid of objective truth.

If the ancient tradition about Homer as a person is so feeble, what can be deduced by modern scholarship, with its wider knowledge of the history and archaeology of the Bronze Age and Early Iron Age and its ability to study the text and texture of the poems in minute detail?

Modern inferences. Modern scholars agree with the ancient sources only about Homer's general place of activity. The most promising information of the ancients was that Homer's descendants, the Homeridae, lived on the Ionic island of Chios. It is theoretically conceivable, of course, that the predominantly Ionic dialect of the poems was simply a convention of the epic tradition and was adopted by Homer for that reason. Yet an east Aegean environment is suggested for the main author of the *Iliad* by certain local references in the poem; *e.g.*, to the peak of Samothrace just appearing over the intervening mass of Imbros when seen from the plain of Troy, to the birds at the mouth of the Cayster near Ephesus, to storms off Icaria and northwest winds from Thrace. East Aegean colouring is fainter in the *Odyssey*, which is set primarily in western Greece; but the poem's vagueness over the position of Ithaca, for example, is not incompatible with the idea of a poet in Ionia elaborating materials derived from the farther side of the Greek world.

Admittedly, there is some doubt over whether the *Iliad* and *Odyssey* were even composed by the same main author. Such doubts began in antiquity itself and depended heavily on the difference of genre (the *Iliad* being martial and heroic, the *Odyssey* picaresque and often fantastic); but they may be reinforced by subtle differences of vocabulary, even apart from those imposed by different subjects. Aristotle's conception of the *Odyssey* as a work of Homer's old age is not impossible; but if the *Iliad* is the earlier of the two (as seems likely from its simpler structure and the greater frequency of relatively late linguistic forms in the *Odyssey*), then the *Odyssey* could have been created after its image, and as a conscious supplement, once the example of monumental composition had been given. In this case the similarities of the two poems could be partly due to the coherence of the heroic poetical tradition that lay behind both. The possibilities are manifold; but at least one can nowadays avoid some of the greater implausibilities of the "Higher Criticism" and the old "Homeric Question" and can reasonably conclude that there *was* a major poet called Homer, that he was an Ionian, that he was substantially the composer of the *Iliad*, and that, whether or not it was directly his work, he was at the very least the inspirer of the *Odyssey*. One other important point about him will be expanded later: that he was an "oral poet," one who lived toward the end of a long line of singers who sang their poems to the ac-

companiment of the kithara, or lyre, and developed a mass of heroic poetry, much of it touching upon the Trojan War, without the use of writing.

Date of the epics

The internal evidence of the poems is of some use in determining when Homer lived. Certain elements of the poetic language, which was an artificial amalgam never exactly reproduced in speech, indicate that the epics were not only post-Mycenaean in composition (that is, later than the end of the Bronze Age around 1100 BC) but also substantially later than the foundation of the first Ionian settlements around 1000 BC. The running together of adjacent short vowels and the disappearance of the semivowel digamma (a letter formerly existing in the Greek alphabet) are the most significant indications of this. At the other end of the time scale the development in the poems of a true definite article, for instance, represents an earlier phase than is exemplified in the poetry of the middle and late 7th century. Both stylistically and metrically, the Homeric poems appear to be earlier than the Hesiodic poems, which many scholars place not long after 700 BC. A different and perhaps more precise criterion is provided by datable objects and practices mentioned in the poems. Nothing, except for one or two probably Athenian additions, seems from this standpoint to be later than around 700; on the other hand, the role assigned in the *Odyssey* to the Phoenicians as traders, together with one or two other phenomena, suggests a date of composition—for the relevant contexts at least—of after 900. A few passages in the *Iliad* may imply a new form of fighting in close formation, dependent on the development of special armour for the foot soldiers (hoplites) after about 750; and references to the Gorgon mask as a decorative motif point in the same direction. It is true that the poems do contain many traditional and archaic elements, and their language and material background are a compound of different constituents originating at different dates. Nonetheless, it seems plausible to conclude that the period of composition of the large-scale epics (as distinct from their much shorter predecessors) was the 9th or 8th century, with several features pointing more clearly to the 8th. The *Odyssey* may belong near the end of this century, the *Iliad* closer to its middle. It may be no coincidence that cults of Homeric heroes tended to spring up toward the end of this same 8th century, and that scenes from the epic begin to appear on pots at just about the same time.

Homer as an oral poet. But even if his name is known and his date and region can be inferred, Homer remains primarily a projection of the great poems themselves. Their qualities are significant of his taste and his view of the world, both as an individual and as heir to the heroic tradition, but they also reveal something more specific about his technique and the kind of poet he was. It has been one of the most important discoveries of Homeric scholarship, associated particularly with the name of an American scholar, Milman Parry (1902–35), that the Homeric tradition was an oral one—that this was a kind of poetry made and passed down by word of mouth, without the intervention of writing. Indeed Homer's own term for a poet is *aoidos*, "singer." The *Odyssey* describes two such poets in some detail: Phemius, the court singer in the palace of Odysseus in Ithaca, and Demodocus, who lived in the town of the semi-mythical Phaeacians and sang both for the nobles in Alcinous' palace and for the assembled public at the games held for Odysseus. On this occasion he sings of the illicit love affair of Ares and Aphrodite in a version that lasts for exactly 100 Homeric verses. This and the other songs assigned to these singers —for example, that of the Trojan Horse, summarized in the *Odyssey*—suggest that ordinary *aoidoi* in the heroic tradition worked with relatively short poems that could be given complete on a single occasion. That is what one would expect, and it is confirmed by the habits of singers and audiences at other periods and in other parts of the world (the tradition of the poet-singers of Muslim Serbia has provided the most fruitful comparison so far). Whatever the favoured occasion for heroic song—whether the aristocratic feast, the religious festival, or popular gatherings in tavern or marketplace—a natural limitation on the length of a poem is imposed by the audience's available time and interest as well as by the singer's own physique and the scope of his repertoire. Such relatively short songs must have provided the backbone of the tradition inherited by Homer, and his portraits of Demodocus and Phemius are likely to be accurate in this respect. What Homer himself seems to have done is to introduce the concept of a quite different style of poetry, in the shape of a monumental poem that required more than a single hour or evening to sing and could achieve new and more complex effects, in literary and psychological terms, than those attainable in the more anecdotal songs of his predecessors.

Poetry sung by the *aoidos*

Poetic techniques. How can one be so confident in classing Homer himself as an oral singer? If he differed from Phemius or Demodocus in terms of length, may he not also have differed radically in his poetic techniques? The very nature of his verse may provide a substantial part of the answer. The style of the poems is "formulaic"; that is, they rely heavily on the use not only of stock epithets and repeated verses or groups of verses—which can also be found to a much lesser extent in a literate imitator like Virgil—but also on a multitude of fixed phrases that are employed time and time again, with occasional minor adjustments, to express a similar idea in a similar part of the verse. The clearest and simplest instance is the so-called noun-epithet formulas. These constitute a veritable system, in which every major god or hero possesses a variety of epithets from which the choice is made solely according to how much of the verse, and which part of it, the singer desires to use up. Odysseus is called divine Odysseus, many-counselled Odysseus, or much-enduring divine Odysseus simply in accordance with the amount of material to be fitted into the remainder of the hexameter (six-foot) verse. A ship is described as black, hollow, or symmetrical not to distinguish this particular ship from others but solely in relation to the qualities of the rhythmical context. The whole noun-epithet system is both extensive and economical—it covers a great variety of subjects with almost no exact reduplication or unnecessary overlap. It would seem that so refined and complex a system could not be the invention of a single poet but must have been gradually evolved in a long-standing tradition that needed both the extension and the economy for functional reasons—that depended on these fixed phrase-units because of its oral nature, in which memory and improvising replace the deliberate, self-correcting, word-by-word progress of the pen-and-paper composer. Admittedly, the rest of Homer's vocabulary is not as markedly formulaic as its noun-epithet aspect (or, another popular example, as its expressions for beginning and ending a speech). Many expressions, many portions of sentences are individually invented for the occasion, or at least so it seems. Even so, there is a strongly formulaic and ready-made component in the artificial language that was used by Homer, including its less conspicuous aspects such as the arrangement of particles, conjunctions, and pronouns.

The monumental poem

It looks, therefore, as though Homer must have trained as an ordinary *aoidos*, who began (like most of the present-day Yugoslav *guslars*) by building up a repertoire of normal-length songs acquired from already established singers. The greatest heroic adventures of the past must already have been prominent in any repertoire, especially the Panhellenic adventures of the Seven against Thebes, the Argonauts, and the Achaean attack on Troy. Some aspects of the Trojan War might already have been expanded into songs of unusual length, though one that was still manageable on a single occasion. Yet the process was presumably carried much further in the making of the monumental *Iliad*, consisting of over 16,000 verses, which would take four long evenings, and perhaps more, to perform. This breakthrough into the monumental, which made exceptional and almost unreasonable demands of audiences, presupposes a singer of quite exceptional capacity and reputation—one who could *impose* the new and essentially nonfunctional concept upon his audiences by the sheer unfamiliar genius of his song. The 8th century BC was in other respects, too, an era of cultural in-

novation, not least in the direction of monumentality; and huge temples (like the early temple of Hera in Samos) and colossal funerary vases (like the mixing bowls and amphoras in the so-called Geometric style from the Dipylon cemetery in Athens) may have found a literary analogue in the idea of a vast poetical treatment of the Trojan War. But in an important sense, Homer was merely extending a tendency of all known oral heroic poetry toward elaboration and expansion. The singer does not acquire a song from another singer by simple memorization. He adjusts what he hears to his existing store of phrases, typical scenes, and themes, and he tends to replace what is unfamiliar to him with something he already knows, or to expand it by adding familiar material that it happens to lack. Every singer in a living oral tradition tends to develop what he acquires. There is an element of improvisation, as well as of memory, in his appropriation of fresh material; and judging by the practice of singers studied from the middle of the last century onward in Russia, Yugoslavia, Cyprus, and Crete the inclination to adjust, elaborate, and improve comes naturally to all oral poets.

Typical elements of an oral poem

Cumulative poetic structure. Homer must have decided to elaborate his materials not only in quality but also in length and complexity. All oral poetry is cumulative in essence; the verse is built up by adding phrase upon phrase, the individual description by adding verse upon verse. The whole plot of a song consists of the progressive accumulation of minor motifs and major themes, from simple ideas like "the hero sets off on a journey"—or "addresses his enemy"—through typical scenes like assemblies of men or gods to developed but standardized thematic complexes like scenes of recognition or reconciliation. Homer seems to have carried this cumulative tendency into new regions; in this as in other respects (for example, in his poetical language) he was applying his own individual vision to the fertile raw material of an extensive and well-known tradition.

The result is even more complex than with an ordinary traditional poem. Understanding the origin and essential qualities of the *Iliad* or *Odyssey* entails trying to sort out not only the separate components of the pre-Homeric tradition but also Homer's own probable contributions, whether distinguishable by their dependence on the specifically monumental structure or by their apparent novelty vis-à-vis the tradition as a whole or by other means. Dialectal and linguistic components must be identified as far as possible—for example, survivals of the Mycenaean language, words used exclusively in the Aeolian cities of the west coast of Asia Minor, or Athenian dialect forms introduced into the poems after the time of Homer; so must details of armour, clothing, houses, burial customs, political geography, and so on, that are likely to be assignable to the Late Bronze Age, the Early Iron Age, or the period of Homer's own activity —at the very least as relatively early or late within the whole range of the poetic tradition down to Homer. These are the tasks of modern Homeric scholarship. Yet such different forms and ideas in Homer are not conveniently separated into distinct sections of the text, which can therefore be assigned to early or late phases of composition. On the contrary, they may co-exist in a single (artificial) linguistic form or a single descriptive phrase. Any member of the tradition, not least Homer himself, may, moreover, have chosen to archaize on one occasion, to innovate on another. One result is that the epics are dubious and equivocal authorities for the reconstruction of historical realities like the attack on Troy or the status of workpeople, just as they are ambiguous source books for early Greek grammar or theology. Another is that they are timeless, generic, not bound to a single world view or period or mode of perception; rather, they have taken on a quality that is genuinely mythic, that unites judgments and experiences never seen together in "real" life into a whole that is literary but nevertheless almost traumatically revealing of the underlying structure of human existence.

Stabilizing the text. An important and difficult question, which affects the accuracy of modern Homeric texts, is that of the date when the epics became really "fixed"—which means reduced to authoritative written form, since oral transmission is always to some extent fluid. An alphabetic writing system reached Greece in the 9th or early 8th century BC; before that was a gap of 200 or 300 years, following the collapse of Mycenaean culture and the disappearance of Linear B (writing with each sign generally representing a syllable), during which Greece seems to have been nonliterate. During that gap, certainly, much of the epic tradition was formed. The earliest alphabetic inscriptions to have survived, a few of them containing brief scraps of hexameter verse, date from around 730 BC onward. Therefore, if Homer created the *Iliad* at some time after 750 BC, he could conceivably have used writing to help him. Some scholars think that he did. Others believe that he may have remained nonliterate (since literacy is not normally associated with oral creativity) but dictated the poem to a literate assistant. Still others believe that the poems may have been preserved orally and not too inaccurately at least until the middle years of the following, the 7th, century, when "literature" in the strict sense appeared in the poetry of Archilochus. There are objections to all three theories, but this much can be generally agreed: that the use of writing was at all events ancillary, that Homer behaved for the most part like a traditional oral poet. Some scholars are convinced that certain of the more subtle effects and cross-references of Homer's poetry would be impossible without the ability to consult a written text. That is doubtful; certainly the capacities even of ordinary oral poets in this direction are constantly surprising to habitual literates.

At least it may be accepted that partial texts of the epics were probably being used by the Homeridae and by professional reciters known as rhapsodes (who were no longer creative and had abandoned the use of the lyre) by the latter part of the 7th century BC. The first *complete* version may well have been that established as a standard for rhapsodic competitions at the great four-yearly festivals at Athens, the Panathenaea, at some time during the 6th century BC. Even that did not permanently fix the text, and from then on the history of the epics was one of periodical distortion followed by progressively more effective acts of stabilization. The most important of these (apart from the widespread dissemination following the growth of the Athenian book trade in the 5th century and the proliferation of libraries after the 4th) were due to the critical work of the Alexandrian scholar Aristarchus of Samothrace in the 2nd century BC, and much later to the propagation of accurate minuscule texts (notably the famous manuscript known as Venetus A of the *Iliad*), incorporating the best results of Greco-Roman scholarship, in the Byzantine world of the Middle Ages. Rare portions of either poem may have been added after, but not long after, the main act of composition: the night expedition that results in the capture of the Trojan spy Dolon and that fills the 10th book of the *Iliad*, some of the underworld scenes in the 11th book of the *Odyssey*, and much of the ending of the *Odyssey* after line 296 of the 23rd book (regarded by Aristarchus as its original conclusion) —these are the most probable candidates on the grounds of structure, language, and style.

Even apart from the possibilities of medium-scale elaboration, the *Iliad* and *Odyssey* exemplify certain of the minor inconsistencies of all oral poetry, and occasionally the composer's amalgamation of traditional material into a large-scale structure shows through. Yet the overriding impression is one of powerful unity.

The story of the *Iliad*

The Iliad. Not only a distillation of the whole protracted war against Troy, the *Iliad* is at the same time an exploration of the heroic ideal in all its self-contradictoriness—its insane and grasping pride, its magnificent but animal strength, its ultimate if obtuse humanity. The poem is, in truth, the story of the wrath of Achilles, the greatest warrior, that is announced in its very first words; yet for thousands of verses on end Achilles is an unseen and unmentioned presence as he broods among his Myrmidons, waiting for Zeus's promise to be fulfilled—the promise that the Trojans will fire the Achaean ships and

force King Agamemnon to beg him to return to the fight. Much of the poetry between the first book, in which the quarrel flares up, and the 16th, in which Achilles makes the crucial concession of allowing his friend Patroclus to fight on his behalf, consists of long but subtly distinct scenes of battle. Moreover it is carefully punctuated by highly individualized episodes and set pieces: the catalog of troop contingents, the duels between Paris and Menelaus and Ajax and Hector, Helen's identifying of the Achaean princes, Agamemnon inspecting his troops, the triumph of Diomedes, Hector's famous meeting back in Troy with his wife Andromache, the building of the Achaean wall, the unsuccessful embassy to Achilles, the night expedition, Hera's seduction of Zeus and Poseidon's subsequent invigoration of the Achaeans. Patroclus' death two-thirds of the way through the poem brings Achilles back into the fight, although not before the recovery of Patroclus' body, the making of new divine armour for Achilles, and his formal reconciliation with Agamemnon. In book 22 he kills the deluded Hector; next he corroborates his heroic status by means of the funeral games for Patroclus; and in the concluding book Achilles is compelled by the gods to restore civilized values and his own magnanimity by surrendering Hector's body to King Priam.

The Odyssey. The *Odyssey* tends to be blander in expression and less vigorous in the progress of the action, but it presents an even more complex and harmonious structure than the *Iliad*.

The story of the *Odyssey*

The main elements are the situation in Ithaca, where Penelope and the young Telemachus are powerless before her arrogant suitors as they despair of Odysseus' return; Telemachus' secret journey to the Peloponnese for news of his father, and his encounters there with Nestor, Menelaus, and Helen; Odysseus' dangerous passage from Calypso's island to that of the Phaeacians, and his narrative there (from book 9 to book 12) of his previous adventures after leaving Troy; his arrival back in Ithaca at the poem's halfway point, followed by his elaborate disguises, his self-revelation to the faithful swineherd Eumaeus and then to Telemachus, their complicated plan for disposing of the suitors, and its gory fulfillment. Finally comes the recognition by his faithful Penelope.

Homer's influence seems to have been at its strongest in some of the most conspicuous formal components of the poems. The participation of the gods in human events confers, at different times, lightness, severity, or universality; it must for long have been part of the heroic tradition, but the frequency and the richness of the divine assemblies in the *Iliad*, or the peculiarly personal and ambivalent relationship between Odysseus and Athena in the *Odyssey*, probably reflect the taste and capacity of the main composer. The manysidedness of battle, the equivocal realism of death in a hundred forms, must have been developed among Homer's predecessors but can never before have been deployed with such massive and complex effect. In the extended similes the strain of heroic action is relieved by the illuminating intrusion of a quite different and often peaceful contemporary world, in images developed for a moment gratuitously and almost longingly beyond the immediate point of comparison. These similes, in their placing and their profusion at least, surely depend on the main composer. And yet, beyond such general intuitions as these, the attempt to isolate his special contributions often becomes self-defeating. The *Iliad* and *Odyssey* owe their unique status precisely to the almost spiritual and therefore unanalyzable confluence of tradition and design, the generic and the particular, the divine or heroic and the intensely human, the crystalline fixity of a formulaic style and the mobile spontaneity of a brilliant personal vision. "Homer" implies, above all, this fusion.

BIBLIOGRAPHY

Greek text: Oxford Classical Texts, ed. by T.W. ALLEN (n.d.).

Translations: by RICHMOND LATTIMORE (*The Iliad*, 1962; *The Odyssey*, 1968), the best close modern translations; by A. LANG, W. LEAF, and E. MYERS (*The Iliad*, Globe edition 1914); and by S.H. BUTCHER and A. LANG (*The Odyssey*, Loeb series 1919), long-established and accurate translations in the Victorian-archaic style.

Commentaries: W. LEAF and M.A. BAYFIELD (1923–24), for the *Iliad;* and W.B. STANFORD (1947–48), for the *Odyssey*. The former is Analytical and the latter Unitarian in approach.

Critical studies: M.I. FINLEY, *The World of Odysseus* (1954); G.S. KIRK, *The Songs of Homer* (1962), abbreviated as *Homer and the Epic* (1965), and (ed.), *The Language and Background of Homer* (1964); A.B. LORD, *The Singer of Tales* (1960); PAUL MAZON (ed.), *Introduction à l'Iliade* (1959); D.L. PAGE, *History and the Homeric Iliad* (1959), *The Homeric Odyssey* (1955); M. PARRY, *L'Epithète traditionnelle dans Homère* (1928); A.J.B. WACE and F.H. STUBBINGS (eds.), *A Companion to Homer* (1962); T.B.L. WEBSTER, *From Mycenae to Homer* (1958); W.J. WOODHOUSE, *The Composition of Homer's Odyssey* (1930).

Allied studies: J. CHADWICK, *The Decipherment of Linear B*, 2nd ed. (1968); M.I. FINLEY, *Early Greece: The Bronze and Archaic Ages* (1970); LORD WILLIAM TAYLOUR, *The Mycenaeans* (1964).

(G.S.K.)

Homer, Winslow

One of the most important of all American artists, Winslow Homer is especially admired for his brilliant watercolours and paintings of the sea. He combined exceptional technical ability with a rare sense of colour, texture, and design. His subjects, often deceptively simple on the surface, dealt in their most serious moments with the theme of man's efforts to establish his humanness in the face of an indifferent universe.

Homer was born on February 24, 1836, in Boston, to an old New England family. When he was six, the family moved across the Charles River to Cambridge, then a rural village, where he enjoyed a happy country childhood. His artistic inclinations were encouraged by his mother, an amateur painter. When he was 19, he was apprenticed to the lithographic firm of John Bufford in Boston. At first, most of his work involved copying the designs of other artists, but within a few years he was submitting his own drawings for publication in such periodicals as *Ballou's Pictorial* and *Harper's Weekly*. In 1859 Homer moved from Boston to New York to begin a career as a free-lance illustrator. The following year he exhibited his first paintings at the National Academy of Design.

Civil War drawings

With the outbreak of the Civil War, Homer made drawings at the front for *Harper's*, but unlike most artist-correspondents his scenes dealt more often with everyday camp life than fighting. As the war dragged on, he concentrated increasingly on painting. In 1865 he was elected to the National Academy of Design. Admirably capturing the dominant national mood of reconciliation, his "Prisoners from the Front" was warmly received when exhibited at the academy shortly after the war ended.

Although Homer's studio was in New York City, the city was rarely his theme. During the warm months he travelled to Pennsylvania, the Hudson River Valley, and New England, camping, hunting, fishing, and sketching. In 1866 he went to France for about a year. Although influenced by French Naturalism, Japanese prints, and contemporary fashion illustration, his work after his return to America did not change markedly, except that the pictures were generally somewhat brighter. Such early pictures as "Snap the Whip" and "Long Branch, New Jersey" depict happy scenes, the former of children frolicking in a meadow after school and the latter of fashionable ladies promenading along the seashore. In a few early pictures a disquieting note of human isolation is struck, premonitory of Homer's later, more powerful work.

In 1873 Homer began to work in watercolour, which allowed him to make rapid, fresh observations of nature. In this demanding medium he explored and resolved new artistic problems, and paintings of the next few years, such as "Breezing Up," reflect the invigorating effect of the watercolours.

Winslow Homer matured slowly as an artist, but his development was constant. With the passage of years his oil paintings became larger, his figures more solitary, his concern for naturalistic detail greater. He painted many

women, increasingly as single figures, intimate, withdrawn, feminine. From the late 1870s Homer began to devote his summers exclusively to direct painting from nature in watercolour. Greater concern for atmospheric effects and reflected light added complexity to the images but at the same time enabled him to achieve greater pictorial unity.

Although Homer received some recognition during his early years, he had not had any real success by midcareer. By 1880 he began to show signs of increasing antisociality, deliberately shunning the company of other people. In 1881 he unexpectedly went to England, where he spent about two years sketching and painting in Tynemouth, a remote fishing port on the North Sea. Here, at the age of 45, his period of greatest artistic growth began. He was intrigued by the life of the hardy fisherfolk of Tynemouth, who struggled against the sea to earn their livelihood, but he did not paint that struggle directly. He depicted instead the robust and courageous women of Tynemouth, who mended the nets, kept house, and waited for their men to return from the sea. The English coastal atmosphere posed a new and difficult artistic challenge, but Homer mastered the diffused light, limited in colour but infinitely varied in tone, in a series of subtle watercolours.

Period of greatest artistic growth

After Homer's return to America in 1883, the sea became the dominant theme in his work. He moved to Prouts Neck, a fishing village on the bleak, desolate coast of Maine. He travelled extensively but always returned to his Prouts Neck studio, with its large balcony facing the sea like the bridge of a ship, to convert his sketches into major paintings. Solitude became for Homer not simply a preference but an absolute necessity, as he turned his mind and his art to subjects dealing with man's fate in confronting the elemental forces of nature.

In the summer of 1883 Homer saw a demonstration in Atlantic City of the use of a breeches buoy for rescue from the sea. The following year he painted his large, impressive, and immediately popular painting "The Life Line," one of several he did at this time on the rescue theme, depicting the dramatic transfer of an unconscious female from a wrecked ship to shore.

Seascape paintings

During the next few years, Homer's interest shifted from the edge of the sea to the sea itself. Perhaps inspired by a putative trip to the Grand Banks of Newfoundland with a fishing fleet, he painted heroic men in the act of pitting their strength, intelligence, and experience against the mighty sea. In the most impressive of these works, "Fog Warning," night is falling, fog is rolling in, and a lone fisherman in a dory calculates the distance and the time remaining for him to get back to his home ship in safety. Although the monumental narrative paintings Homer produced in his studio in the mid-1880s lack the freshness of his earlier works, Homer simultaneously painted innumerable brilliantly coloured watercolours during his travels north to Canada and south to the Caribbean.

While Homer's fishermen and their women are heroic in their confrontations with the physical world, the artist occasionally took a more jaundiced view of his fellowman. In "Huntsman and Dogs" of 1891, set in a cheerless autumnal landscape, a sullen-faced young hunter, pausing on a hillside levelled by timbering and blackened by fire, epitomizes man as a despoiler of nature, killing for trophies rather than food.

Homer abandoned the human subject entirely in "The Fox Hunt" of 1893. A fox ventures forth to forage for berries on the snow-covered land, and a sinister line of starved black crows converges to attack him. The ensuing life-and-death struggle will be over quickly, but the pulse of nature that drives the winter ocean against the cliffs in the distance will go on forever. "Northeaster" (1895) distills this theme, and only the viewer witnesses the endless struggle between the irresistible sea and the immovable rocky shore. In "Northeaster," Homer successfully wedded the freshness of his watercolours to the power of his oils to achieve an impressive pictorial effect that, as in many of his later works, transcends the subject matter.

"The Gulf Stream" (1899) stands at the apex of Homer's career. A black man lies inert on the deck of a small sailboat. A hurricane has shredded the sails, snapped off the mast, and snatched away the rudder. Unlike the boys in "Breezing Up" or the fisherman in "Fog Warning," this man is powerless to control his vessel. He is at the mercy of the elements. Sharks circle the boat, a waterspout hovers in the distance, and a boat on the distant horizon passes by unseeing and unseen. As in Stephen Crane's comparable short story, *The Open Boat*, nature is seen as not caring whether a man lives or dies.

Homer, ever more crusty and isolated in his old age, continued to paint vigorously and adventurously through the first decade of the 20th century. Similar in subject matter to his earlier work, although with more emphasis on pure seascape, his late paintings, in their unconventional composition and brilliant colour, reflect increasing concern with the abstract and expressive possibilities of art. Homer died in his Prouts Neck studio on September 29, 1910. Although by the 1890s he had become generally recognized as one of the leading American painters, and his work brought top prices, his passing was but briefly noted, and appreciation of his artistic achievement came only in the years following his death.

MAJOR WORKS

PAINTINGS IN OIL: "Pitching Quoits" (1865; Fogg Art Museum, Cambridge, Massachusetts); "Prisoners from the Front" (1866; Metropolitan Museum of Art, New York); "Croquet Scene" (1866; Art Institute of Chicago); "The Bridle Path, White Mountains" (1868; Sterling and Francine Clark Art Institute, Williamstown, Massachusetts); "Long Branch, New Jersey" (1869; Museum of Fine Arts, Boston); "High Tide" (1870; Metropolitan Museum of Art, New York); "The Country School" (1871; City Art Museum, St. Louis, Missouri); "Snap the Whip" (1872; Butler Institute of American Art, Youngstown, Ohio); "Breezing Up," or "A Fair Wind" (1876; National Gallery of Art, Washington, D.C.); "The Gale" (1883, repainted 1893; Worcester Art Museum, Massachusetts); "The Life Line" (1884; Philadelphia Museum of Art); "Fog Warning" (1885; Museum of Fine Arts, Boston); "The Herring Net" (1885; Art Institute of Chicago); "Eight Bells" (1886; Addison Gallery of American Art, Andover, Massachusetts); "Huntsman and Dogs" (1891; Philadelphia Museum of Art); "Northeaster" (1895; Metropolitan Museum of Art, New York); "The Lookout—'All's Well'" (1896; Museum of Fine Arts, Boston); "The Gulf Stream" (1899; Metropolitan Museum of Art, New York); "Huntsman and Dogs" (1891; Philadelphia Museum of Art); "West Point Prouts Neck" (1900; Clark Art Institute, Williamstown, Massachusetts); "Searchlight Harbor Entrance, Santiago de Cuba" (1901; Metropolitan Museum of Art, New York); "Early Morning After a Storm at Sea" (1902; Cleveland Museum of Art).

PAINTINGS IN WATERCOLOUR: "Fresh Air" (1878; Brooklyn Museum, New York); "Gloucester Harbor and Dory" (1880?; Fogg Art Museum, Cambridge, Massachusetts); "The Green Dory" (1880?; Museum of Fine Arts, Boston); "Inside the Bar, Tynemouth" (1883; Metropolitan Museum of Art, New York); "Tynemouth" (1883; Fogg Art Museum, Cambridge, Massachusetts); "Conch Divers" (1885; Minneapolis Institute of Arts, Minnesota); "Shark Fishing" (1885; private collection, New York); "Street Scene, Santiago, Cuba" (1885; Philadelphia Museum of Art); "After the Hunt" (1892; Los Angeles County Museum of Art); "Adirondacks" (1892; Fogg Art Museum, Cambridge, Massachusetts); "The Adirondack Guide" (Museum of Fine Arts, Boston); "The Turtle Pound" (1898; Brooklyn Museum, New York); "After the Hurricane, Bahamas" (1899; Art Institute of Chicago).

BIBLIOGRAPHY. PHILIP C. BEAM, *Winslow Homer at Prout's Neck* (1966), useful documentation of Homer's later years; WILLIAM HOWE DOWNES, *Life and Works of Winslow Homer* (1911), a pioneering study published the year after the artist's death; JAMES THOMAS FLEXNER, *The World of Winslow Homer, 1836–1910* (1966), an imaginative presentation with good plates that sets the artistic context within which Homer worked; ALBERT TEN EYCK GARDNER, *Winslow Homer, American Artist: His World and His Work* (1961), an attempt to demonstrate the decisive influence of Homer's year in France upon his subsequent work; LLOYD GOODRICH, *The Graphic Art of Winslow Homer* (1968), a basic catalog of Homer's graphics; *Winslow Homer* (1944), the most comprehensive and important study of Homer to date; *Winslow Homer* (1959), a useful picture book with concise text.

(J.D.Pro.)

Hominidae

Hominidae (superfamily Hominoidea, infraorder Anthropoidea, order Primates) is the taxonomic family to which man (*Homo sapiens*) belongs. *Homo sapiens* is the sole genus and species of living primates included in the family Hominidae. The Hominidae also include extinct populations or lineages of primates known only from fossil remains. At least three genera usually are recognized: *Ramapithecus*, a late Miocene–Pliocene genus; *Australopithecus*, a Pleistocene form; and *Homo*, a Pleistocene–Recent form. The living species *Homo sapiens* is, perhaps, the most successful mammalian species to have evolved. It is highly variable and exhibits much greater sexual dimorphism than do most other primates. Since the days of Charles Darwin (1809–1882) it has been clear that *Homo sapiens* is a zoological object, an animal that reached its present biological status through genetic mutation, natural selection, and adaptation to changing environments on earth. These processes characterize the evolution of all other living species. The most generally accepted and the most empirically plausible thesis is that man and the extinct Hominidae are part of a natural unit of quite closely related animals, the mammalian order Primates. In view of the evidence, the formerly sacrosanct opinion that man is a unique, separate creation sharply waned, many religious leaders endorsing the evolutionary interpretation.

In the not too distant past it was believed a tenable hypothesis that Hominidae, containing only *Homo sapiens*, was a taxonomic entity of considerably higher rank in the Linnaean hierarchy than that of a taxonomic family. According to this hypothesis, ancestors of *Homo sapiens* were a lineage that had evolved separately since the earliest differentiation of the primates, stemming directly from a tarsier-like prosimian of the Eocene Epoch. This is no longer considered tenable, particularly since the fossil record suggests that the Pongidae (anthropoid apes) and the Hominidae had a common ancestor no later than the Oligocene Epoch or early Miocene (roughly 30,000,000 years ago) and that each was clearly distinct by late Miocene times (perhaps 15,000,000 years ago).

Difficulties in classifying man and other hominids

Difficulties in interpreting the evolutionary history of the Hominidae largely stem from attitudes conditioned by the more traditional studies of human variation. *Homo sapiens* ranges from less than four feet tall among so-called Pygmies of Africa and some islands of East Asia, to almost seven feet among some residents of Central Africa, northern Europe, and North America. Fossil Hominidae were smaller than most living men; australopithecines were probably under five feet tall, although some were massively built. Variability in stature is matched to some extent in modern man by that of other features, such as pigmentation or shape of nose.

Homo sapiens is extraordinarily egocentric and finds fossil hominids almost as fascinating as his living fellows. Every little variation among members of his species is given weighty consideration by some scholar. One trait excessively emphasized in attempts to classify *Homo sapiens* is skin colour. Considerable variation is found as well in eye colour, hair colour and form, and skull form. All, and more, have been used to define or describe subspecies or races of man. At least 15 subspecies of *Homo sapiens* have been proposed formally since the days of Linnaeus, and many more informally.

No wonder, then, that differences in teeth, jaws, and skulls among primate fossils were treated as generic, familial, and even superfamilial distinctions. Many such fossils, only now being related to the ancestral lineage from which pongids and hominids diverged, were left to languish in taxa that suggested no relationship at all to the Hominidae. Several fossils now referable to the Hominidae were placed in the lineage of living anthropoid apes, being given generic status within the Dryopithecinae (family Pongidae, superfamily Hominoidea). William King Gregory seems to have been the only investigator of the first half of the 20th century to suspect something amiss, stating that the ancestors of modern man would be found within that group.

Homo sapiens occupies every environment on the planet: windy, cold, desolate Tierra del Fuego; the high Andes; the frigid Arctic; the deserts of Africa; the steppes of Central Asia; the humid, hot, tropical rain forests; the temperate climates of Europe and North America. Antarctica appears to be the only major land mass on which *Homo sapiens* did not live until recently.

Hominid age, distribution, and abundance

Ancient Hominidae also were widely distributed; fossil hominids from 1,000,000 to 5,000,000 years old are found in South Africa, East Africa, North Africa, and Asia. The Dryopithecinae, related to the common ancestor of Pongidae and Hominidae, also are found widely distributed in the Old World. Hominids of between 600,000 and 800,000 years ago lived in northern Europe where the climate was particularly rigorous as well as in the more hospitable regions of Asia and Africa. The hominid genus *Ramapithecus* has been found in Africa and India. Several authorities accept evidence that dates *Ramapithecus* as at least 14,000,000 years old, suggesting that species of the Hominidae in their earliest forms were widely distributed over the Eurasian–African land mass. They apparently did not penetrate the New World, however, until less than 100,000 years ago, well after the most modern form of *Homo sapiens* developed.

Australopithecine fossil fragments found in southern and eastern Africa as well as those from North and Central Africa and Asia constituted several hundred individuals by the 1970s. At least 200 examples of later fossil Hominidae were known, sometimes colloquially called Peking man, Java man, and Neanderthal man (all quite referable to the species *Homo sapiens*). The distribution and number of fossil fragments suggest that the Hominidae, although not excessively abundant until recently, were present over most of the major land masses of the Old World and attained respectable population size. Rather fragmentary remains of *Ramapithecus* from East Africa and India support the view that hominids were widely distributed as early as the late Miocene.

Unsettled problems in classification

Controversy on the number of genera and species within the Hominidae continued in the 1970s. The usual position—that Hominidae consist of the genera *Ramapithecus*, *Australopithecus*, and *Homo*—has been noted above. A more conservative interpretation favoured including the genus *Paranthropus*, an australopithecine. A more radical view was to eliminate the genus *Australopithecus*, placing both *Australopithecus* and *Paranthropus* in the genus *Homo*.

PHYSIOLOGICAL AND BIOLOGICAL CHARACTERISTICS

Body structure. Skeleton and musculature among hominids are adapted for erect posture, bipedal gait, and extreme functional differentiation of the limbs. The hand of the forelimb is uniquely suited for grasping and for very fine, skillful manipulation. The hindlimb's foot is a pedestal for the weight of the body, the foot and leg together providing a strong propulsive lever for such activities as walking, standing, running, and ballet dancing.

Orthograde posture. Orthograde (upright) posture is common to all Primates. While man is an erect biped, all other living primates are best characterized as orthograde quadrupeds. Similarities in the structure of the trunk found among Primates are likely the result of this upright posture, a fundamental characteristic of all members of the order. Orthograde posture is habitually assumed by even the most primitive of Primates but has been so refined in mature humans that erect posture is the only one normally taken. Far too little is yet known of the postcranial skeletons of the earliest hominids to estimate how long almost wholly erect posture has been a trait of the Hominidae. The existing evidence suggests that australopithecines were essentially erect, their immediate predecessors having been well on the way to upright posture. Alternative views have been based on differing interpretations of the fragmentary skeletal remains of a single foot from the australopithecine deposit at Olduvai Gorge, Tanzania.

Bipedal locomotion. Hominidae are distinguished by evolutionary trends that are particularly evident in progressive modifications of the skeleton for erect posture and bipedal gait. The most significant include proportionate lengthening of the lower extremity and changes in proportion and in morphological details of the pelvis, femur, and foot (Figures 1 and 2). All are related to the mechanical requirements of bipedal locomotion.

From J. Buettner-Janusch, *The Origins of Man* (1966); John Wiley & Sons, Inc.

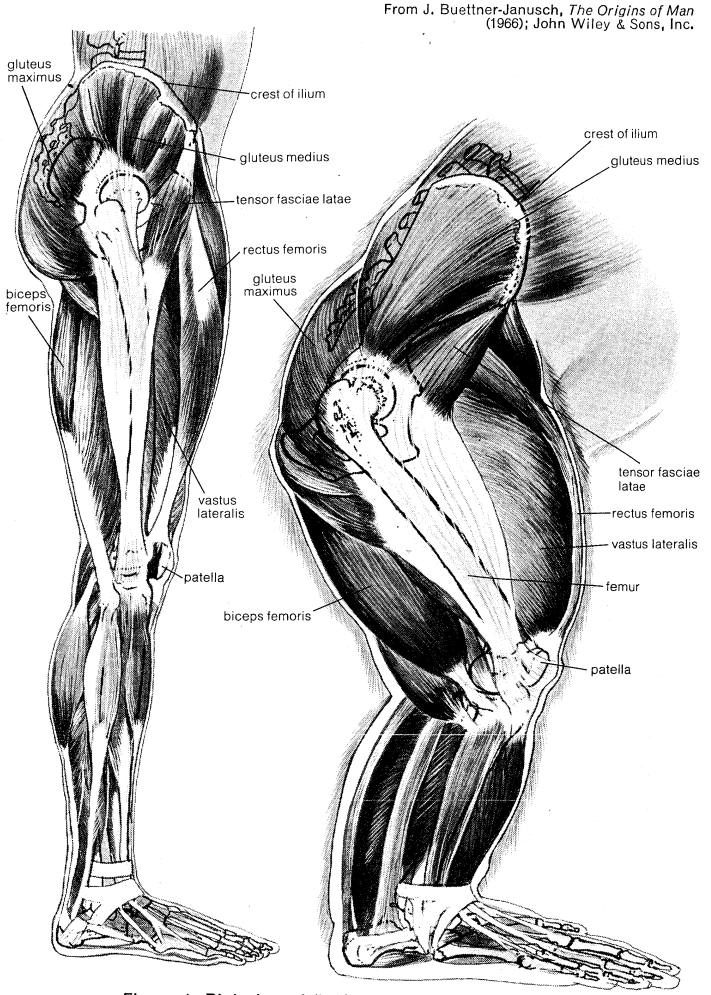


Figure 1: Right leg of (left) man and (right) gorilla.

The hominid pelvis includes at least three bones (ilium, ischium, and pubis) that mirror evolutionary change. The form of the pelvis in man is a functional compromise that permits erect and stable bipedal locomotion, providing support to keep the trunk from falling backward when one stands upright. Including a birth canal adequate for the large-brained human fetus, pelvic structure also serves to support a variety of viscera.

The earliest bipedal hominid seems to have been *Australopithecus*, on the basis of inferences from several well-preserved pelvic fragments and an almost intact fossil pelvis. Australopithecine pelvic and limb bones differ from those of *Homo* in their relationships among several anatomical landmarks, exhibiting forward prolongation in the region of the anterior superior spine of the ilium and a relatively small sacroiliac surface. The australopithecine ischial tuberosity is usually low, and there is a marked forward prolongation of the intercondylar notch of the femur (see SKELETAL SYSTEM, HUMAN).

Inferences from pelvic bones

Analyses of australopithecine and modern human pelvises allow us to postulate evolutionary changes that led to bipedalism: the ilium became shortened and bent back on the ischium; the angle between ilium and ischium grew smaller. Such bone changes seem to have been part of a process that brought the gluteus maximus muscle behind the hip joint and made it a powerful extensor of the leg.

Although many postural and locomotor demands have been met, the human pelvis is not the efficient structure an engineer might design. Many disorders in man arise from incomplete adaptation of the back and pelvic girdle to erect bipedalism. The lumbosacral region of the spine is structurally weak, and mechanical strains will produce such severe ailments as herniated intervertebral disks and spondylolisthesis (forward displacement of individual vertebrae) and less severe, but common, chronic lower back pain.

Efforts to trace the evolution of gait and posture among the Hominidae face the critical question of whether the australopithecine pelvis is that of a hominid or more like that of a pongid. At least one characteristic of australopithecine pelvic remains is clearly hominid—the ilium is bent back, shortened, and broadened so that the iliac blade has the characteristics expected of a pelvis that is part of the bipedal complex. Investigators who favour the view that the earliest known hominid pelvis is that of the australopithecines are well aware that it is not exactly like that of the contemporary hominid pelvis in other respects. It is the major functional change, however, that assumes significance in attempts to reconstruct the past.

Inferences from bones of the foot

The foot is essential in the locomotor adaptation of the Hominidae. Structural modifications of the ancestral prehensile primate foot produced man's plantigrade platform; *i.e.*, sole and heel both touch the ground. The foot of erect bipedal man must completely support the body and be strong enough to lift the load by its lever action. Specializations that permit this include the shape of the arch and the position and robustness of the big toe. Bones associated with the lever action of the foot are in such proportion that they provide support adequate to man's unique walk; man strides. In striding, the repeated sequence in which the foot bears the weight is heel, lateral edge of the foot, big toe. The big toe bears all the weight at the end of one step and at the beginning of the next. The metatarsal and phalangeal bones of man's big toe are more robust than are those of the other toes.

A foot skeleton of *Australopithecus* can be reconstructed from a dozen bones recovered at Olduvai Gorge. Many of the proportions and articulations among the bones are typically hominid. The sturdy first metatarsal of the Olduvai australopithecine resembles that of contemporary man. While the bone is not as robust relative to the other metatarsals as is that of *Homo*, the Olduvai hominid was a much smaller animal.

Forelimb structure and manipulation. Man's hand is an impressive organ that distinguishes him from all other living primates. Beyond nonlocomotor and nonprehensile function (*e.g.*, shoving, clubbing, or poking with the fingers), prehension and the precise movements of the human hand are unique, generated by what probably is the most elegantly skillful biological organ ever evolved. All Primates except man rely on their forelimbs and hands as major organs of locomotion. Man alone has freed his hands for manipulation by specializing for erect posture and bipedal locomotion.

Fossil evidence for the evolution of the hands of hominids other than man was meagre in the 1970s: a few australopithecine bones from Olduvai Gorge. The reconstructed hand appears to have more similarity to those of juvenile pongids and of adult *Homo sapiens* than to hands of adult great apes.

Man's opposable thumb

The most characteristic feature of the hand is a truly opposable thumb—that is, one that rotates at the carpometacarpal joint so that it opposes the other four fingers. A thumb that does not rotate (moving only in one plane at this joint) is called pseudo-opposable.

Facial structure. The hominid face is already distinct in the australopithecines. Subnasal prognathism (protrusion of the jaws) is reduced, and ultimately there is an early disappearance of the facial component of the premaxilla (bones of the upper jaw) through fusion. The contour of the forehead, the supraorbital tori (bony ridges above the eyes), the high position of the zygomatic arch of bone below the eyes, the orientation of the eye sockets, and the reduction of the temporal process on the

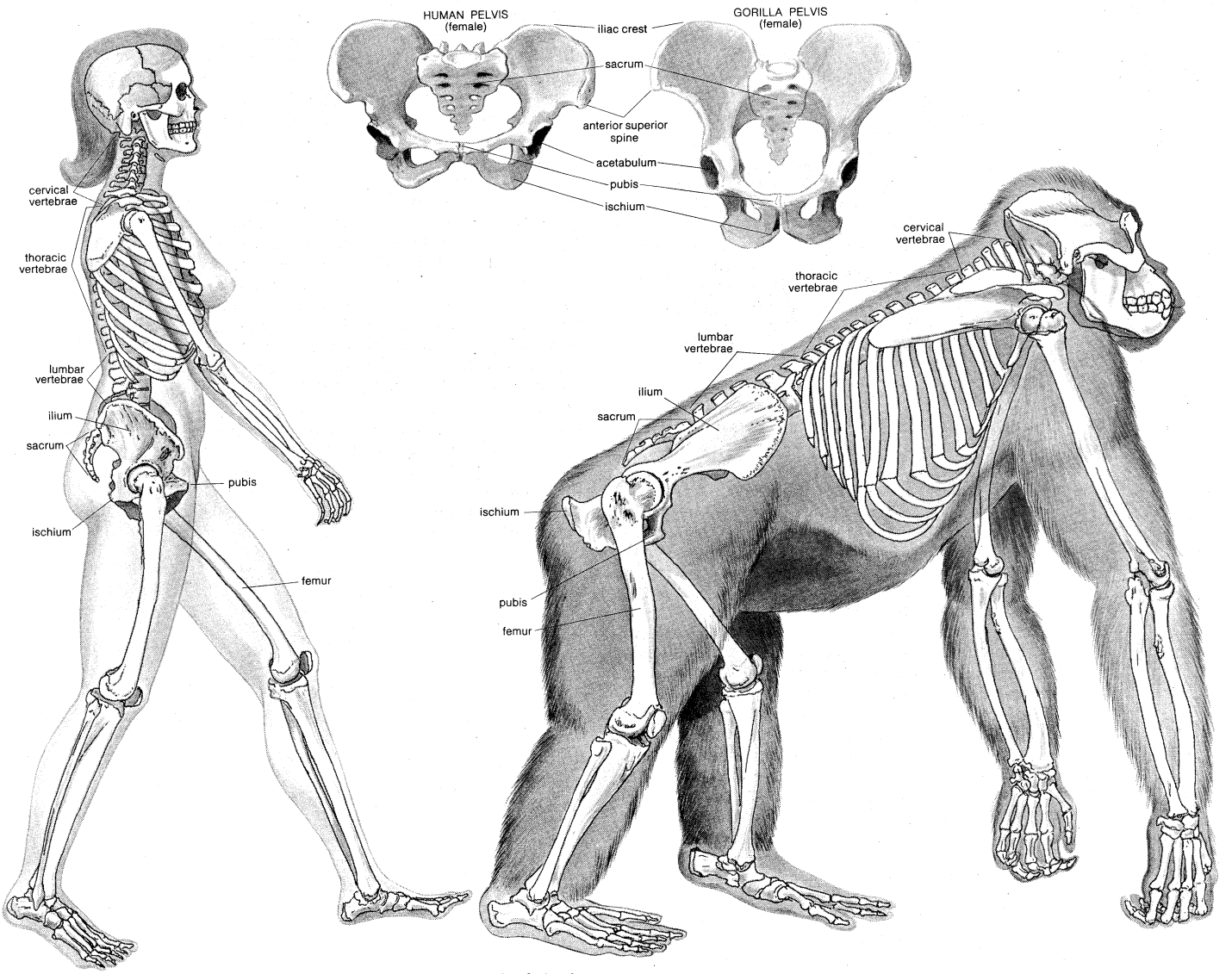


Figure 2: Locomotion of man and gorilla relative to skeletal structures.
From J. Buettner-Janusch, *The Origins of Man* (1966); John Wiley & Sons, Inc.

zygomatic bone comprise a total pattern quite distinct from that found in pongids of similar size (Figures 3, 4, and 5). This pattern is consistent with that found in the skulls of later Pleistocene hominids such as the Peking, Java, and Neanderthal fossils. The face of *Ramapithecus* has been reconstructed only from maxillary fragments,

From Sir Wilfrid E. Le Gros Clark, *Man-Apes or Ape-Men?: The Story of Discoveries in Africa*, Copyright © 1967 by Holt, Rinehart and Winston, Inc.; reprinted by permission

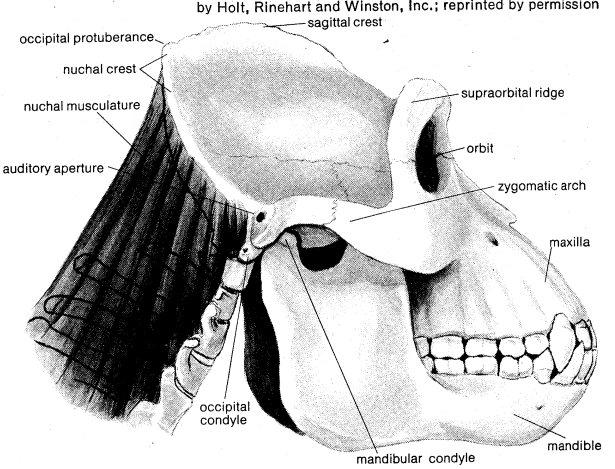


Figure 3: Massive neck muscles of gorilla, typical of pongids.

and the most that can be said with confidence is that *Ramapithecus* had a short face and an arcuate (arched) palate.

The mandible (lower jaw) has undergone major modifications since the Hominidae became distinct, being massive and large among early hominids. The association of large mandible, small braincase, and relatively high face once led many to conclude that both *Ramapithecus* and *Australopithecus* had clear pongid affinities. Among australopithecines, marked prognathism (although reduced compared to that of pongids) and the occurrence of a small sagittal crest atop some skulls reflect a large jaw powered by rather massive temporal muscles. Morphological differences between pongid sagittal crests and those found on some fossil hominid skulls are clear. The latter are the evolutionary consequence of upward displacement on the skull of bony supports for the temporal muscles. These supports developed during growth, occurring only on some skulls.

Dentition and diet. Hominid dentition is evident in both *Ramapithecus* and *Australopithecus*, especially in the reduction of the anterior teeth (canines and incisors) believed to signal a major change in diet and in behaviour (Figure 6). The massive anterior dentition of the Pongidae is most suitable for tearing and breaking up such vegetable foods as wild celery and bamboo, and their great canine teeth are socially important in displays of emotional states. As inferred from limited knowl-

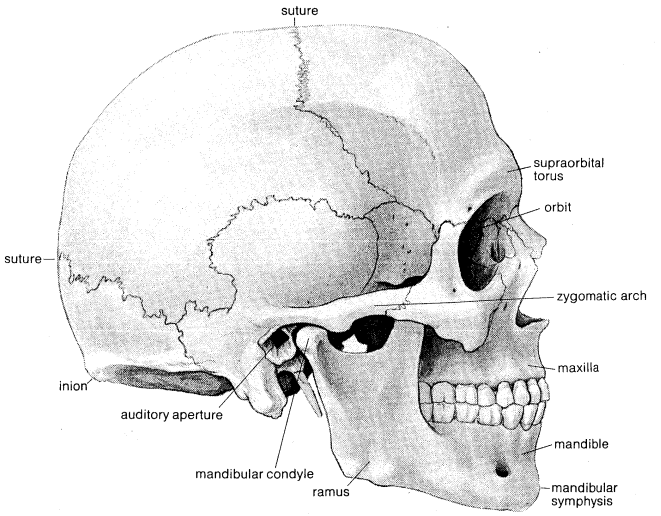


Figure 4: Right side of hominid (human) skull.
From J. Buettner-Janusch, *The Origins of Man* (1966); John Wiley & Sons, Inc.

edge, particularly of patterns of interstitial wear on the teeth, the principal hominid diet was not tough, succulent, green vegetation; meat may have been eaten, but it is believed that the essential shift was to consuming small, hard objects such as seeds.

Brain and nervous system. While the hominid brain is known directly only from *Homo sapiens*, deductions may be made from Pleistocene (*e.g.*, australopithecine) cranial fragments and intact skulls, and from Peking, Java, and Neanderthal remains. The brain increased in volume, and ratio of brain weight to body weight probably grew larger. The cerebral cortex expanded markedly and evolved with modern man to a large, highly complex structure for integration and control. Almost nothing is known about other parts of the nervous system among extinct hominid genera, although unreliable speculations based on reconstructions from fossil remains often are made. The same may be said for evidence of reproduction and growth based on the fossil record, almost all information being derived from man and contemporary pongids (see MAN, EVOLUTION OF).

Biomolecular characteristics. Chromosomes, hemoglobins, blood groups, many serum proteins, and red-cell enzymes (among other genetically controlled traits) have been studied extensively in contemporary *Homo sapiens*.

From J. Buettner-Janusch, *The Origins of Man* (1966); John Wiley & Sons, Inc.

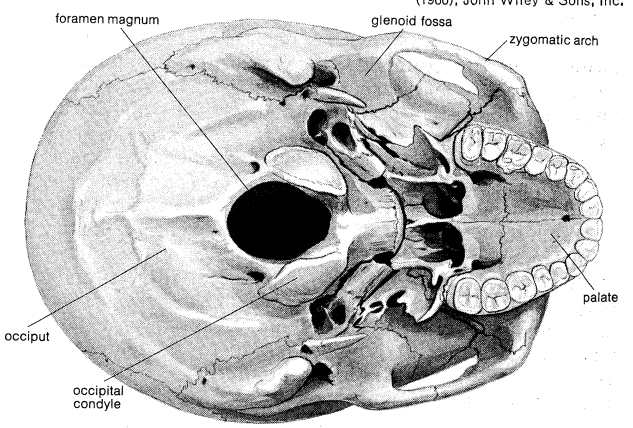


Figure 5: Base of hominid skull.

No reliable information on homologous traits of any extinct hominid exists, since chromosomes and proteins do not fossilize. Data available for man and living pongids, however, are compatible with modern notions of the affinities between hominids and pongids. Precise immunological and biochemical comparisons, representing efforts to refine views of the phylogenetic relationship of pongids and hominids, have been inconclusive at best. Use of these data to specify the time of divergence of the Hominidae and Pongidae seems premature, since results thus far are inconsistent with geochronological and chronometric estimates of the ages of hominid and pongid fossils.

BEHAVIORAL CHARACTERISTICS

Mentality is a zoological characteristic of major interest in any assessment of the Hominidae. Since fossils are not capable of such behaviour, however, most of the relevant evidence stems from *Homo sapiens*. The most salient characteristic of hominids is their extraordinarily different path of neurological evolution as contrasted with other families of primates. The distinct cognitive feature of hominids (especially man) is the ability to use symbols —to give arbitrary meaning to material objects. This feature is well illustrated by White's classic example that no chimpanzee can ever distinguish holy water from drinking water but that any man can be taught the difference. An eloquent account of the act of symbolling is found in the autobiography of Helen Keller, describing the moment when her teacher finally showed her that the word water meant the liquid she hitherto knew only by touch and taste. This is a superb, if metaphoric, indication of what the process of becoming a human hominid must have involved.

Symbolic behaviour

From J. Buettner-Janusch, *The Origins of Man* (1966); John Wiley & Sons, Inc.

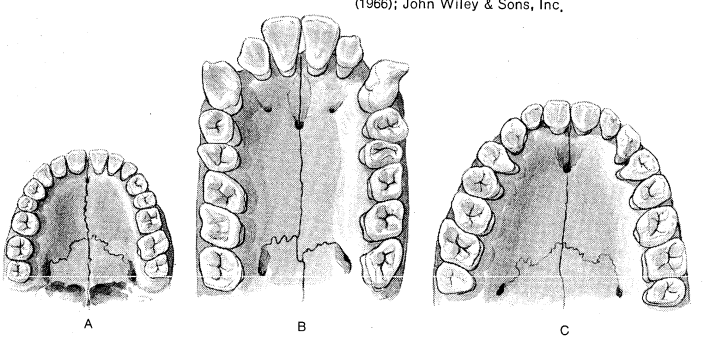


Figure 6: *Upper jaws of various primates.*
(A) *Homo*, contemporary man, family Hominidae. (B) *Pan*, living chimpanzee, family Pongidae. (C) *Ramapithecus*, restoration of a Miocene fossil form; note that *Ramapithecus* jaw is more nearly similar to that of man than is the jaw of the chimpanzee.

Toolmaking capabilities. In their difficult attempt to discover how and when the distinctive attributes of the brain of man became manifest in evolutionary history, most scholars look to the hominid fossil record and its associated archaeological sequences. Tools are obvious remnants, and signs of their planned, systematic production give evidence of symbolic communication among individuals and generations.

The use of tools—that is, employment of physical objects to perform certain tasks—is not unknown among so-called lower animals. Sea otter can be said to use rocks as tools to smash open oysters or clams against their chests as they float on their backs. Such investigators as W. Köhler (1925) and Jane van Lawick-Goodall (1964) showed that chimpanzees use objects as tools; other Primates, notably baboons, have been reported to behave similarly. Chimpanzees will lick a bit of grass or a stick to poke into termite nests, eating the insects that cling to the moistened surface. Vultures are known to crack ostrich eggs with rocks to feast on the contents. While such acts may be called tool *use*, these special and isolated manipulations are not equivalent to hominid tool*making*. Clearly, it would be most speculative to say that these animals think about making tools in a hominid manner homologous to man. Since firm evidence of traditions or styles of toolmaking has not been found among such animals, it would appear that only a rudimentary ability to explore the environment with tools is derived from the general primate heritage of the Hominidae.

It is unlikely that the first appearance of toolmaking ability will ever be precisely established. Any hominid fossil found with recognizable tools clearly comes after the cultural stage at which the first tools were made. These are most apt to have been bits of stick or other perishable material that no longer survive.

Brain evolution and culture

Specifically human behaviour evolved from an ability to manipulate symbols, a product of neuroanatomical change. There is ample evidence that the cerebral cortex increased in size during primate evolution. Major trends were the elaboration of the special senses and the expansion of cortical centres for conscious control over complex behaviour and muscles. Studies of comparative neuroanatomy suggest that cortical association areas that process complex sensory data and that mediate voluntary actions have increased in size. Although only the crudest estimate of changes in the cortex has resulted from study of fossil hominids, the neurological basis for culture, as in toolmaking, seems to be found in the evolution of the cortex.

Cranial volume does not have major significance as a measure of the degree to which fossil hominids approached the human condition. Even the ratio of brain weight to body weight is suggestive rather than definitive. It is not the amount of neural tissue in the cranium but the way it is differentiated and organized that is critical. The fossil record yields little beyond total volume, a crude estimate of the weight ratio, and (from an occasional endocranial cast) a few hints about the gross morphology of the brain surface. It seems reasonable, however, that by the time the potential for toolmaking developed, the size of the brain was not critical.

Acceptable evidence that a particular species made tools implies that symbolic communication and education were in the repertory of the hominid (whose fossil bones are found with the tools), whatever the brain size. If an australopithecine with about 600 cubic centimetres of brain volume is found with tools, this signifies a hominid brain large enough for symbolic communication and toolmaking. Tools indeed have been reported with australopithecine fossils, and it is therefore presumed that toolmaking was part of the behavioral repertory of the genus *Australopithecus*. There is little doubt that hominid fossils from the Middle and Late Pleistocene are those of toolmakers. It is a question of greatest interest whether sufficient postcranial material and tools of *Ramapithecus* ever will be found together *in situ*.

Language and related symbolic behaviour. The origins of symbolic vocal communication (oral language) in man probably will not be found in the fossil record despite the evidence of symbolic communication (making tools to a plan, occurrence of tool styles and traditions). Vocal and facial displays of other living primates do offer limited sources of information for an inquiry into the origins of spoken language. Man's imitation of what is heard, including the sounds of other animals, must be one root of origin. Vocal language must have arisen as a modification of already existing vocal systems. It can reasonably be postulated that language arises in a species in which auditory control over vocalization is sufficiently developed to permit individuals to imitate each others' sounds.

Vocalization

Control of vocalization in man largely involves controlling changes in resonance properties of the buccal (mouth) cavity. Humanoid grunts take advantage of all the resonant properties of the mouth and upper respiratory tract. Ability to contract the orbicularis oris (the circular muscle of the lips) would have developed through natural selection to produce the vocal range and variability available to modern man. Any change that makes transmission of information by vocal or facial expressions more explicit and less ambiguous also enhances social interaction, and selection would favour a coherent and well-organized society. The grunts of *Homo sapiens* may have evolved in the ancestral hominids by conveying vocal information simultaneously with facial expressions, just as the grunts of baboons seem to have evolved.

While human vocal cords are more blunt on the edges than those of baboons or apes, the principal difference in vocal apparatus is the position of the larynx, relative to the rest of the respiratory tract. Man's larynx is lower in the throat and farther from the soft palate than in other primates, directly reflecting development of erect posture and expansion of the braincase. The position of the larynx changed as the foramen magnum (the skull's aperture for the spinal cord) moved forward and as the mandible grew smaller. The foramen magnum, in turn, moved in response to the way the head is balanced on the neck and to the expansion of the posterior portion of the base of the skull. The consequent descent of the larynx created a long, tubular resonating cavity that permits the low-pitched speech of man. Auditory discrimination in man probably evolved along with the vocal apparatus, as it is most acute in discriminating pitch within the normal human range. Symbolic meaning, however crude, would not be attached to particular sounds if one animal could not imitate vocalizations of others. The development of vocal mimicry must have been important in the origin of language and may well have depended in part on the ability to manipulate the behaviour of others through vocalization. Chimpanzees give particular calls that direct the attention of others to specific objects. Although the chimpanzee larynx allows rich, variable vocal function, its structure prevents reproduction of many human phonetic elements. Although no fossil larynges have been found, the morphology of the base of the fossil cranium suggests that the human condition was developing in the earliest hominids.

Facial displays also are related to the evolution of vocal language, since changes in such facial muscles as the orbicularis oris and zygomaticus (that control lip movements) were of critical importance (Figure 7). Primate

From J. Buettner-Janusch, *The Origins of Man* (1966); John Wiley & Sons, Inc.

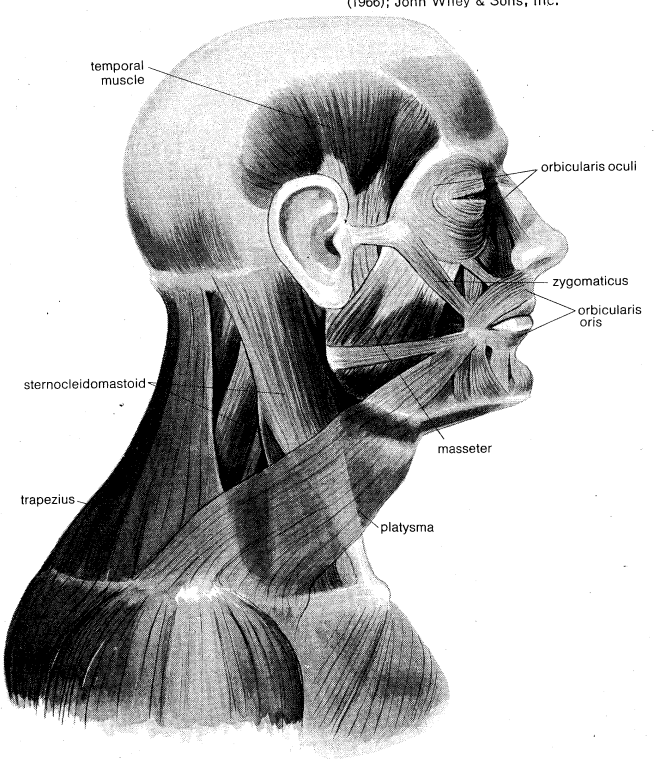


Figure 7: Muscles of human head and neck.

facial displays have evolved to the point that they can occur without vocalization; among chimpanzees and *Homo sapiens* the sight of a desired object may elicit a silent smile.

Social organization. Social behaviour or social groupings, of course, do not fossilize. Speculations and rather tenuous deductions in reconstructing the social life of early hominids are based on fossil morphology, on living nonhuman primates (*e.g.*, baboon troops and anarchic chimpanzee bands), and on contemporary *Homo sapiens*.

Small, erect, bipedal hominids, with dentition no longer suitable for tough, succulent stems and leaves ripped in foraging the dense humid forest, probably developed a mode of life suitable to hunting and to gathering grains, seeds, and fruits. The socially organized activity that this shift may be said to imply required differentiation of the roles of males and females. Relatively stable integrated family units very likely arose and, with cooperative hunting and gathering, further division of roles. Though it is fairly clear that these social developments occurred in populations that eventually generated human hominids, that australopithecines were so organized is uncertain.

CONTRASTING ADAPTATIONS OF HOMINIDAE AND PONGIDAE

Hominidae are distinguished from Pongidae (anthropoid apes) by evolutionary trends that illustrate the adaptations of each for different environmental situations. Obvious morphological contrasts in skeleton and musculature between living pongids and hominids suggest the functional significance of homologous features observed in fossil representatives of the two families. The contrasts are particularly evident in progressive skeletal modifications in adapting to erect bipedalism. Most significant in hominids are a proportionate lengthening of the lower extremity and changes in the proportions and morphological details of the pelvis, femur, and pedal skeleton—all related to the mechanical requirements of erect posture and bipedal gait. Associated soft tissues also are modified, most particularly the hamstring tendons, gluteus muscles, and the complex (iliopsoas muscle) that flexes the trunk and thigh. The hominids preserve a well-developed pollex (thumb) and have lost opposability of the hallux (big toe).

Locomotion and posture

Pongid skeletal modifications sometimes are interpreted as deriving from a rather small brachiating (arm-swinging) ape (presumably one of the dryopithecine fossils of the Miocene Epoch). The proportionate segmental and overall lengthening of the upper extremity in pongids is believed to support this view. The pongid pollex is reduced, there is a strong and opposable hallux, and details of the limb bones are modified. The forelimb and hand of some pongid genera seem to be adaptations to knuckle or fist walking. Increase in size and mass of some individuals in the pongid lineage (including the large, extinct genus *Gigantopithecus*) probably made a shift from arboreal or terrestrial quadrupedal locomotion to the knuckle-walking gait a distinct advantage.

The pongid pelvis retains the main proportions characteristic of quadrupedal mammals. Although to some the anatomy of the pectoral girdle of pongids suggests a brachiating ancestor, all Miocene Hominoidea (Hominidae and Pongidae) recovered and described by the 1970s fail to show specific adaptations for brachiation. At least, the Miocene hominoids do not resemble modern gibbons (*Hylobates*), whose arboreal acrobatics led to the definition of brachiation. Analyses of both living pongids and extinct forms suggest that the basic adaptation is dominated by forelimb locomotor and feeding behaviour, knuckle walking distinguishing pongids from their Oligocene or early Miocene ancestors.

The hominid skull bears marks of a way of life that contrasts strongly with that of pongids. In fossil skulls of australopithecines the occipital condyles (protuberances on the base of the skull that articulate with the vertebral column) are more anterior than in pongids. Their position is associated with increased flexure of the basicranial axis, leading to upward displacement of the braincase relative to the face, with resultant increase in cranial height. These changes often are cited as serving to maintain the balance of the head, consequent to development of erect posture and bipedal gait. They are also a result of the expansion of the braincase.

The pongid skull is that of a quadrupedal, orthograde primate. There is marked prognathism and, in larger species, massive jaws associated with strong muscular ridges on the skull. The extensive nuchal (nape) area of the occiput, the relatively high inion (occipital protuberance), the position of the occipital condyles well behind the level of the auditory apertures, and the limited degree of flexion of the basicranial axis are all features consistent with knuckle or fist walking. These characteristics of the skull are found in Miocene and Pliocene Pongidae as well.

That hominid dentition differs significantly from that of the pongids is now viewed as an adaptation to a diet that varies markedly from that of the pongids. Significant hominid features include reduction of the incisors and canines, appearance of bicuspid premolars, and changes in the occlusal relationship of the jaws. The canines have diminished to a spatulate form and interlock slightly or not at all. There is no pronounced sexual dimorphism of canines, and the spaces between these teeth (diastemata) largely have vanished. First premolars adapted for cutting are replaced by bicuspids, with secondary reduction of the inner (lingual) cusp in later hominid forms. Occlusal alterations tend to promote wear in all teeth to a flat, even surface at an early stage of attrition. The dental arcade is even and rounded, and there is a marked tendency in later stages of the fossil record (as in modern man) toward a reduction in molar size. Deciduous teeth appear to be replaced earlier (relative to eruption of permanent molars) and there is progressive molarization of the first deciduous premolar. This is the dentition of a plains-living animal that had forelimbs not primarily used in locomotion and that hunted and gathered seeds, fruits, and grasses.

Differences in dentition

Pongid dentition is clearly established in the fossil record of the Miocene Epoch and probably originated earlier. There appears to be progressive increase in the size of the incisors and widening of the symphysial region of the mandible with eventual formation of the simian shelf (a distinctive bony part of the anthropoid ape mandible). Strong, conical, sexually dimorphic canines that interlock in diastemata are found throughout the pongids (except for gibbons). The cutting function of the first lower premolar is accentuated with the development of a strong anterior root. Postcanine teeth preserve parallel or slightly divergent alignment in relatively straight rows, as opposed to the rounded hominid dental arcade. First deciduous molars remain predominantly unicuspid, and there is no apparent acceleration in the eruption of permanent canines. Pongid dentition belongs to an animal that feeds on large stalks of vegetation (tearing and rending them with large molar teeth), using enormous canines to lever, chisel, and support.

Skeletal (especially cranial) comparison of living and fossil Hominidae and Pongidae once led to widely varying conclusions. Relatively few cranial characteristics provide evidence for obviously different evolutionary directions for the two families. For example, the cranial capacity of *Australopithecus* is relatively small, ranging from 450 cubic centimetres to more than 600 cubic centimetres. There are strongly built supraorbital tori, and among larger individuals a low sagittal crest occurs at the vertex of the skull (frontoparietal). When hominid adaptation was better understood, these features no longer were held to indicate pongid affinities. Yet it is understandable that early in the 20th-century authorities were reluctant to classify australopithecines as Hominidae. As long as notions of what constitutes an early hominid prompt a search for a tiny version of *Homo sapiens*, most fossil hominids will be rejected. Nevertheless, hominid adaptations are manifest in *Australopithecus*. Dental characteristics that are hominid as opposed to pongid also occur in *Ramapithecus*. Clearly, elements of the adaptive complex that is called hominid were present in the late Miocene, and perhaps earlier.

Cranial distinctions

ECOLOGICAL ASPECTS OF HOMINID ORIGINS

Based on sparse evidence, discussions of the ecology of hominid origins are fraught with controversy and confusion. Clearly different from all living pongids, the only modern hominid is terrestrial. No living primate seems to provide an obvious model for reconstructing the ecological situation of the immediate ancestor of the hominids. One popular view is that a hominoid arboreal bra-

chiator (from which it is believed modern pongids also developed) was a transitional stage between prehominids and hominids. Another view is that hominid progenitors were terrestrial, with a highly social way of life resembling that of modern baboons. Others believe that the prehominid hominoid was arboreal, clinging to and feeding from small branches at the ends of limbs of tall trees in the forests. It also has been suggested that a semiterrestrial, semibrachiator preceded hominids. This typological view, despite temporary popularity, is based on dubious assumptions.

How is a decision to be made among alternatives of this sort? Each assumes a specific environment, a distinctive ecological system. Anatomic data (the best evidence available) suggest to some investigators that muscle and bone arrangements in the shoulder of *Homo* are a recent derivation from a lineage that brachiated. Other equally respected authorities interpret most of the anatomy as derived from the basic structure of the orthograde trunk of all primates, rather than from an anatomy specifically adapted for brachiating. Man's hand resembles that of terrestrial Primates, and some take this as supporting the view that the immediate ancestor of the Hominidae must have been terrestrial. Others point to startling resemblances between the hands of gibbons (true brachiators) and those of men. The notion that progenitors of the Hominidae occupied the small branches of trees is based on analysis of rather specialized modern primates. The few available postcranial bones of early Miocene Hominoidea once were believed to belong to brachiating primates. This view, however, is untenable, since all known Miocene hominoids are quadrupedal and are not brachiators in the sense that gibbons are. An arboreal brachiator probably should not be postulated as either the structural or the ecological progenitor of the hominids.

The inert fossil record can show no direct evidence for determining most of the structural and behavioral adaptations of the prehominids and earliest hominids. But it is possible to speculate about the sort of animal implied by early Miocene fossil fragments attributed to the Hominoidea.

The Miocene hominid *Ramapithecus* unfortunately is known only through maxillae and mandibles. Although the dentition in many ways is more like that of *Homo* than that of *Australopithecus*, firm statements about the mode of locomotion and general ecological situation of this beast are not yet supportable. Nonetheless, the small anterior dentition is probably good evidence that hominid dietary adaptation had already been achieved. If this is so, then it is not unreasonable to suggest that *Ramapithecus* was structurally adapted for hominid locomotion and for the hominid segment of the habitat. The hypothetical population from which *Ramapithecus* diverged was probably not specifically adapted to a brachiator's environment, but equally at home in trees or on the ground. Besides having a dentition approximating that of *Ramapithecus*, it must have had hands that were relatively free of locomotor tasks, and its pelvic girdle must have been less specialized than that of any pongid.

To speculate further: suppose that an early Miocene or late Oligocene hominoid (medium sized, smaller and slighter than modern chimpanzees, with smaller front teeth and defter hands than most other anthropoids) is faced with major environmental change. There is some evidence that (in Africa, at least) the distribution of great forests grew restricted and savanna grasslands expanded. The first of the prehominid hominoids might well have been able to respond to a changing environment by changing posture and locomotion. It seems reasonable that ecological conditions provided the selective advantage for the shift from orthograde quadrupedalism to erect bipedalism. In such a new adaptive zone, hunting and the gathering of small, hard food objects (seeds for instance) would become essential. It would include a change in ecology from thick forest cover to open woodlands, bush, and savanna. Clearly, a change in primate social organization is indicated: from random food gathering, like that of modern pongids, to a pattern of organized hunting and gathering. Indeed, almost all contemporary human groups that live by hunting are basically organized to subsist on vegetable material gathered by specifically assigned members.

The development of such cooperative groups that could exploit the savanna and grasslands very likely provided the selective pressure that emphasized erect posture, bipedal gait, skillful use of the hands, symbolic behaviour, and humanoid social structure. Erect bipeds with small anterior dentitions and talented hands would not have been able to create cooperative hunting and gathering groups without modifying their social organization. The rather loose social groupings of chimpanzees or the somewhat more organized baboon troops suggest what early prehominid hominoids may have been like socially and that a change in social organization was required. Organized foraging as a characteristic of a species imposes some differentiation in the social roles of those who care for infants and those who hunt. Cooperation among hunters develops, and family groups tend to become permanent and well integrated. Awareness must develop that allows for planning and discussion; the prey to be hunted, the tools needed for future hunting, and the probable harvest of wild grains and fruits. (Baboons and chimpanzees are justly characterized by the phrase, "out of sight, out of mind.") A lineage of primates in which these capacities emerged would be under strong pressure to continue to develop along such lines in an environment that presumably was changing rapidly from forest to bush and savanna. The transition from hominoid to early hominid to human involved reorganization of hand, pelvis, and foot, as well as brain. The unique social life of *Homo sapiens* seems to have its roots in a shift made by a prehominid hominoid from random food gathering to organized hunting and gathering on the African and Asian plains.

The probable origins and evolution of the Hominidae are better known than they once were. The number and variety of fossils available for study have increased; more importantly, preconceptions that hindered earlier investigations are dissipating. Major problems that remain include the need for finding more fossils and for refining the analyses of the anatomical, ecological, and, even, behavioral complexes these fossils imply.

BIBLIOGRAPHY. W.E. LEGROS CLARK, *The Antecedents of Man*, 2nd ed. (1962), *The Fossil Evidence for Human Evolution*, 2nd ed. (1964); and J. BUETTNER-JANUSCH, *Origins of Man* (1966), are the most comprehensive general books. W.K. GREGORY, *The Origin and Evolution of the Human Dentition* (1922), demonstrates a very progressive view, although it was published years ago. G.G. SIMPSON, "The Meaning of Taxonomic Statements," in S.L. WASHBURN (ed.), *Classification and Human Evolution* pp. 1–31 (1963), is an excellent account of systematics and taxonomy. Accounts of the origin of the Hominidae are found in D.R. PILBEAM, *Tertiary Pongidae of East Africa: Evolutionary Relationships and Taxonomy* (1968), and with E.L. SIMONS, "Preliminary Revision of the Dryopithecinae (Pongidae, Anthropoidea)," *Folia Primat.*, 3:81–152 (1965); and also in E.L. SIMONS, "On the Mandible of *Ramapithecus*," *Proc. Natn. Acad. Sci. U.S.A.*, 51:528–536 (1964). A.H. SCHULTZ, *The Life of Primates* (1969), is a general view of living primates that stresses similarities and differences among nonhuman primates and man. Excellent papers about specific problems in the origin of the Hominidae are R.J. ANDREW, "The Origin and Evolution of the Calls and Facial Expressions of the Primates," *Behaviour*, 20:1–109 (1963); C.J. JOLLY, "The Seed-Eaters: A New Model of Hominid Differentiation Based on a Baboon Analogy," *Man*, 5:1–9 (1970); G. KELEMEN, "Anatomy of the Larynx and the Anatomical Basis of Vocal Performance," in *The Chimpanzee*, vol. 1 (1969); J.R. NAPIER, "Studies of the Hands of Living Primates," *Proc. Zool. Soc. Lond.*, 134: 647–657 (1960); W.L. STRAUS, "Fossil Evidence of the Evolution of Erect Bipedal Posture," *Clin. Orthop.*, 25:9–19 (1962); and R.H. TUTTLE, "Knuckle-Walking and the Evolution of Hominoid Hands," *Am. J. Phys. Anthrop.*, 26:171–206 (1967). A rather different view of some of the processes involved is found in J.T. ROBINSON, "The Australopithecines and Their Bearing on the Origin of Man and of Stone Tool-Making," *S. Afr. J. Sci.*, 57:3–13 (1961); and in P.V. TOBIAS, *Olduvai Gorge* (1967).

(J.B.-J.)

Homo Erectus

One of the extinct species belonging to the human family (see HOMINIDAE) is *Homo erectus* (formerly called *Pithecanthropus*). Members of this species differ in a number of respects from modern man, who is classified as a distinct primate species called *Homo sapiens*. The history of *Homo sapiens* may go back for 100,000 years, or perhaps even 200,000 or 300,000 years. Earlier than that, the fossil record tells of the existence of another kind of man, classified as *Homo erectus*. His remains have been found over large parts of the Old World—Africa, Asia, Europe—and he may well have inhabited these parts of the globe for as long as 500,000 to 1,000,000 years. His "moment" in geological time lies in the Pleistocene Epoch (2,500,000–10,000 years ago) and more particularly, in a rather vaguely defined subdivision, the middle Pleistocene.

Both in space and time, *Homo erectus* is well placed to have been the direct ancestor of *Homo sapiens*. Herein lies his special importance in the general theory of human evolution, for he seems undoubtedly the strongest claimant to fulfill the position of modern man's immediate predecessor. This reading of the evidence is widely accepted today; nevertheless, there are differences of opinion on such details as where and how *Homo erectus* evolved into *Homo sapiens*. That a development from the earlier *erectus* species into the later *sapiens* species did take place, however, is the consensus among most students of human evolution. Some of the possible pathways by which the evolutionary transition from *H. erectus* to *H. sapiens* may have occurred will be examined here.

Homo erectus is not, by any means, the earliest representative of the hominids (*i.e.*, the family of man). More archaic hominids are known from the early part of the middle Pleistocene, from the preceding lower Pleistocene, and even from the Pliocene Epoch (perhaps 4,000,000 years ago), which is the second last epoch of the Cenozoic Era, last of the great divisions of time in the earth's geologic record. This earlier phase of emergent humanity consists of a diverse array of upright-walking (though small-brained) creatures ranging from very large toothed, heavy-muscled, robust-boned forms to lighter, smaller toothed, more petite hominids. The most important of this group of fossil progenitors is called *Australopithecus*; and in body structure and function, as well as in their occurrence in time and space, some populations of *Australopithecus* (*q.v.*) fulfill the requirements for the ancestry of later forms of man.

In other words, the fossil evidence available today strongly suggests that one or another branch of *Australopithecus* evolved further, eventually giving rise to *Homo erectus*. Again, there is no absolute unanimity; a few dissident voices have placed a somewhat different interpretation on the sequence of events to be inferred from the fossils. These alternative interpretations are also reviewed, and the most probable reconstruction of events is considered.

If one accepts the majority view for the moment, *Homo erectus* may be interpreted as a transitional form of mankind, poised briefly (as geological eons go) between apeman, on the one hand, and modern man on the other. According to this view, in *Homo erectus* was wrought the revolution that converted an intelligent animal into a thinking, forward-looking, planning, imagining, and speaking man.

Ever since the time of the great Swedish naturalist Carolus Linnaeus (or Carl von Linné), it has been the practice to consider the world of living things as divided into species; several closely related species, in turn, are grouped into a genus (plural genera). To name a species, Linnaeus (1759) proposed that the genus and species be designated—like a surname and forename—in that order. The domestic dog, for instance, is called *Canis familiaris;* this refers to the *familiaris* species of the genus *Canis*. There are other related species in this genus, like *Canis lupus*, the wolf, and *Canis aureus*, the jackal. Modern man is called *Homo sapiens;* he belongs to the *sapiens* species of the genus *Homo*. The fossil hominid species that is of concern here is *Homo erectus*, another species of the genus *Homo*.

For more than three-quarters of a century, fossils have been coming to light showing more or less of the anatomical features of *Homo erectus*. Traditional in this field of study is the tendency for investigators to endow newly discovered fossils with newly invented names. In this way, the discoverers of newly found fossil specimens often have given the impression that their finds are completely different from any other specimens previously known. Not only have species names for hominids multiplied apace but even the proposed genera, within which the species are grouped, have presented a bewildering array—to confound the unfortunate student and to obscure the broader evolutionary trends. Thus, fossils that today are recognized by many workers as belonging to *Homo erectus*, at one time or another have been given such generic (genus) names as *Pithecanthropus, Sinanthropus, Atlanthropus*, and *Telanthropus*.

With the passage of time, paleoanthropologists (people who study the fossil history of man) have learned increasingly about the variability of living and fossil species. Minor anatomical differences—which once were taken as the definitive criteria for a new species or even a new genus—have come to be appreciated as individual or, at most, regional variants within the same species. Gradually, as the number of available samples of hominid fossils has grown with new discoveries, emphasis has been transferred from minor aspects of anatomical dissimilarity to the major points the ancient bones have in common. It has come to be realized that a whole group of hominid fossils has more in common with one another than any of them has with *Homo sapiens* or with *Australopithecus*. If these early fossil men were even nearly as variable as are modern species, then it seems to most authorities that they must all have belonged to a single species (now called *Homo erectus*).

It may be tempting to be snobbish and cynical about the name-making and species-inventing propensities of earlier workers; but it should not be forgotten that it was only the growth of understanding about variability in modern species, as well as the accumulation of additional fossils, that led to the more recent tendency to lump many different fossils into a single species bearing a single name. The changeover in approach was a consequence of history, rather than the exclusive result of any heightened understanding of biology among today's anthropologists.

HISTORY OF DISCOVERIES OF HOMO ERECTUS

Of the many names earlier given to representatives of *Homo erectus*, one is of more than passing interest: *Pithecanthropus*. Indeed, this name had already been coined before the discovery of the fossil to which it came to be applied. The name was invented by the German zoologist, Ernst Heinrich Haeckel, in the 19th century. He was one of those biologists who enthusiastically adopted Charles Darwin's theory of natural selection as the cardinal mechanism of evolutionary change. Haeckel postulated the former existence of a yet-to-be-discovered ancestral hominid, a kind of missing link that would be found to bridge the gap in the evolutionary chain between man and ape; and he called this creature of his scientific imagination *Pithecanthropus*.

The missing link

Years later, in 1891, Dutch army surgeon Eugène Dubois discovered the vault (braincase) of a skull and a few other fossilized bones at a place called Trinil in central Java. Considering its low-set braincase, retreating forehead, prominent eyebrow ridges, and a marked constriction of the skull behind the eye sockets, Dubois concluded that this cranium showed anatomical characteristics intermediate between those of men, as then understood, and those of apes; he revived Haeckel's old term and named his discovery *Pithecanthropus* (literally, apeman). Near the cranium he found a distinctly hominid thighbone (femur). Since this long, straight bone was so much like a modern human femur, Dubois decided that its owner must have walked erect. He assumed that the femur had belonged to the same individual, or at least the

same kind of hominid as did the skull—and so he named the species represented by the little cache of remains *Pithecanthropus erectus.*

The scholarly world's cynical reaction to the first claims about *Pithecanthropus erectus* eventually changed; and so did Dubois's own thinking—for at the time he died, many years later, he was much more inclined to stress the apelike features and to play down the hominid affinities of his discovery.

Distribution of Homo erectus. Subsequent discoveries gradually established the case for a new and separate species of fossil man. At first, these discoveries were centred largely in Asia. At several different places in Java (now part of Indonesia), essentially similar fossils were found: the sites are Trinil, Kedung Brubus, Modjokerto, and Sangiran. Another series of finds was made in China, especially in a famous cave called Chou K'ou-Tien near Peking (Peiping), although virtually all the remains from there were subsequently lost during the Sino-Japanese War, about 1939–40. Newer discoveries have since been made in this cave, while two new Chinese sites, in the Lan-T'ien district of Shensi Province, have yielded remains attributable to *H. erectus.*

Other finds in Asia and Africa

By the end of World War II, the freakish pattern of early discovery had given rise to an idea that *H. erectus* was a peculiarly Asian expression of early mankind. Then new discoveries in Africa served to change this view, and it came to be realized that Europe, too, may have harboured *H. erectus.*

In Africa, excavations at a place called Ternifine, east of Mascara, Algeria, in 1954–55, yielded remains whose nearest affinities seemed to be with the Chinese form of *H. erectus.* Other hominid fragments from northwest Africa—parts of a skull found in 1933 near Rabat in Morocco and jaws and teeth from Sīdī ʿAbd ar-Raḥmān (1954) in Morocco—show features reminiscent of *H. erectus*, though they are rather more advanced in structure than are those of Ternifine and of Asia. Another fossil probably related to *H. erectus* is a cranium found in 1960 at Koro Toro, Chad. Unfortunately, this cranium is fractured, distorted, and very worn. The most complete and convincing evidence for the existence of *H. erectus* in Africa came with the discovery in 1961 of a characteristic *H. erectus* braincase in Olduvai Gorge in Tanzania (so-called Olduvai hominid 9).

A further African claimant to membership in the species *H. erectus* is represented by skull fragments from a limestone cave deposit at Swartkrans in the Transvaal, South Africa. These fragments, found alongside remains of a robust *Australopithecus* population, were at first called *Telanthropus* but have since been assigned by many authorities to *Homo erectus.* Their incompleteness makes positive identification difficult, however.

The realization that Africa as well as Asia was apparently peopled by a form of mankind classifiable as *H. erectus* led to a re-examination of some of the earliest hominid fossils from Europe. An isolated lower jawbone had been found in a sandpit at Mauer, close to Heidelberg, Germany, in 1907. Although it had been given a variety of names over the years, its exact affinities to other fossils remained uncertain, since no associated cranium was found. In recent years, a number of investigators have come to regard the Mauer jaw as representing a member of the species *Homo erectus.* Although its geological age would appear to be slightly greater than that of Trinil in Java, Chou K'ou-Tien in China, and Ternifine in Algeria, this skeletal fragment from Europe shows more modern structural features than do the Asian and African jaws of *H. erectus.* The exact significance of these features in the Mauer jaw is still being debated; they could be the mark of an individual variant, highlighting the fact that it is not yet known how variable in bony structure *H. erectus* was. Alternatively, the Mauer specimen could represent a subspecies or race of *H. erectus* that is slightly more advanced in anatomic structure than are the African and Asian populations; or the Mauer man could have been a member of a very early population of *Homo sapiens.* Evidence for the latter has been provided by a newer discovery made in Hungary.

Possible European specimens

In 1965, remains of two individuals—a child and an adult—were found in a travertine quarry at Vértesszőllős, about 50 kilometres (roughly 30 miles) west of Budapest in Hungary. The remains of the child are milk teeth, and enough of them is preserved to show affinities with the Chinese *H. erectus* of Chou K'ou-Tien. The adult is represented by a large part of an occipital bone (at the back of the head) from a large-brained skull. While showing some features reminiscent of *H. erectus*, the general form and the capacity of the Hungarian cranium also suggest an affinity with an early branch of *H. sapiens.* In fact, it appears to be related to somewhat later middle Pleistocene skulls (perhaps 200,000 years old) from Europe, such as those of Swanscombe and Steinheim, which are accepted as early members of *H. sapiens.* Since such twofold affinities are exhibited by the Vértesszőllős group of remains, authorities differ as to whether to call the population they represent *H. erectus* or *H. sapiens.* The same uncertainty applies to some of the fossils found in North Africa; the Sīdī ʿAbd ar-Raḥmān and Rabat remains are regarded by some experts as late surviving members of *H. erectus* and by others as forms transitional between *H. erectus* and *H. sapiens.*

The European fossils that are of comparable age to those of Ternifine, Chou K'ou-Tien, and Trinil (namely the jaw from Mauer and the isolated milk teeth and occipital bone from Vértesszőllős) appear to have more structural features in common with early *H. sapiens* than do their Asian or African contemporaries. Some investigators would say that these anatomical differences are negligible compared to inferences they make that the remains from Mauer and Vértesszőllős are contemporary with *H. erectus* elsewhere, and that they represent a man whose way of life or cultural status was equivalent to that of *H. erectus* elsewhere. Even if these two inferences were accepted as facts, however, the approach taken by these investigators would represent a departure from the way in which paleontologists usually decide whether two groups of specimens belong to the same or to separate species.

Uncertainties

At this stage in the understanding of the early middle Pleistocene populations of Europe, it would probably be safest to conclude that the case is not yet convincingly established for the existence in Europe of *H. erectus*—as known from Asia and Africa.

The antiquity of Homo erectus. To reconstruct the position of *H. erectus* in hominid evolution, it is essential to define his place in time as precisely as possible. Modern developments in such disciplines as physics have placed at the disposal of the paleoanthropologist a variety of techniques that permit increasingly accurate assessments of the absolute age of fossils. Many of these methods are based upon the effectively constant (or absolute) rate at which radioactive isotopes of such elements as potassium and argon decay (see DATING, RELATIVE AND ABSOLUTE). When the newer methods cannot be applied, investigators may still ascribe a relative age to a fossil. This can be done by noting the contents of the layer of rock or the deposit in which the fossil was found; a layer containing, for instance, evidence of the remains primarily of extinct animals is probably older than one containing signs of predominantly recent or still living forms.

Such lines of evidence have led to the tentative conclusion that the species *Homo erectus* is essentially of early middle Pleistocene age. The oldest appearing fossil examples unfortunately all lack absolute dates; but, by relative dating from associated remains of other animals, it would seem probable that the hominid remains from Lan-T'ien in China, Sangiran and Modjokerto in Java (more especially, fossils from older deposits called the Djetis beds), and those from Swartkrans in South Africa (if, indeed, these are correctly to be designated *H. erectus*) are the oldest representatives of *H. erectus* thus far discovered.

On the other hand, the youngest accepted hard-core representatives of *H. erectus* in the fossil record would seem to be the group from Peking in China, Trinil in Java, Ternifine in Algeria, and the braincase of Olduvai

hominid 9 from Tanzania. Repeated potassium-argon datings of the Trinil beds has yielded an estimate of their age in years as 550,000 BP (before present).

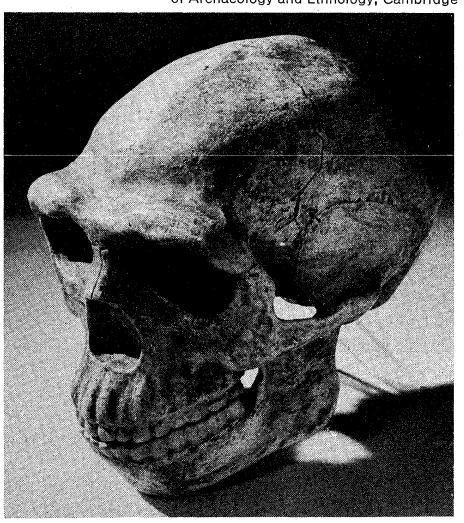
Until more complete evidence is available it would seem reasonable to suggest 1,500,000 to 500,000 BP as a time range for *Homo erectus*. Beyond this time range are a group of earlier fossils that some scientists have called *Homo habilis*, a few members of which are regarded by other investigators as representing an early expression of *H. erectus*. These fossils are best known from the Olduvai Gorge in Africa, and the oldest specimen from there seems to be about 1,800,000 years old. On the other hand, there is a group of later specimens that show some features of *H. erectus* but that are commonly regarded as transitional forms or as members of *H. sapiens;* these include later middle Pleistocene specimens from Europe (discovered at sites at Swanscombe, Steinheim, and Montmaurin) and from Africa (Sīdī ʿAbd ar-Raḥmān and Rabat). Still later forms suggest *H. erectus* biological ancestry, such as certain upper Pleistocene fossils from Asia (Solo man from Ngandong in Java) and from Africa (sites called Tēmara, Tangier, Lake Eyasi, Broken Hill [Kabwe], Cave of Hearths, and Hopefield). Thus, backward and forward in time from the dating of the hard-core *H. erectus* specimens, the problem of recognizing fossils as belonging to *H. erectus* becomes more difficult; the "lower" and "upper" boundaries of the species become blurred. These are the transitional zones in which a predecessor species seems to have been grading imperceptibly into its evolutionary product, *Homo erectus*, and in which *H. erectus* apparently was undergoing further evolutionary change into its descendant species, *H. sapiens*, to which modern man belongs.

Homo habilis

THE BODILY STRUCTURE OF HOMO ERECTUS

It is striking that the anatomical differences observed between *H. erectus* and *H. sapiens* have been confined to the skull and teeth. The limb bones thus far discovered have been indistinguishable from those of *H. sapiens;* on this evidence it has been inferred that *H. erectus* was an upright-walking man of medium stature.

By courtesy of the University Museum of Archaeology and Ethnology, Cambridge



Reconstructed skull of Peking man (*Homo erectus pekinensis*).

An estimate of the size of his brain may be obtained from the capacity of the interior of the fossil braincase (cranium). Such endocranial (interior) capacity measurements show that *H. erectus* was smaller brained than is modern man. The average capacity for 14 crania of *H. erectus* from Java, China, and Africa is 941 cubic centimetres (cc; 57 cubic inches); the smallest and largest capacities determined for this group of fossils are 750 and 1,225 cc. The average capacity in modern *H. sapiens* is 1,350 cc (82 cubic inches), but the range for modern man is appreciable, perhaps 1,000 to 2,000 cc (61 to 122 cubic inches). Manifestly, the upper part of the range for *H. erectus* overlaps the lower part for *H. sapiens.*

Some difference in estimated brain size is apparent between the Javanese and the Chou K'ou-Tien (Peking Chinese) populations of *H. erectus*. Thus, for seven Javanese crania, the average is 883 cc (54 cubic inches), with a range from 750 to 1,030 cc (46 to 63 cubic inches); while for five Chou K'ou-Tien crania, the capacity ranges from 915 to 1,225 cc (55 to 75 cubic inches) and averages about 1,045 cc (64 cubic inches). That is, the mean capacity in the Peking fossils of *H. erectus* exceeds that of the Javanese by about 160 cc (10 cubic inches). Investigators note with interest that the Lan-T'ien cranium (also Chinese), which comes from an earlier period than do those of Chou K'ou-Tien, and is an approximate contemporary of the earlier fossils from Java, shares with the Javanese group a smaller cranial capacity (780 cc). Theoretically, the difference in brain size between the two groups of Asian fossils of *H. erectus* may be the consequence of further evolutionary increase in brain size in later populations of *H. erectus*. Alternatively, it may simply be interpreted to represent differences in average features between two different "races" or subspecies of *H. erectus*. The solitary African value available is that of Olduvai hominid 9, which has a capacity of 1,000 cc (61 cubic inches).

Differences in brain size

While the cranial capacity of *H. erectus* falls short of that of *H. sapiens* by about 470 cc (29 cubic inches) on the average for the Javanese group, and about 300 cc (18 cubic inches) for the Peking group, *H. erectus*, in turn, exceeds the australopithecine capacities (see AUSTRALOPITHECUS) by an average of about 430 cc (or 26 cubic inches) in the case of the Javanese fossils and about 600 cc (or 37 cubic inches) for Peking man. Thus, the cerebral gap between *Australopithecus* and *H. erectus* is slightly greater than that between *H. erectus* and *H. sapiens*. Into the former gap fit the cranial capacities of that rather mysterious and controversial little group of fossils known in the literature as *H. habilis*. Clearly, the last word has not been written on their affinities. The possible evolutionary place of the cranial capacity of *H. erectus* seems to emerge clearly in the following list (progressing from some of the most ancient fossils to today's man).

Average capacity of braincase	
Smaller *Australopithecus* (6 fossil examples)	450 cc
Robuster *Australopithecus* (2)	515 cc
Homo habilis (3)	656 cc
Javanese *H. erectus* (7)	883 cc
Chinese (Peking) *H. erectus* (5)	1,043 cc
H. sapiens (many modern men)	1,350 cc

Apart from their characteristically small capacity, the skulls of *H. erectus* show a series of distinctive features. The braincase is low, with sides that taper upwards, and the bones of the cranial vault are thick. Over the eye sockets is a strongly jutting ledge of bone, called a supraorbital torus, while a markedly thickened shelf of bone (occipital torus) adorns the hind end of the skull. There is a receding forehead—a wit has called these men the original lowbrows—and the front part of the cranium immediately behind the supraorbital torus is appreciably constricted from side to side.

The nose of *H. erectus* is wide, the jaws and palate being broad and somewhat prominent. The teeth are on the whole larger than those of *H. sapiens*, though smaller than those of *Australopithecus*. The front teeth (incisors and canines) are especially large for a hominid, some even exceeding in size those of *Australopithecus*. A few *H. erectus* specimens from Java have so well-developed an eyetooth (or canine) that the tip protrudes beyond the crowns of the adjacent teeth. In terms of evolutionary theory, this is considered an archaic trait, since it occurs in apes, but in no other hominids, extinct or living. Several other primitive characteristics are preserved in the teeth, such as the presence of three roots on a few of the upper first premolars and an increasing length of the lower molars from first to third. Considered as a totality, however, almost all authorities agree that the dentition of *H. erectus* is hominid in character, rather than pongid (of the ape family).

The total pattern of the bodily structure of *H. erectus*, as preserved in his bones, shows a provocative blend of advanced features—such as his lower limbs or locomotor apparatus, which suggest that he stood upright and walked on two legs, as is typical of *H. sapiens*—and of primitive features—such as his archaic dental traits (testifying to his probable genetic inheritance from more ape-like forebears) and his brain size, which is smaller than that of modern man, though not so small as that of *Australopithecus*. In addition, in his thick skull bones and extraordinarily developed eyebrow ridges and occipital torus, some investigators say they see unique, specialized features, not characteristic either of his presumed ancestors or of apes and not pointing to *H. sapiens* as the direction of subsequent evolution.

Some scientists even infer that these last traits show *H. erectus* to have specialized so far off the modern human line that it could not have been ancestral to *H. sapiens*. Certainly it is notable that the more ancient *Australopithecus* had thin skull bones and only modest protuberances on his cranium; while the later *H. sapiens* also tends to have thin skull bones with a marked diminution in the size of crests and ridges sculpted on the surface of the cranium. It is often said that only two choices are open in interpreting this situation: either *H. erectus*, with his thick cranium and monstrous adornments on the skull, could not have been on any direct evolutionary line from *Australopithecus* to *H. sapiens;* or else, it must be postulated that *H. erectus* evolved his specialized features from *Australopithecus* and then lost them again (or underwent a kind of evolutionary reversal) to produce the thin, smooth cranium of modern *H. sapiens*.

Yet, the two choices offered for solving the problem are judged by some authorities to be over simplified. It has been observed that the exponents of this view probably fail to take into account that there is very little evidence about the variability of these features—cranial thickness and external embellishments of the skull—among members of even one population of *H. erectus*, let alone among different populations of *H. erectus* dispersed through two or three large continents. Then, too, practically nothing is known about the climatic or ecological conditions under which cranial thickening occurred, or of the effect on skull growth of the brain enlargement that was so striking a feature of the evolutionary advance of early middle Pleistocene Man to the men of the late middle and of the upper Pleistocene. These and many other questions need answers before *H. erectus* can be written off as an ancestor of *H. sapiens*. In the meantime, another hypothesis that meets most of the available evidence is that the anatomical structure of *H. erectus* was a blend of ancient and modern features, precisely because he belonged to a people in transition. It is suggested that he was in the process of evolving from pre-*Homo erectus*, probably *Australopithecus* and *Homo habilis*, to post-*Homo erectus;* that is, to *H. sapiens*. In most details, his bodily structure fulfills what might have been predicted for an intermediate between *Australopithecus* and *H. sapiens*.

Unanswered questions

INFERRED BEHAVIOUR OF HOMO ERECTUS

To understand this species of fossil man more completely requires looking past his bones and beyond his dispersal in time and space. What evidence exists that bears upon the behavioral or cultural pattern of his daily existence?

First, the discovery sites themselves may throw light on this question. At Chou K'ou-Tien, the remains of *H. erectus* were found in a cave deposit; this in itself does not prove that these men were consistent cave dwellers. But the additional evidence of associated remains of stone and bone that seem to have been accumulated by these creatures—charred animal bones, collections of seeds, and what could be ancient hearths and charcoal—all point to *H. erectus* as having spent appreciable periods of time as a troglodyte (cave dweller) at Chou K'ou-Tien. The remains of Lan-T'ien, Trinil, Sangiran, Modjokerto, as well as Ternifine and Olduvai, were all found in open sites, sometimes in stream gravels and clays, sometimes in river sandstones, in conglomerate and volcanic rocks, or in lake beds. These suggest that *H. erectus* also lived in open encampments along the banks of streams or on the shores of lakes; proximity to water was crucial to the survival of man. The remains of Vértesszőllős Man showed that he occupied mud flats deposited along with minerals around springs in a tributary of the Danube River. These open presumed campsites revealed by excavation contain abundant stone implements and stone chips that seem to have resulted from their manufacture, fractured and partly burnt bones of animals that could have been hunted for food, and traces of what appears to have been a hearth.

Thus, both Chou K'ou-Tien and Vértesszőllős have shown signs that early middle Pleistocene Man had a controlled mastery of fire. Indeed, outside the cave at Chou K'ou-Tien, charcoal was found along with traces of a stone toolmaking industry in an open gully deposit that seems to be slightly older than the cave deposit itself (containing the bones of *H. erectus*). In this region of China, therefore, it has been observed that the earliest convincing indication of the use of fire by men immediately precedes the earliest example of a cave being occupied by them. This supports the notion that successful cave dwelling by human creatures depended on their first having mastered fire.

Mastery of fire

It was not only cave dwelling that mastery of fire seems to have made possible. Able to keep warm, man was now apparently able to move into colder climes; indeed, this factor may have speeded the migrations of ancient men into the chilly, often glaciated regions of prehistoric Europe. Sooner or later, too, man started cooking his food, thus reducing the work demanded of his teeth. This, in turn, may have played an important part in minimizing the evolutionary advantage of big teeth—cooked food needs far less cutting, tearing, and grinding than does raw food. This relaxation of the evolutionary selective pressure that favoured the survival of people with strong, big teeth may, in turn, have led directly to a diminution in the size of the teeth—one of the features, it will be recalled, that distinguish *H. sapiens* from *H. erectus*.

Other signs of the culture of *H. erectus* are the implements found in the same deposits as his bones. Chopping tools made from split pebbles characterize both the Chou K'ou-Tien and Vértesszőllős deposits; both are members of a so-called chopper–chopping-tool family of industries described for prehistoric Southeast Asia in the scientific literature. At both sites, stone flakes and bone artifacts showing human workmanship occur too. This evidence has been thought by some to support efforts to classify Vértesszőllős Man as belonging to *H. erectus*. The uncertainty entailed in trying to infer from cultural evidence to biological classification is shown by the cultural associations of the Ternifine man. This northwest African *H. erectus* was found in association with totally different kinds of stone implements; these comprise bi-faced hand axes and scrapers that have been characterized as representing what archaeologists call an early Acheulean industry (Acheulean II). This is part of a great Acheulean hand ax complex of human industry, remnants of which are found widely spread over large parts of Europe, Africa, and Asia (except Southeast Asia). An Acheulean industry is known also from Olduvai Gorge, as is a local ancient form of stone-chopper manufacture known as the Oldowan culture; but the exact cultural associations of these stone tools with the African *H. erectus* (as exemplified by Olduvai hominid 9) are uncertain.

Hence, *H. erectus* has been found associated in some parts of the world with a chopper–chopping-tool tradition and in other places with an Acheulean bifaced hand ax industrial complex. This serves as a timely reminder that hereditary aspects of race or bodily structure are not inextricably interlinked with such cultural manifestations as toolmaking traditions. It is most reckless, and indeed fallacious, to infer from stone implements the physical type of the man who might have made them. Apparently *H. erectus* had considerable versatility in manufacturing implements. Even during this early phase

Reckless inferences

of stone toolmaking activities, a well-diversified suite of tool types is recognizable, within both the east Asian chopper-tool industries and the African Acheulean complex.

Numerous nonhominid animal bones have been found with *H. erectus* remains; sometimes they seem to have been cooked, deliberately broken, and even fashioned. From this evidence, it seems that *H. erectus* was a hunter. His bodily (including cerebral) endowment and his manufactured equipment were so much superior to those of *Australopithecus* that it is highly probable his food-collecting techniques, including hunting, were better, too. Many scientists hold that *Australopithecus* was more of a scavenger than a hunter, perhaps at best an opportunistic hunter who seized his chance when a weak, young, sick, or aged animal crossed his path; indeed, many of the animal bones found in australopithecine deposits are of juvenile and aged creatures. *H. erectus*, on the other hand, seems to have been a confirmed hunter, a habitual eater of fresh meat; and his prey included animals of all age groups.

It can credibly be supposed that, as with present-day hunters (such as the Kalahari Bushmen and the Australian Aborigines), meat from the hunt formed only a part of the diet of *H. erectus*. Other juicy morsels may have been furnished by snakes, birds and their eggs, locusts, scorpions, centipedes, tortoises, mice and other rodents, hedgehogs, fish, crustaceans, and a myriad other edible forms of life. Many of these even children could have caught—as they do in the Kalahari today, before they are allowed to accompany the older men on the hunt. Vegetable food also must have played a big part in the diet of *H. erectus*, in the form of fleshy leaves, fruits, nuts, and roots. Accumulations of hackberry seeds, for example, were found in the Chou K'ou-Tien cave deposit.

There seems little doubt that *H. erectus* must have been omnivorous (as *H. sapiens* is today), for such a diet is the most opportunistic of all, and modern man is the most opportunistic of all living primates. Emancipated from too narrow an environmental dependence, from too restricted a dietary regimen, man has come to live off many diets, in many surroundings. He is *par excellence* the creature that lives with an eye for the main chance. *H. erectus* was probably one of the earliest of the great opportunists; and it is likely that his very opportunism endowed him with evolutionary flexibility, with adaptability, with a very plastic survival kit.

Another question one may ask about his culture is whether there is any evidence of ritual among these extinct people. There is no sign yet that *H. erectus* buried his dead; no complete burials have been found, no graves, no grave goods, no red ochre (a mineral used as a paint by later forms of man) on or around the bones. That cannibalism was practiced seems most probable; it has been pointed out that the human bones of Chou K'ou-Tien are in the same broken and splintered state (perhaps for eating the marrow) as are those of other animals. The preponderance of human skulls in the Chou K'ou-Tien deposit has been interpreted as evidence that *H. erectus* went in for headhunting. The site has yielded thousands of pieces of nonhominid animal remains pertaining to every part of the skeleton; but, although the bones of *H. erectus* are those of more than 40 individuals, the overwhelming majority are represented by parts of skulls, there being only a very few fragments of human limb bones. Human heads would seem to have been selected, much as one finds them in the lairs of later headhunters.

Signs of cannibalism and headhunting

Evidence of collecting skulls is not the only sign of ritual observances. In no single *H. erectus* cranium known (whether from China, Java, or Africa), is the base of the cranium intact. Every one shows damage to the region around the foramen magnum (the hole in the base of the skull through which the stem of the brain passes on its way to become the spinal cord). This damage appears to have been deliberately inflicted; that this is one of the thinnest parts of the cranium is an unconvincing argument for accidental damage; indeed, even in the much thinner walled crania of *Australopithecus* this region is typically found intact. On the other hand, very similar signs of intentional mutilation are found on the skulls of early and later Neanderthal men, as well as on those of the Bronze Age of Germany. Present-day headhunters of Borneo and New Guinea inflict very similar cranial mutilations on special ritual occasions; they extract the brain through the aperture they have made and eat it ceremonially, holding that this permits the name of the deceased victim to be handed on to the son of the ritual cannibal. At any rate, some authorities consider that the damage to the base of the crania of *H. erectus* was not only deliberate but that it is an indication of ritual mutilation with ritual (rather than nutritional) cannibalism. If this interpretation is correct, perhaps headhunting, ritual mutilation of the skullbase, and, possibly, ritual cannibalism were the earliest semblance of ritual in the life of *H. erectus*.

RECENT CONTROVERSIES CONCERNING THE RELATIONSHIP BETWEEN HOMO ERECTUS AND HOMO SAPIENS

Erectus as a direct ancestor of sapiens. One leading worker in recent times has energetically opposed the view that *Homo erectus* was the direct ancestor of *Homo sapiens;* paleontologist L.S.B. Leakey (1903–72) held that the anatomical features of *H. erectus* are too specialized for the species to have evolved into *H. sapiens*. In his view, the lightly constructed and unspecialized *Homo habilis* was the direct forebear of *H. sapiens*. Those of the lower Pleistocene hominids that evolved into *H. erectus* had, according to Leakey, gone out on an evolutionary limb, doomed to extinction and destined to make no contribution to *H. sapiens*.

Most other workers do not accept this view and recognize *H. erectus* as ancestral. Anthropologist P.V. Tobias has proposed a compromise view; he acknowledges that there is a small but significant body of evidence for the existence of middle Pleistocene hominids that did *not* show the classical features of *H. erectus*—such as thick cranial-vault bones, large brow ridges and occipital ridges, and big teeth. Examples are provided by the Kanjera crania from Kenya, the remains of Vértesszőllős, Swanscombe, Steinheim, Mauer, and the little group of East African fossils known as *H. habilis*.

The existence of what appear to be two forms of man in the middle Pleistocene is susceptible of several interpretations other than that proposed by Leakey. The remains may all belong to *H. erectus*, but the species may have been very much more variable than was suggested in earlier definitions (like that of Sir Wilfrid Le Gros Clark)—so variable, in fact, as to make the old anatomical definition of *H. erectus* virtually meaningless. Or, a dichotomy into a "specialized" line—the classical *H. erectus*—and a slender, unspecialized, thin-skulled line may have occurred. In this view, neither line would be excluded from contributing to the genetic heritage of *H. sapiens*. Some populations, it is suggested, could have reached *sapiens* status via the *H. erectus* pathway, others by way of the non-*erectus*, *habilis*, and pre-*sapiens* route.

More likely, the two lines did not remain distinct everywhere; crossing may have occurred to mingle the genes of the two forms. In that event, it is unlikely that one would be justified in calling each of the two lines a separate species; this comes back to the idea that the two supposed lines may represent nothing more than two subspecies, or "races," of the same species (that is, of *H. erectus*). According to this view, *H. erectus*—like *H. sapiens* today—was a polytypic species; that is, one divided into two or more different forms, each known as a subspecies. Indeed, it has been observed that a number of subspecies of *H. erectus* are recognized (*e.g., H. erectus erectus, H. erectus pekinensis, H. erectus mauritanicus*).

Supporting the compromise view (rather than Leakey's) is the appreciable number of *H. erectus* features to be found in some early members of *H. sapiens* of the late middle Pleistocene and the upper Pleistocene, and even in some modern men. It is argued that at the very least these features prevent the exclusion of *H. erectus* from having contributed to the ancestry of modern man.

The erectus–sapiens threshold. The U.S. anthropologist C.S. Coon, in his book *The Origin of Races*, has suggested that the passing over from *H. erectus* to *H. sapiens* represented the crossing of a kind of evolutionary Rubicon; that is, a clearcut boundary, or threshold, marked by not only bodily changes but striking cultural changes. This supposed interrelation of cultural achievement and the shape and size of teeth, jaws, and brain is a theorized state of affairs with which many students of fossil man strongly disagree. Throughout the human fossil record, there are examples of dissociation between skull shape and size, on the one hand, and cultural achievement on the other.

A smaller brained fossil *H. erectus* from China seems to have been among the first men to tame fire, while much bigger brained people in other regions of the world, and living later in time, have yielded no evidence that they knew how to handle fire. (Differences in inferred cultural activities between African and Asian members of *H. erectus* have already been emphasized above.)

In fact, evolutionary theory recognizes continuity between one species and a succeeding species in the same lineage. Such successive species in the evolutionary sequence are called chronospecies. The "boundaries" between chronospecies seem almost impossible to determine by any objective anatomical or functional criteria; one is reduced to the guesswork of "ruling off" at a moment in time to draw the boundary. Thus, competent authorities have seriously suggested that, in the last analysis, such a chronological boundary may have to be drawn arbitrarily between the last survivors of *H. erectus* and the earliest members of *H. sapiens*. The problem of defining the limits of chronospecies is not peculiar to *H. erectus;* it is one of the most vexing questions in paleontology today.

This kind of vague boundary, drawn for convenience at a moment in time, with the earlier hominids grading imperceptibly across this line into the later, is a different conception from Coon's theory that there is a firm, specific, biologically and culturally definable threshold, or boundary, between *H. erectus* and *H. sapiens*. It is with the latter assertion that many workers have expressed disagreement since the appearance of Coon's book; the concept he proposed has not so far gained wide acceptance.

Multiple erectus–sapiens evolutions. Here again, Coon has advanced a theory with which most workers concerned with human evolution have found themselves in serious disagreement. Assuming the validity of his notion of an *erectus–sapiens* threshold, Coon has claimed that *H. erectus* evolved into *H. sapiens* not once but five times, as each subspecies of *H. erectus*, living in its own territory, passed the postulated critical threshold. Since this part of Coon's theory depends on accepting his supposed *erectus–sapiens* threshold as correct, those who find this threshold concept at variance with the modern genetic theory of evolutionary change cannot accept this idea of Coon's either.

Coon's belief that Europeans evolved before Africans

A corollary of this concept is Coon's suggestion that different subspecies (or "races") of man evolved at different rates and, hence, crossed the *erectus–sapiens* threshold at different times. In particular, Coon has suggested that "the step from the ancestral *Homo erectus* to the modern *Homo sapiens* was taken by Caucasoid man in Europe no less than 200,000 years before the same step was taken by Negro man in Africa." This part of Coon's reasoning is not only based on a concept that is at odds with modern evolutionary theory, but some of the assumed "facts" on which it is based are highly problematical. He has claimed that the races of today had already appeared before the *erectus–sapiens* transition. Apparently by allocating two very early fossil human skulls from Europe (Swanscombe from England and Steinheim from Germany) to the European, or Caucasoid, "race," although they lived 250,000 years ago, he has concluded that the supposed *erectus–sapiens* transition had occurred *among Europeans, or Caucasoids,* as long ago as 250,000 BP. Few if any paleoanthropologists would agree with Coon that these early skulls are to be recognized as those of members of the Caucasoid, or European, "racial" group, as generally understood.

Turning to Africa, Coon has claimed that Broken Hill man (represented by parts of two skeletons found in Zambia in 1921) belonged to *H. erectus*. Since it is probable, on archaeological evidence, that Broken Hill man lived perhaps 30,000 or 40,000 years ago, Coon (by attributing these Zambian bones to *H. erectus*) has claimed that *H. erectus* was still present in Africa long after he had disappeared from the European scene. The crux of this argument hinges on controversial judgments of the species to which Broken Hill man is to be assigned. In contrast to Coon, most paleoanthropologists regard Broken Hill man as a member of a Neanderthal-like African subspecies of *H. sapiens*, commonly dubbed *H. sapiens rhodesiensis*. If that one set of Zambian bones has been misdiagnosed by Coon as *erectus*, one major prop of his argument—that the *erectus–sapiens* threshold was crossed in Africa much later than elsewhere—falls away.

The evidence for the late survival of *H. erectus* in Africa is questionable at best. No less suspect is the evidence on which Coon has based his claim that the first *H. sapiens* appeared in Africa very late. In the 1930s, Leakey discovered the heavily fossilized remains of a group of human skulls at a place called Kanjera on the shores of Lake Victoria in Kenya. The remains have relatively smooth brows and suggest affinity with *H. sapiens*. At least two possible interpretations have been advanced. The Kanjera cranias may represent a relatively smooth-browed variant of the polytypic species, *H. erectus*, since the Kanjera deposits in which they belong are of middle Pleistocene age; or they may represent an early African branch of *H. sapiens*. One is confronted here with the same twofold possible interpretation as was faced (above) in discussing the Hungarian fossils from Vértesszőllős. Coon has classified the Kanjera fossils as *H. sapiens* but has allocated them a date of 40,000 years, a dating for which there seems to be no evidence. As mentioned, the Kanjera deposit is known on the basis of associated remains of nonhuman animals to be of middle Pleistocene age. If the fossilized human skulls found there indeed do belong to the same time period as do the other animal bones, then Kanjera Man would be of the same age as are the earliest *sapiens* skulls found in Europe. Chemical tests so far performed support the view that the human and animal bones of Kanjera were contemporaneous. If this is correct, the age of the Kanjera crania would be far greater than Coon's figure of 40,000 years. Yet it is precisely the latter, dubious date that supports Coon's claim that the earliest known African representative of *H. sapiens* dates from this relatively recent, upper Pleistocene date.

Since considerable doubts surround Coon's statements about both the Broken Hill (Kabwe) and Kanjera skulls, his claim (based on this "evidence") that the supposed *erectus–sapiens* threshold was crossed in Africa very recently, and perhaps 200,000 years later than in Europe, has not been sustained.

This controversy about *H. erectus*, in which Coon's concepts have understandably gained no serious support, rests both on evolutionary theory and on the fossil facts. There seems to be no valid evidence for saying that African man has been *H. sapiens* for only 30,000 or 40,000 years and that Europeans of that species have been present for 250,000 years. One should be obliged on present evidence to conclude that Africa's earliest claimant to *sapiens* status (Kanjera) is of about the same age as the earliest *sapiens* men in Europe.

The controversy underlines the continuing need for more fossils (to establish the range of physical variation of *H. erectus*) and for more discoveries in good archaeological contexts that permit increasingly precise dating. Hopefully, additions to these two bodies of data will settle the remaining controversies and bring the unsolved problems of phylogeny and classification of *H. erectus* nearer to resolution.

BIBLIOGRAPHY. B.G. CAMPBELL, *Human Evolution* (1967), a general text with a functional approach to human evolution; W.E. LE GROS CLARK, *The Fossil Evidence for Human Evolu-*

tion, 2nd ed. (1964), an invaluable systematic synthesis, with definition and lucid exposition on *H. erectus;* C.S. COON, *The Origin of Races* (1962), an encyclopaedic compilation, with much useful factual data on *H. erectus* fossils, attempting to relate them to later human races; M.H. DAY, *Guide to Fossil Man* (1965), catalogs essential facts about selected hominid fossils, including *H. erectus;* W.W. HOWELLS, "Homo erectus," *Scient. Am.*, 215:46–53 (1966), a crisp, popular summary; F.S. HULSE, *The Human Species*, 2nd ed. (1971), an introductory text with brief but mature comments on *H. erectus;* G. KURTH (ed.), *Evolution and Hominisation*, 2nd ed. (1968), a volume of contributed articles, several of which discuss *H. erectus;* K.P. OAKLEY, *Frameworks for Dating Fossil Man*, 3rd ed. (1969), an important reference work, giving evidence for the dating of *H. erectus* and other fossil men; K.P. OAKLEY and B.G. CAMPBELL (eds.), *Catalogue of Fossil Hominids*, pt. 1, *Africa* (1967), includes African finds attributed to *H. erectus;* P.V. TOBIAS, *The Brain in Hominid Evolution* (1971), containing up-to-date data on endocranial capacity in *H. erectus;* S.L. WASHBURN (ed.), *Social Life of Early Man* (1961), symposium volume including an article on the awakening of ritual and ideology in *H. erectus.*

(P.V.T.)

Homoptera

The members of this large group of sucking insects, considered a suborder of the Hemiptera by many entomologists but a separate order by others, exhibit considerable diversity in body size and number more than 32,000 species. All of the Homoptera are plant feeders, with mouthparts adapted for sucking plant sap from a wide assortment of trees and wild and cultivated plants. Many homopterans cause injuries or destruction to plants, including fruit trees and grain crops; others are vectors of plant diseases; and a few provide secretions or other products that are beneficial and have commercial value to man. Most members of the Homoptera fall into one of two large groups; the Auchenorrhyncha contain the cicadas, treehoppers, froghoppers or spittlebugs, leafhoppers, and planthoppers or fulgorids; the Sternorrhyncha include aphids or plant lice, phylloxerans, coccids, scales, whiteflies, and mealybugs.

From Richard A. Pimentel, *Invertebrate Identification Manual* (1967); Reinhold Publishing Corporation

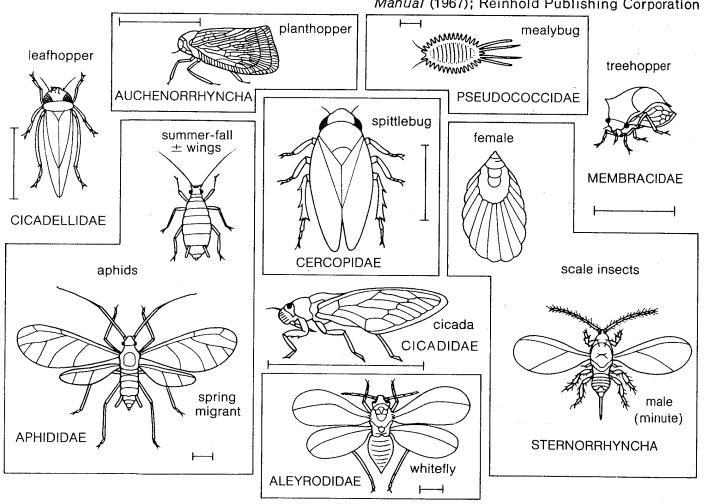


Figure 1: Representative homopterans. Line scales indicate approximate sizes of insects.

GENERAL FEATURES

Size range. The majority of homopterans range from 4 to 12 millimetres in length. There are cicadas in Borneo and Java, however, that are eight centimetres long with wingspreads of 20 cm. The large fulgorid or lanternfly can attain this size also. On the other hand, some scale insects are only ½ mm in length.

Distribution and abundance. Although Homoptera are distributed throughout the world, the relative numbers of individual species vary in a given locale. Only one cicadid species is known in Great Britain, fewer than 12 in all Europe; however, more than 200 cicadid species are known in North America, and about 180 in Australia.

The abundance of any species in a given environment depends upon the biotic potential of the insect, the abundance of the food plant, and other factors favourable for development of large populations. Certain species never reproduce in excessive numbers, while others, considered pests, produce millions of offspring. Insect species that feed on available crops or other plants present in quantities sufficient to support them normally develop large populations; for example, the oyster shell scale (*Lepidosaphes ulmi*) on fruit trees and ornamentals; the greenbug (*Toxoptera graminum*) on wheat; the potato leafhopper (*Empoasca fabae*) on potatoes, beans, and alfalfa; and grape leafhoppers (*Erythroneura*) frequently develop large populations that result in heavy plant losses.

IMPORTANCE

Homopterans, because they feed on sap sucked from plants, often cause injuries or destruction to the plants that nourish them. When such plants happen to be cultivated crops (*e.g.*, grains or fruit trees) or ornamentals valued by man, the economic loss resulting from infestations is severe. In addition, some homopteran species act as vectors of virus- and bacteria-caused diseases of their plant hosts (see DISEASES OF PLANTS). The check exerted upon insect pests by other insects is an important mechanism of natural control of populations. Predacious insects feed on small, weak species; parasitic insects live on or in a host and feed at its expense. Aphids, for example, are parasitized principally by members of the Hymenoptera; two important aphid predators are ladybird beetles and lacewings. Pests also may be controlled by chemical and biological methods (*e.g.*, development of resistant plants, as with European grapevines; see PEST CONTROL).

Plant injury

The homopterans are responsible for injuring numerous plants of economic importance to man. Cicadas or dog-day harvestflies are well-known pests frequently misnamed locusts. Characterized by their large size and the strident song of the male, periodical cicadas emerge every 13 or 17 years in large numbers, swarm in trees, mate, and lay eggs in green twigs. Of economic importance is the permanent damage to fruit twigs caused by egg deposition slits; when the weakened twigs mature into fruit-bearing limbs, they break under the weight of the fruit, and the crop is lost. Failures of this sort can be avoided by not planting young fruit trees in years of cicada emergence.

Leafhoppers alone cause various types of plant injury by interfering with the normal physiology of the plant. The salivary secretion of the potato leafhopper, for example, causes leaf cell hypertrophy that impairs transport of sugars; the resulting sugar accumulation in the leaves destroys chlorophyll and causes the leaves to turn brown and die. This injury, termed "hopper burn," can result in complete loss of a potato crop if not controlled. Another type of injury is caused by leafhoppers that feed upon plant mesophyll tissue. In addition to removing excessive amounts of sap, these insects also destroy the plant's chlorophyll, resulting in yellow spots on the leaves that eventually turn yellow or brown. *Erythroneura, Typhlocyba,* and *Empoasca* species cause this injury to apple trees and grapevines. Grape leafhoppers check growth and foliage function and cause formation of grapes that are inferior in size, colour, flavour, and sugar content. Plants are injured also when insects lay eggs in green twigs; for example, the egg punctures of several leafhoppers and treehoppers reduce the flexibility of plant limbs. Plant stunting and severe curling of leaves occur when the leafhopper *Empoasca fabae* punctures the undersurfaces of leaves and veins of bush beans and inhibits growth; the same leafhopper attacks alfalfa and causes leaves to turn yellow and drop off. In the same way, aphids and mealybugs cause leaf curling on potatoes and many types of ornamental plants, and the potato psyllid feeds on potato leaves and causes curling and yellowing known as "potato yellows."

The froghoppers, often called spittlebugs because immature stages live in spittlelike masses, feed upon a vari-

ety of plants. One important species, *Philaenus leucophthalmus*, a meadow insect that feeds extensively upon clover and alfalfa, causes severe stunting that can result in loss of up to 50 percent of a crop. Scale insects, unless parasitized, produce enormous populations on green twigs, young limbs, leaves, and fruit; when tree bark or shrubs become encrusted with one or more layers of scales, the entire plant often dies. Damage is caused to apples by the rosy apple aphid. Females of the third seasonal generation remain on the apple leaves until after small apples have formed. Many aphids crawl onto these tiny apples and puncture them causing dimpling of the fruit and normal incision of tiny apples. The cluster of apples, known as aphid apples are small and gnarled.

More than 100 species of leafhoppers are known organisms causing plant diseases. Some important plant disease viruses transmitted by leafhoppers are aster yellows (transmitted by *Macrosteles fascifrons*); potato yellow dwarf (several species of *Aceratagallia* and *Agallopsis*); and phony peach disease and Pierce's disease of grape (species of *Cuerna*, *Homalodisca*, and *Oncometopia*). Corn stunt is transmitted by species of *Dalbulus;* curly top of sugar beet by *Circulifer tenellus;* eastern and western x-disease by species of *Colladonus;* and elm phloem necrosis by *Scaphoideus luteolus*. One species of spittlebug is a vector for a yellow virus of peaches. Aphids are vectors for several virus mosaic disease organisms. A membracid species transmits the virus that causes pseudocurly top of tomato and tobacco, and two species of fulgorids are vectors of virus disease organisms of rice. The whitefly *Bemisia tabaci* transmits the virus that causes tobacco leaf curl, and species of mealybugs are vectors of the virus that causes pineapple wilt. The bacteria that cause fire blight disease on pear, apple, and quince trees are transmitted by several types of insects including leafhoppers. The bacterial pathogen, *Neofabraea perennans*, that causes perennial canker of apple is transmitted by the woolly apple aphid.

Because homopterans suck more sap from plants than they need, the surplus is excreted from the tip of the abdomen as sweet droplets known as honeydew. If the insect is feeding on apple foliage and honeydew falls on apples, a sooty fungus grows in each droplet; and the apples become black spotted and are no longer marketable.

Beneficial aspects

Of great economic importance are insects that secrete lac on twigs in tropical and subtropical regions. The lac is refined and used in preparing shellac and varnishes. More than 4,000,000 pounds of lac are refined annually. Other waxes secreted by aphids and scale insects are used in candlemaking and in medicines.

Although few homopterans produce food for man, the tamarisk manna scale, *Trabutina mannipara*, is thought to have produced the biblical manna for the children of Israel. The females produce large quantities of honeydew that solidify in thick layers on plant leaves in arid regions. This sugarlike material, still collected by natives of Arabia and Iraq, is considered a great delicacy. The term manna often refers to plant products also. Certain species of scale insects produce a gum that was used as chewing gum by tribes of North American Indians. Female root-inhabiting scale insects, species of *Margarodes*, enclose their bodies in gold and bronze coloured wax cysts that are used in strings of beads. Certain colour patterns and designs of the forewings of tropical species of leafhoppers and planthoppers have been used in artwork by various peoples. For many generations the Mexican Indians have used a black, white, and red colour design in their art. The design is that of the forewings of a brilliantly coloured *Agrosoma* leafhopper, found on bushes along streams.

NATURAL HISTORY

Life cycle. Generally, homopterans are bisexual, have periods of mating, and produce eggs; however, individual life cycles vary in length and complexity. Metamorphosis is simple or gradual; immature stages resemble adults except that the latter usually have wings. The life cycles of most homopterans are short. A typical example is the common meadow spittlebug, *Philaenus leucophthalmus;* it has one generation a year. Eggs, laid in late summer on stems or sheaths of host plants, hatch the following spring. Adults appear and mate during the summer.

Periodical cicada. The life cycle of three species of periodical cicadas is the longest known; it lasts 17 years. In the temperate zone enormous numbers of orange-winged adults emerge in spring and mate within a week; the females lay eggs a few days later. Using her strong ovipositor, the female cuts deeply into green twigs and through the harder wood of deciduous trees; she inserts 12 to 14 eggs through drilled slots into each of two chambers separated by a thin partition of wood. The female drills slots until she has deposited a total of 400 to 600 eggs. Injury to oaks, hickories, and young fruit trees is severe; and branches usually die beyond the point of egg insertion. Although eggs may be deposited in some 75 different kinds of trees or shrubs, the females prefer hickory, oak, apple, peach, pear, and grape.

Tree damage

The eggs hatch after two to six weeks; the young drop or crawl to the ground, enter the soil using their large digging claws, and begin a subterranean life, feeding on suitable tree and shrub roots for 16 years (periodical cicada). In apple orchards the young feed at depths of 2 to 18 inches; in wooded areas, at depths of 18 to 24 inches. "Harvestflies," common black and green species, appear in summer and feed on tree roots from two to five years. The periodical cicadas that live in central areas of the United States have a 17-year cycle, but three southern species complete their development in 13 years, and there are species that complete the cycle in one year. Since enormous numbers of nymphs feed on tree roots, many trees would die if the metabolic rate of the insect were not low; however, sap is taken from roots very slowly over a period of several years, and most trees survive. Although nymphs are almost full grown in eight years, they continue to feed and develop until the 13th or 17th year when mature nymphs emerge from the soil, climb any convenient tree or post, and attach themselves firmly. The dorsal line of the integument splits, and the adults emerge slowly through this opening. Adults live only a few weeks. Broods of both the 17- and 13-year cicadas have been studied. The largest and most widespread brood of the 17-year form occurs in abundance over much of the northeastern quarter of the United States.

Leafhopper. Many leafhoppers (*e.g.*, *Empoasca maligna*, *Gyponana mali*) have cycles that involve passing the winter as eggs inserted in apple twigs. Other leafhoppers, however, such as *Empoasca recurvata* and *Erythroneura*, hibernate as adults during the winter. The sugarbeet leafhopper, *Circulifer tenellus*, winters as an adult in desert areas and produces an early spring generation on desert plants. As desert plants become unfavourable for feeding, the leafhoppers migrate to available cultivated plants where from one to four summer generations are produced. When the crop is harvested or the plants become unfavourable for feeding, the leafhoppers return to desert plants. Although definite alternation of desert and cultivated host plants occurs in this life cycle, no specific plant serves as a primary or secondary host. Plant selection by migrating leafhoppers is determined largely by the amount of rainfall and succulence of plants, both wild and cultivated. While most species have one generation a year, a few have two or three. The life cycle of planthoppers or fulgorids is similar to that of leafhoppers, while the pear psylla, *Psylla pyricola*, hibernates as an adult and can produce four generations of nymphs.

Alternation of host plants

Whiteflies. The whiteflies, common on citrus trees and in greenhouse plants, do not survive winter out of doors in the North but produce several generations a year in the South. The metamorphosis of whiteflies varies from the typical gradual form. In the first instar (interval between molts) the young, active, wingless forms are usually called larvae; during three subsequent instars, the sessile and scalelike nymphs undergo internal wing development. The molt from last larval instar to pupa occurs inside the last larval skin, which forms a puparium. Metamorphosis is essentially complete.

Aphids. The aphids or plant lice, soft-bodied insects that develop large populations, have several types of complex life cycles. Generally aphids overwinter in the egg stage on twigs or plant buds, usually designated a primary host. In the spring the eggs hatch into females that reproduce parthenogenetically, giving birth to living young. Several generations may be produced during the season in this way. Early generations are usually wingless, but by the third generation winged individuals appear. In many species these winged forms migrate to a secondary host plant, usually an annual, and the same type of reproductive process continues. In the latter part of the season, winged aphids of both sexes appear and migrate back to the primary host where mating occurs, and the females lay the overwintering eggs. There are two distinctive characteristics in the aphid life cycle: first, seasonal alternation of food plants involving a primary host (typically a perennial) during the winter and a secondary host (an annual) during the summer; second, there is alternation between sexual and asexual cycles, with eggs resulting from sexual mating and living young, usually females, being produced asexually.

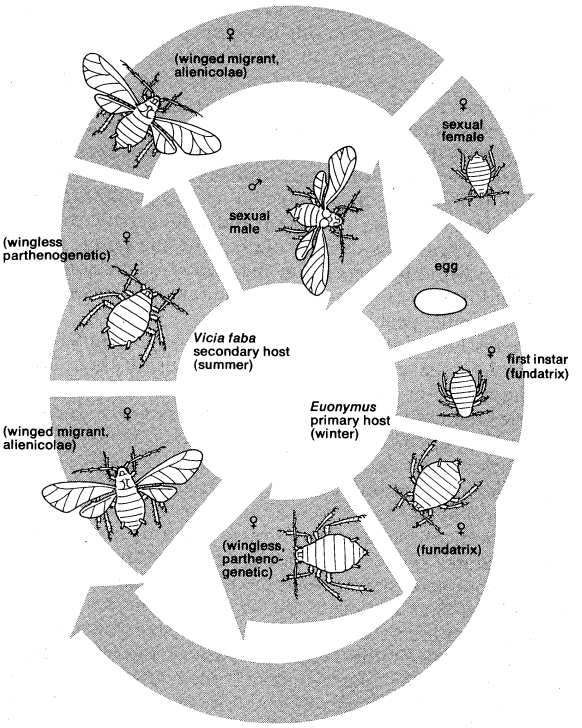


Figure 2: Life cycle of *Aphis fabae* (see text).

Scale insects. The scale insects have modified life cycles also; for example, the oyster shell scale, *Lepidosaphes ulmi,* typically passes the winter as an egg beneath a secreted scale covering; however, the San José scale *Aspidiotus perniciosus* produces living young. In either case newborn young, active crawlers, leave the scale covering to search for food. After a few days they molt, losing their legs, antennae, and anal spines, and retaining poorly developed eyes. They secrete a hard scale about their soft bodies, insert their mouthparts into a plant, and remain sessile. As the females mature, they increase in size, enlarging the scale covering periodically, but do not change form or develop wings.

Young males also have a crawler stage but become sessile and inactive after the second molt, passing through a more complete metamorphosis beneath the scale covering. The last preadult instar has two external wings and is called the pupa. Adult males have two wings and two small knobs or halteres where the second pair of wings would normally develop. Some males have three pairs of eyes. Adult males seek out wingless females, concealed beneath the scale covering, and mate with them; eggs form in the female. As many as three males may mate simultaneously with one female.

Reproduction and growth. Reproduction is bisexual among the homopterans; but asexual reproduction occurs in the aphids, in a few primitive leafhoppers, and possibly in species whose life cycles are not known in detail. An unusual situation occurs in the normally hermaphroditic cottony cushion scale *Icerya purchasi;* both sexes are present in one individual, and the eggs of any individual may be fertilized by its own sperm.

Egg laying

In the Auchenorrhyncha eggs are laid by the female, who uses an egg-laying structure, the ovipositor, to insert eggs into plant tissue. In the Sternorrhyncha (*e.g.,* aphids) the female places her eggs on the surface of the plant. The eggs of scale insects are retained in the body of the female or remain under the scale covering if separated from the female. In mealybugs and certain "cottony" scales, eggs are extruded from the body and remain in a mass enclosed by waxy plates or shreds. In most homopterans, each female produces a few hundred eggs. Exceptions occur in some scale insects (*e.g.,* cottony maple scale) where a female may lay 5,000 eggs.

Growth is gradual and is accompanied by periodic molting. The nymphal stages, or instars, between egg and adult usually number five in leafhoppers and related species. Wings, if present, develop when the fifth instar molts and the adult emerges.

Sexual dimorphism. Sexual dimorphism occurs in most groups of Homoptera. Males and females often are coloured differently; for example, the male leafhopper *Arundanus nacreosus,* a species common on cane, is orange, the female, milk white. Size and form vary between males and females also; the male marsh leafhopper *Hecalus lineatus* is not only a different colour than the female but also only half as long. In treehoppers the pronotum (the dorsal sclerite of the prothorax) often is so different in shape and size between the two sexes that they appear to be two species; examples of this are *Umbonia crassicornis* and *Philya inflata.* Among scale insects most females lack wings, legs, and antennae, while males have all three; males and females are so diverse in appearance that previous knowledge is necessary to associate two sexes of the same species. Most homopterans lack defense mechanisms; however, one scale insect, *Phenacoccus echeveria,* extrudes a honeydew-like material from the posterior ostioles as a defense mechanism.

Ecology. *Habitats.* Every insect lives in a habitat defined by specific physical, chemical, and biological conditions. If these conditions are changed sufficiently, the insect cannot survive and will either migrate to available acceptable conditions or perish. Temperature and humidity are important climatic factors in determining geographical regions and local habitats of specific homopterans. The distribution of homopterans is influenced also by conditions that favour distribution of host plants.

Influence of plant distribution

Plant distribution is determined largely by rainfall-evaporation ratios; insects with specific host relationships occur in the same regions where the plants are found. Other climatic factors may limit the insect to a smaller range within the host plant range; for example, selection of food plants by the desert species of sugar beet leafhopper depends on the abundance of rainfall during one season. Host plants of a given species may be closely related, as legumes on which eggs are deposited and adults live; or the life cycle may be divided between alternate unrelated host plants. The fact that most species are specific in their plant relationships determines habitats such as swamp, marsh, bog, meadow, prairie, desert, deciduous or coniferous forest. Certain species occur only on sagebrush or rabbit brush in the desert, on blueberry bushes in a bog, on white oak in a deciduous forest, or on white pine in a coniferous forest.

Moisture or humidity relationships also affect the habitats of homopterans. The eggs of most auchenorrhynchans are deposited in tender plant stems or in the undersides of leaf ribs or veins. Thus, the incubation period is passed in saturated humidity. After hatching, the nymphs feed on the under surface of the leaf and remain in high

relative humidity since most of the stomata, through which transpiration occurs, are on the under surface. A reduction in relative humidity due to reduced transpiration can destroy large field populations. Certain leafhopper and fulgorid species, although they are not adapted for aquatic life, can live on plants and produce normal populations under conditions of periodic tidal submergence, even in cold waters.

Formation of galls. Insect galls, abnormal growths of plant tissue, are caused usually by the mechanical or chemical stimulus of egg laying in plant tissue and by subsequent activities of the hatching young. The young usually live and feed inside the gall and complete their development before emerging. Some 60 species of homopterans, including aphids, psyllids, and coccids, cause plant galls, although aphids are responsible for a majority of them. Galls frequently seen on foliage include aphid leaf galls, caused by the grape phylloxera *Phylloxera vitifoliae;* the leaf petiole gall of poplar, caused by the aphid *Pemphigus populitransversus;* and the elm cockscomb gall on elm leaves, caused by the aphid *Colopha ulmicola.* Different species cause the formation of different types of plant galls.

Associations with other insects. There are insects that attack homopterans as predators or parasites or use them to provision their nests. Colonies of aphids and scale insects are prey for several kinds of ladybird beetles. Female beetles lay their eggs on leaves or twigs where aphids are feeding; when the beetle eggs hatch, the larvae feed upon the homopterans in the colony. One larva of *Hippodamia convergens* can consume 300 aphids in a two week developmental period, while the adult female devours several thousand aphids in three months. Certain species of flower flies or syrphids commonly lay their eggs on leaves or twigs where colonies of aphids are feeding. The hatching larvae thrust their piercing mouth structures into the bodies of the aphids and devour them by extracting visceral and body fluids. Another predator is the aphidlion larva, a chrysopid with mandibles like the ladybird larva; however, instead of chewing the aphids, the aphidlion larva inserts its mandibles into the body of the aphid and sucks fluids through a channel or groove on the inside of its mandible. Green winged adult aphidlions lay eggs in aphid colonies, placing them on stalks, so that when the young larvae hatch, there is an adequate food supply nearby. Most larvae of the chamaemyiids (*i.e.*, aphidflies) feed on aphids, scale insects, and mealybugs; the larvae of Drosophila known as pomace flies, are predacious on mealybugs and other small Homoptera; and the larvae of a few gall midges (*i.e.*, cecidomyiids) prey on aphids and scale insects. Certain Diptera have parasitic larvae that feed on the internal tissues of homopterans including certain scale insects, leafhoppers, and planthoppers. Some moth larvae are parasites of fulgorids, while other larvae are internal parasites of female gall-like coccids of the genus *Kermes.*

Homopteran predators

Among the Hymenoptera, certain wasps are parasites of planthoppers, leafhoppers, and treehoppers. The larvae of dryinid wasps develop internally in the host although part of the body of the larva protrudes from the body of the host, forming a saclike structure between the abdominal and thoracic segments. Most encyrtid wasps are parasites of aphids, scale insects, and whiteflies. The female eulophid wasp develops as a parasite of scale insects, while the male, developing as a hyperparasite, attacks parasites of the scale insects (often females of its own species). The thamid wasps have habits similar to those of the eulophids in that both parasitize scale insects and whiteflies or are hyperparasites of chalcid wasps that parasitize homopterans.

Another interesting insect association concerns the sand wasps. They paralyze homopterans by stinging them, then store them in burrows, lay eggs, and rear young using the homopterans as food. Best known of these is the large cicada-killer wasp (*Sphecius speciosus*); the female digs a burrow in well drained soil, stings a cicada until it ceases to struggle, places it in the bottom of the burrow, lays her eggs on the cicada, covers the burrow, and dies. The larvae develop on the cicada, remain in the burrow until the following spring or summer, and emerge as adult wasps. Other wasps also burrow in the soil and provision their burrows with one or more kinds of Homoptera, particularly leafhoppers, planthoppers, or treehoppers. Aphid wasps use the same method of provisioning their nests, while squareheaded wasps usually use leafhoppers of one species to provision burrows in decomposed wood.

Subterranean life. Many species of adult and young aphids are subterranean and feed on the roots of plants. In some species the alternate food plant is no longer used, and the aphids no longer develop wings. Some entire colonies spend years below the surface of the soil; other species spend most of each year underground; and a few species appear above ground, locate a new host plant, and immediately seek roots. The woolly aphid can live indefinitely on the roots of apple trees but can exist only part of the year on elm, the alternate host. The strawberry root louse has a sexual cycle in which eggs are laid, but these aphids are dependent upon ants for survival; not only do the ants care for the eggs in their nests but they also carry the young aphids from plant to plant. In some subterranean aphids the sexual cycle, and with it the egg-laying stage, has disappeared entirely. Subterranean aphids have no predators and few parasites. Other root feeders are young cicadas, certain young cercopids, some cixiid nymphs of the fulgorids, and immature stages of a few leafhoppers.

FORM AND FUNCTION

External features. *Polymorphism.* Polymorphism is marked in several groups of Homoptera. The Cicadidae are similar in form but vary in size and coloration. The Cercopidae include different types. The small *Clastoptera* are short, ovate, and froglike in appearance and are called froghoppers; the more elongated *Philaenus* are often called spittlebugs. The Membracidae, or treehoppers, have an enlarged prothorax that often covers the head, thorax, and abdomen; it may protrude forward resembling a coarse spine, project anteriorly at the sides so that the insect appears to possess a pair of pointed horns, or produce a large keeled hump above the abdomen. The most bizarre and curiously shaped treehoppers are found in tropical American countries; here the prothorax develops into chitinous adornments and processes.

From H. Weber, *Grundriss der Insektenkunde* (1954); Gustav Fischer Verlag

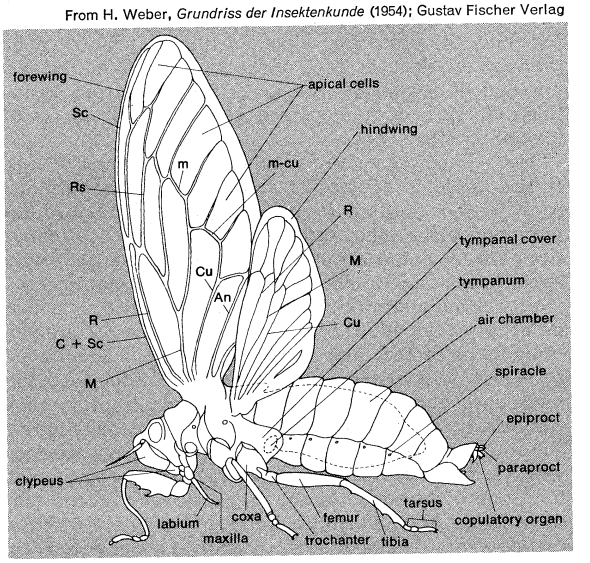


Figure 3: External features of the cicada. Wing veins: Cu, cubitus; M, media; m, medial; R, radius; C, costa; Rs, radial sector; An, anal; m-cu, mediocubital; Sc, subcosta.

The bodies of the Cicadellidae, or leafhoppers, are dorso-ventrally flattened or cylindrical; and the head varies in shape and size from short, broad, and rounded to long, thin, and bladelike. The head size and structure of ful-

gorid genera vary. Species of *Scolops* have a long, slender, anterior projection of the head that resembles a beak or snout; however, the true mouth structures are beneath the head. In the genus *Apache* the head is flattened laterally and projects as a vertical thin leaflike structure, while in *Cyrpoptus* the head, flattened dorso-ventrally, is horizontal. Dimorphism in the Aleyrodidae or whiteflies occurs only when inactive and sessile immature stages succeeding the first instar pass through a quiescent stage (a pupa) that has no resemblance to the winged adult.

Typically aphids have both winged and wingless adults. Some species ("woolly" plant lice) have waxy plates or fibres over their bodies. Most aphids, however, do not secrete waxy materials. Some aphids are root feeders, some are gall formers, but most are leaf and stem feeders. Polymorphism can result from host-induced variation; for example, progeny of a European fruit *Lecanium* female develop into morphologically different insects depending on the host plant. Generally, homopteran body form is similar to that of other insects. It consists of head, thorax, and abdomen, all covered with a chitinous exoskeleton (see INSECTA).

Head. The head is usually fitted with a pair of large compound eyes, but in certain male scale insects three pairs of eyes are present. Simple eyes or ocelli usually occur on the head and probably function as organs of light perception. Cicadas normally have three, while other homopterans have two or none. Vision is variable among the homopterans. Reactions to visual stimuli are greatest in leafhoppers, spittlebugs, planthoppers, treehoppers, and jumping plant lice; reaction is slower in cicadas, much slower in aphids, and practically nonexistent in scale insects and mealybugs. Leafhoppers as well as some spittlebugs, planthoppers, and tropical cicadas are attracted to lights at night. Members of other groups seldom respond to light.

A pair of antennae, arising below and between the eyes, are usually short and bristlelike, varying in length throughout the group. They are probably the most important sensory structures and are of taxonomic significance in species identification of aphids. Mouth structures of homopterans arise at the back of the head. The beak (or proboscis), elongate and segmented, is composed of a sheathlike labium that encloses four piercing stylets, two mandibles, and two maxillae. The stylets alone enter the plant tissue; the mandibles do the piercing; and the two inner stylets, the maxillae, fit together to form a sucking tube composed of two channels, one for conducting food, the other for saliva. In the Auchenorrhyncha the beak arises at the back of the head, while in the Sternorrhyncha it appears to arise between the front coxae; this difference in point of origin of the beak also is of taxonomic significance.

Legs. Each segment of the thorax bears a pair of legs. In the cicadellids, fulgorids, cercopids, membracids, and psyllids, the hindfemurs are enlarged and adapted for jumping; in the other groups the femurs are normal in size. The type and arrangement of spines on the femur are of taxonomic importance in separating certain families. The mesothorax and metathorax both bear a pair of wings in the adult stage; the wings are usually membranous or of the same texture. Adult female scale insects and most female aphids lack wings. Male scale insects have one pair of wings on the mesothorax; wing rudiments, called halteres, are found on the metathorax. The first pair of wings are hyaline, opaque, pigmented with various colours, or covered with waxy secretions in the form of powder, shreds, or plates. The second pair of wings are always membranous.

Abdomen. The abdomen, typically 11-segmented, appears to have only 7 or 8 segments because the last few segments are modified as specialized genital structures. Genital segments 8 and 9 bear structures associated with external openings of genital ducts. In the male these structures are modified for copulation and transfer of sperm to the female; in the female they are modified for oviposition. Although external genital structures, these are usually enclosed in a genital chamber. The female genitalia in the Auchenorrhyncha consists of an ovipositor, formed by the appendages (gonopods) of segments 8 and 9. The ovipositor, a pair of basal plates and three pairs of elongate bladelike structures, generally is used to pierce or drill slots in plant tissue for oviposition. The variable external genitalia of the male, often complex, are frequently of considerable taxonomic value. In the Auchenorrhyncha they are contained in a genital chamber consisting of portions of the ninth segment. The genitalia consist of the styles and the aedeagus, equipped with a gonopore through which sperm are discharged during mating. When the male mates, the aedeagus and styles are exposed directly to the base of the female ovipositor, where the sperm are transferred.

Internal features. In general the internal organs and systems are similar to those of other insects. Although the respiratory systems of homopterans and heteropterans

From R.E. Snodgrass, *Principles of Insect Morphology* (1935); McGraw-Hill Book Co.

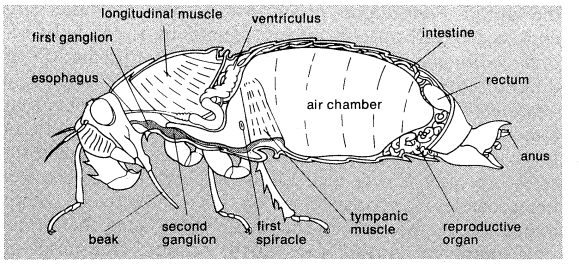


Figure 4: Internal features of the cicada.

are adapted for terrestrial life, certain species of both groups can live on submerged plants. The circulatory system is open, and blood circulates freely in the body cavity. The nervous system is composed of a ventral nerve cord with ganglionic masses for almost every segment.

The alimentary system is composed of three major parts, the foregut or stomodaeum, the midgut or mesenteron, and the hindgut or proctodaeum. The structure and function of the alimentary canal differ from other insects because homopterans feed entirely upon plant sap and ingest large amounts of it. Little absorption of food can take place in the foregut. The midgut, where digestion and absorption occur, is lined with epithelial cells that produce enzymes and absorb food after digestion. The residue passes into the ileum (small intestine) where, together with the waste products from the malpighian tubules, it passes to the colon for excretion.

Physiology and biochemistry. *Honeydew.* Plant sap contains a large quantity of water. In order to extract sufficient nutrients to survive, a large quantity of sap must be ingested. The filter chamber is a modification of the alimentary tract that allows nutrients to be concentrated in the midgut and small intestine as excess water (containing some sugar and waste materials) bypasses the midgut and small intestine to be exuded from the rectum as honeydew. It attracts ants and other hymenopteran and dipteran insects that feed on sweet nutrients.

Aphids are often called ant's cows. One well-known association is the corn root aphid and the corn field ant. The ants collect eggs in autumn, carry them to their nests, maintain the eggs through the winter, and place the young aphids on the roots of small weeds and grasses in the spring. As soon as newly planted corn seeds germinate, the ants place the aphids on corn roots and obtain honeydew by stroking the aphids with their antennae. The aphids are almost totally dependent upon the ants and are almost helpless in finding their preferred host, the roots of corn plants, without assistance. In a similar manner virgin female *Acropyga* ants carry in their mandibles on their nuptial flight a fertilized female mealybug as a source of honeydew for the new nest.

Spittle. Exuded from the alimentary tract by nymphs of the Cercopidae (*i.e.*, spittlebugs) are spittle masses commonly found on stems of meadow plants. The spittle fluid is voided from the anus after it has been mixed with a mucilaginous substance excreted by epidermal glands

of the seventh and eighth abdominal segments. Air bubbles are introduced into the spittle by means of the caudal appendages of the nymph. Immature spittlebugs rest head downward on the plant; as spittle is voided, it covers the nymph and is not easily dislodged, even by heavy rains. Adults do not produce spittle.

Wax and lac

Glandular secretions. Wax, produced by numerous wax glands and secreted by cornicles, is secreted by many aphids and scale insects. Mealybugs, whiteflies, woolly aphids, and cottony scales are named for white wax on their bodies or wings. Probably the best known wax producers are males of the Chinese wax scale *Ericerus pe-la;* they secrete large amounts of pure white wax useful in making candles. The Indian wax scale *Ceroplastes ceriferus* secretes a wax that is used for medicinal purposes.

There are several lac insects; some secrete highly pigmented wax. The Indian lac insect *Laccifer lacca* is important commercially. It is found in tropical or subtropical regions on banyan and other plants. The females, globular in form, live on twigs in cells of resin; their bodies are covered with exudations of lac, sometimes so heavy that the twigs become coated to a thickness of ½ to 1⅓ inch. The twigs are cut; lac is melted off, refined, and used in shellac and varnishes.

A group of small scale insects that typically live on desert cacti and resemble mealybugs are known as cochineal insects. *Dactylopius coccus* is the source of a natural crimson or scarlet dye called cochineal dye, originally used by the Indians of Mexico. Mature females are brushed from the cacti and dried; the pigments are extracted from the dried bodies. The Spanish used these dyes as early as 1518, and they were exported to Europe until replaced by the aniline dyes about 1870. The crimson colour of cochineal dye is attributed to cochinealin or carminic acid.

Sound production. At one time it was thought that the familiar call of male cicadas was the only sound produced by homopterans. It is now known, however, that sound production is common among other Auchenorrhyncha (leafhoppers, treehoppers, planthoppers, and spittlebugs) although their songs cannot be detected by the human ear unless amplified. Sound has been observed also in a few aphids and in one psyllid (both Sternorrhyncha).

The auchenorrhynchan Homoptera have evolved the most complex insect sound-producing mechanism known, the timbal organ. A pair of timbals, circular membranes supported by heavy chitinous rings, occur on the dorsolateral surface of the first abdominal segment. Contraction of a large timbal muscle attached to the membrane causes distortion of the timbal, producing a sharp click or pulse. The timbal springs back by its own elasticity when the muscle is relaxed. If the rates of muscle contraction and relaxation are rapid, the sound seems continuous to the human ear. The frequency contractions of the timbal muscles range from 120 to 480 per second. Associated with timbal organs in cicadas are large tracheal air sacs that open to the exterior and have resonant frequencies comparable to timbal vibration frequencies.

Although the timbal organ is similar in all species studied, the songs produced are variable; apparently this variation is caused by actions of tensor muscles and air sacs. The tensor muscles control pulse repetition frequency and sound intensity; abdominal movements control expansion and contraction of the air sacs and the consequent resonance frequency. The timbal is the only sound-producing mechanism evolved in the Homoptera except for a scraping of tibial stout hairs over reticulations of the abdomen of aphid colonies (*Toxoptera coffeae*); the sound produced is a rhythmic and synchronous scraping.

Cicada songs

Each cicada species has a characteristic song, often useful in identification. The analysis of periodical cicada songs has been the basis for morphological separation and determination of geographical range for several 13- and 17-year species. Thirteen-year species sometimes occupy parts of the same geographical areas as 17-year species. The primary song in cicadas, produced by the male, is a mating or pair-forming (aggregating) call. Female cicadas have no sound-producing organs. Males, attracted to calls of other males, stimulate each other to sing in chorus. Most mating takes place under these conditions. Male cicadas are stimulated to sing in the presence of tape recordings of songs of their species. Courtship songs or signals occur after pair formation or aggregating calls. Courtship interruption calls occur also for pair reforming. Calls produced when the insect is attacked, trapped, or in "distress" have been observed and are known as "dying yells."

Sound-producing organs occur in males of some cercopids, membracids, fulgorids, and cicadellids and in females of certain cicadellids. In *Doratura* both sexes have well developed sound-producing organs. The female of *Paropia* has a striated timbal that is poorly developed in the male. The sound-producing organ in the female is probably a primitive condition. Unlike cicadas, several leafhopper males produce calls in darkness and commonly produce mating calls when females are near. Rivalry calls between males also have been observed, usually accompanied by leg movements (kicks) that are attempts to strike and drive away a rival male.

EVOLUTION AND PALEONTOLOGY

Paleontologists do not agree on the exact or relative ages of either the Homoptera or Heteroptera (see HETEROPTERA). While some entomologists consider each group a separate insect order, others feel they have a common origin and classify them as suborders of the order Hemiptera. Although characteristics of the earliest Homoptera are not known, it is probable that the Protohomoptera had three tarsal segments, three ocelli, two pairs of wings about equal in size and shape with complete venation, an alimentary tract lacking a filter chamber, and male genitalia fitted with harpogones and subgenital plates.

Proto-homoptera

Based on the primitive nature of the ovipositor, the primitive sucking pump, and simple alimentary canal, the fulgorids are considered to be different from other Auchenorrhyncha and probably evolved earliest. The ovipositor of *Scolops pungens* is more primitive than that found in many of the Orthoptera. Thus the Fulgoridae are a combination of specialized sucking mouthparts and a primitive ovipositor.

The similarity of the thoracic sterna combined with jumping hind legs places the cicadellids, membracids, and cercopids together and differentiates them from the cicadas, which have different thoracic characters, lack the enlarged hindfemurs, and have a third (median) ocellus on the head. The Cercopidae show some relationship to the Cicadidae by having a complete tentorium in which the anterior tentorial arms are connected with the posterior arms; but in this respect they differ from the Cicadellidae and Membracidae, in which the tentorial structure is reduced. The hind legs of both Cicadellidae and Membracidae bear rows of spines that are absent on the hind legs of Cercopidae. Therefore, although the cicadellids, membracids, and cercopids are related and differ from other Auchenorryncha, the Cicadellidae and Membracidae are more closely related to each other than are the Cercopidae to either group. Furthermore, the cicadas and cicadellids, both of which retain different combinations of primitive characteristics, cannot be related through the cercopids, which lack these characteristics entirely. The cicadellids, in their structural and biological diversity, differ from other Homoptera and show a greater array of evolutionary stages in various combinations. Unlike other groups, cicadellids contain groups that stabilized at different evolutionary levels.

The Sternorrhyncha were probably separated from the Auchenorrhyncha as early as the Lower Permian (about 270,000,000 years ago). Although small in size, many fossil psyllids are found in the Upper Permian (about 230,000,000 years ago) strata and onward. If fossil remains have been identified properly, the aleurodids date from the Upper Permian also. They are highly specialized, both biologically and structurally. The aphids exhibit various degrees of polymorphism. The female geni-

talia are usually reduced; however, two groups, Adeligidae (part of the family Chermidae) and the family Phylloxeridae, have retained a true basic ovipositor. A fossil wing of *Permaphidopis sojaneusis* from the Permian resembles the wing of recent aphids.

Female coccids, wingless and sessile, are unlike the tiny winged males. Although the structural characters that have not been lost provide little information on phylogeny, they do show varying degrees of specialization. Most coccids, for example, have a single tarsal segment, but all species of *Xylococinae* have two; the male genitalia is a simple tubular, heavily sclerotized organ similar to the genitalia of aleurodids. The Peloridoidea represents a primitive form, probably a Paleozoic relic (about 225,000,000 to 570,000,000 years old). Highly specialized in some respects, it still has some primitive characteristics.

The Fulgoridae were the earliest group differentiated from the base of the Auchenorrhyncha stem; the cicadellids probably were next, apparently in Late Permian or early Triassic (about 220,000,000 years ago). The cercopids probably were derived from this stock; certainly the membracids were, later. The cicadas probably arose from the early cicadellid stem but are not found in fossils until the Cretaceous (about 65,000,000 to 136,000,000 years ago). If fossil forms have been properly placed, cicadellids and cicadas differentiated no later than the Permian; and cercopids and fulgorids must have had an even earlier origin. In the Sternorrhyncha group the psyllids, probably the earliest group to be differentiated, are known from abundant fossils in the Late Permian. The aleurodids also apparently arose in the Late Permian; the aphids date back to the Late Triassic (about 200,000,000 years ago); and the precretaceous fossil, *Mesoccus asiatica* Bekker-Migdisova, from the Permian seems to place the coccids in this geologic age.

CLASSIFICATION

Distinguishing taxonomic features. The beak and the wings are the most distinctive features of homopterans. The beak (or proboscis), fastened rigidly to the head, appears to arise from the ventral margin and consists of two pairs of stylets (mandibles and maxillae) adapted for piercing and sucking. In most homopterans, both pairs of wings are transparent or slightly thickened; the front pair have a uniform structure throughout. When at rest, the forewings are held rooflike over the dorsum with a slight overlapping on the inner margin near the tip. The digestive tract is complex, forming a filter chamber in most groups.

The suborders are distinguished by point of origin of beak, length and appearance of antennae, and number of tarsal segments. Separation of families of the Auchenorrhyncha is based on characters of the ocelli, position of antennae, form of pronotum, and spination of legs. Families of Sternorrhyncha are separated on the basis of number of tarsal segments, structure and venation of wings, and presence or absence of cornicles.

Annotated classification.

ORDER HOMOPTERA

Mostly small (4–12 mm); wings, when present, number two or four; sucking mouthparts; plant feeders; more than 32,000 species; worldwide distribution.

Suborder Coleorrhyncha

Origin of beak at antero-ventral extremity of face; propleura form a sheath for base of beak; hind wings absent; forewings held flat over abdomen when at rest; no flight function; prothorax with paranota; digestive tract lacks filter chamber.

Family Peloridiidae

Most primitive Homoptera; Tasmania, New Zealand, South America.

Suborder Auchenorrhyncha

Beak arises at antero-ventral extremity of the face, not sheathed by propleura; antennae with one to three basal segments, with a terminal seta; forewings rooflike when at rest; filter chamber present; all males apparently produce sound.

Family Cicadidae (cicadas)

Also called dog-day harvest-flies; usually large; three ocelli on face; front wings membranous; male with audible sound-producing organs on ventral base of abdomen, non jumping.

Family Membracidae (treehoppers)

Usually less than 12 mm in length; two ocelli; enlarged pronotum extends over head, thorax and all or part of abdomen; jumping hindtibia; wings largely concealed by pronotum.

Family Cercopidae (froghoppers or spittlebugs)

Less than 15 mm in length; two ocelli; jumping hindtibia with one or two stout spines, and circlet of stout spines at apex; hindcoxae short, conical.

Family Cicadellidae (leafhoppers)

Two ocelli or none; variable in size, 2 to 25 mm; jumping hindtibia with two or more rows of spines; hindcoxae transverse.

Family Delphacidae (planthoppers)

Hindtibia with broad movable apical spur; sexes often dimorphic, very different.

Family Derbidae (planthoppers)

Anal area of wing not reticulate; without cross veins; terminal segment of beak not more than 1½ times as long as wide.

Family Cixiidae (planthoppers)

Head not prolonged in front; carina of head median or absent; tegulae present; claval suture distinct, abdominal terga 6–8, rectangular.

Family Kinnaridae (planthoppers)

Terminal segment of beak at least twice as long as broad; front wings usually not overlapping as apex of clavus; head not prolonged in front; tegulae present; median ocellus usually present; abdominal terga 6–8, chevron shaped.

Family Dictyopharidae (planthoppers)

Head prolonged in front; or frons with two or three carinae or the tegulae absent; claval suture obscure.

Family Fulgoridae (planthoppers)

Second segment of hind tarsus large, apex with row of small spines; anal area of hind wing reticulate with cross veins.

Family Achilidae (planthoppers)

Terminal segment of beak at least twice as long as wide; claval vein extending to apex of clavus; body somewhat flattened; forewings overlapping at apex.

Family Tropiduchidae (planthoppers)

Second segment of hind tarsi with two apical spines, one on each side; apex reduced or conical; front wings longer than abdomen; cross veins between costal margin and apex of clavus.

Family Flatidae (planthoppers)

Costal and/or apical border of wing with numerous cross veins; wings longer than body, in repose held almost vertically at sides of body; clavus with numerous small pustule-like tubercles.

Family Acanaloniidae (planthoppers)

Also called Amphiscepidae; hindtibia without spines except at apex; front wings very broad, costal margin broadly rounded, venation reticulate; wings longer than body, at repose held almost vertically at sides of body.

Family Issidae (planthoppers)

Wings usually shorter than body, if longer than abdomen, usually oval; clavus without numerous small pustule-like tubercles; costal border of wings usually without numerous cross veins.

Suborder Sternorrhyncha

Beak appears to arise either between fore coxae or behind them; antennae usually long, filamentous, without a well differentiated terminal seta.

Family Psyllidae (jumping plant lice)

Beak long; mouthparts well developed in both sexes; tarsi two-segmented with two claws; antennae 5 to 10 (usually 10), segmented; front wings often thicker than hindwings, not exceeding 7 mm in length.

Family Aleyrodidae (whiteflies)

Very small; covered with a white powdery, waxy material; wings opaque; not jumping insects.

Family Aphididae (aphids or plantlice)

Wings membranous, Rs vein present in forewing; cornicles usually present; sexual females oviparous, parthenogenetic females viviparous; females and usually males with functional mouthparts; without abundant wax glands.

Family Eriosomatidae (woolly and gall-making aphids)

Aphididae in part; Rs vein present in forewing; cornicles indistinct or lacking; M vein in forewing not branched; wax glands usually abundant; sexual forms with the mouthparts atrophied and not functional.

Family Chermidae (pine and spruce aphids)
Adelgidae and Phylloxeridae in part; feed on needles, twigs, and leaves of conifers; Rs vein in forewing absent; cornicles absent; all females viviparous; Cu_1 and Cu_2 in forewing separated at base; apterous parthenogenetic females covered with wax.

Family Phylloxeridae (phylloxerans)
Chermidae in part. Rs in forewing absent; cornicles absent; all females oviparous; Cu_1 and Cu_2 in forewing stalked at base; apterous parthenogenetic females not covered with wax.

Family Margarodidae (giant coccids, ground pearls, cottony cushion scales)
Males with compound eyes and ocelli; anal ring reduced, without pores or setae; females wingless and legless.

Family Ortheziidae (ensign coccids)
Male with ocelli only; abdominal spiracles absent; anal ring distinct and flat, bearing many pores and 6 long setae; females wingless.

Family Diaspididae (armoured scales)
Apical segments of female fused, forming a pygidium; female with scale covering separate from body; legs absent; beak 1-segmented; antennae rudimentary.

Family Coccidae (soft scales, wax scales, tortoise scales)
Females flattened, elongate oval; exoskeleton hard, smooth, or wax covered; legs present or absent; antennae absent or much reduced. Females often tortoise-shaped; males winged or wingless; anus covered by two dorsal plates.

Family Aclerdidae (aclerdid scales)
Scales attacking grasses; openings of wax glands rarely 8-shaped; pygidium absent; male with ocelli only; anus covered by a single dorsal plate.

Family Lacciferidae (lac scales)
Females globular in form, legless; antennae 3- or 4-segmented, minute; body enclosed in cells of resin; tropical or subtropical.

Family Asterolecaniidae (pit scales)
Females without pygidium; beak with more than 1 segment; posterior end of body not cleft; abdomen not narrowed posteriorly or produced into an anal tube; wax gland openings 8-shaped, usually in rows; legs vestigial or absent.

Family Pseudococcidae (mealybugs)
Covered with a white powdery secretion; wax gland openings on dorsum, not 8-shaped; anal ring with four or more setae; dorsal ostioles and usually 1 to 4 circuli present.

Family Eriococcidae (scales)
Pseudococcidae in part; anal ring with 4 or more setae; dorsal ostioles and ventral circuli absent; body not covered with powdery secretion.

Family Dactylopiidae (cochineal insects)
Occur on cacti; abdomen not narrowed posteriorly; wax gland openings on dorsum; anal ring absent; wax gland ducts minute, arising from centre of cluster of sessile pores; setae stout and cut off at end.

Family Kermidae (gall-like coccids)
Females spherical, hemispherical or oval; legs absent in adult; antennae 6-segmented; anal ring absent; wax gland ducts not minute, openings not 8-shaped.

Critical appraisal. The Homoptera, along with the Heteroptera, are considered by many entomologists as suborders of the order Hemiptera, mainly on the bases of similar types of piercing–sucking mouthparts and on the general type of gradual metamorphosis. This system, however, places only minor importance on the distinct differences in structure and in details of metamorphosis that have led some workers to propose separate ordinal rank for Homoptera and Hemiptera and abandonment of the term Heteroptera. The mouthparts vary considerably in detail; in the Heteroptera the beak arises from the front of the head and is movable, while in the Homoptera the beak is fastened rigidly to the head, cannot be moved, and appears to arise from the dorsoventral portion. Little considered is the fact that certain Homoptera (*e.g.*, scale insects and whiteflies) pass through a stage in their development resembling complete metamorphosis. The names Heteroptera and Homoptera are derived from their different wings. In the Heteroptera only the apical portion of the wing is membranous and has visible veins; the basal portion of the wing is thickened and leathery. When these wings are at repose, they are held flat upon the dorsal portion of the abdomen; and the apical portions of the wings completely overlap. On the other hand, the forewings of the Homoptera are either membranous or of the same texture and contain visible veins throughout. At repose these wings are held at an angle rooflike over the abdomen, overlapping only slightly on the inner apical margin. Those who classify these two groups as suborders of the Hemiptera place only minor emphasis upon these distinct differences.

BIBLIOGRAPHY. D.J. BORROR and D.M. DELONG, *An Introduction to the Study of Insects* (1970), a modern text dealing with the orders and families of insects; C.T. BRUES and A.L. MELANDER, *Classification of Insects*, 2nd rev. ed. (1954), taxonomic treatment of orders and families; H.H. ROSS, *A Textbook of Entomology*, 3rd ed. (1965), general discussion of morphology, physiology, and embryology; E.O. ESSIG, *College Entomology* (1942), comprehensive text; A.D. IMMS, *General Textbook of Entomology*, 9th ed. rev. by O.W. RICHARDS and R.G. DAVIES (1957), comprehensive morphological and taxonomic treatment; R.E. SNODGRASS, *Principles of Insect Morphology* (1935), comparative internal and external morphology of insects; Z.P. METCALF, *A Bibliography of the Homoptera (Auchenorrhyncha)*, 2 vol. (1943), *General Catalogue of the Homoptera* (1954–68), *Fulgoridae of Eastern North America* (1923), taxonomic treatment with keys to families and excellent illustrations; K.C. DOERING, "Synopsis of the Family Cercopidae (Homoptera) in North America," *J. Kans. Ent. Soc.*, 3:53–64, 81–108 (1930), a synoptic treatment of all known North American cercopids; P.B. LAWSON, "The Cicadidae of Kansas," *Kans. Univ. Sci. Bull.*, 12: 309–376 (1920), taxonomic and distributional discussion of Kansas cicadas; P.W. OMAN, *The Nearctic Leafhoppers (Homoptera: Cicadellidae): A Generic Classification and Check List* (1949); H. OSBORN, *The Membracidae of Ohio* (1940), taxonomy of treehoppers of Ohio and proximity; D.L. CRAWFORD, *Monograph of the Jumping Plant Lice or "Psyllidae" of the New World* (1914), a taxonomy of jumping plant lice of North America; G.F. FERRIS, *Atlas of the Scale Insects of North America*, 4 vol. (1937–42), a comprehensive study of genera and species, excellent illustrations; A.D. MACGILLIVRAY, *The "Coccidae"* (1921), synoptic treatment of scale insects; F.C. HOTTES and T.H. FRISON, *The Plant Lice or Aphididae of Illinois* (1931), aphid taxonomy; J.W. EVANS, *A Natural Classification of Leafhoppers (Jassoidea, Homoptera)*, 3 pt. (1946–47), a taxonomy of leafhoppers with emphasis on higher categories.

(D.M.DeL.)

Homo Sapiens

In the past it has often been the task of philosophers, theologians, politicians, and ideologists to define true man, to describe the human condition, and to set the boundaries of humanity. On the whole, such writers have been undeterred by their frequent lack of scientific information about man, with the result that their conclusions were often biassed or based upon false premises. But since the 1920s, with the rapid progress made in the field of human evolution, the problem of defining true man—*Homo sapiens*—has attracted the attention not only of anthropologists and anatomists but also of geneticists, biochemists, biostatisticians, taxonomists, and a number of other biological specialists. This concentration of scientific effort and expertise has resulted in a critical reappraisal of the typological concept of evolution, which depends on the classification of individuals according to a narrow definition of the species type. This concept has largely been replaced by the idea that a population, not an individual, is the evolving unit. When this distinction was made clear, differences between the fossil remains of individual prehistoric human beings fell into perspective, aided by modern knowledge of the range of normal variability in human populations. Still the realization that populations, themselves, vary widely and cannot be narrowly typed brings us face-to-face with the problem of classification, a problem foreseen by Charles Darwin, who wrote in his *Descent of Man:*

> In a series of forms graduating insensibly from some ape-like creature to man as he now exists, it would be impossible to fix on any definite point when the term man ought to be used.

Nonetheless, several attempts have been made to establish a point in his evolution at which man's prehistoric

ancestors are to be considered human beings (*Homo sapiens*) by making use of a so-called human attribute as an absolute criterion.

Scottish anatomist and physical anthropologist Sir Arthur Keith regarded the relatively large size of the human brain as of paramount importance, arguing that mental activities of a higher order would result and that these activities, in turn, would lead to more successful sexual, maternal, and social behavioral patterns. Thus, an arbitrary line was drawn: ". . . any group of the great primates which has attained a mean brain volume of 750 cc. and over should no longer be regarded as anthropoid, but as human." This was Keith's "Rubicon 'twixt ape and man." Such a view could not last, however, because man cannot be defined by any single anatomical criterion.

A more plausible approach was suggested by the British paleontologist Kenneth P. Oakley, who made use of a primarily behavioral criterion to define man: the ability to make tools. The systematic making of tools for future as well as present use implies a capacity beyond that of modern apes, who will freely improvise tools from sticks or branches but will nearly always use them to solve by perceptive means an immediate and visible problem. Clearly "man the toolmaker" as a definition involves not only a hand structure capable of toolmaking and tool use but also a brain with a capacity for conceptual thought. Here again, functional anatomy and behaviour are inseparable and highly correlated.

Taxonomic definition of *Homo sapiens*

The problem here, however, is not to define humanity but *Homo sapiens*, the human species, and it was the British anatomist W.E. Le Gros Clark who emphasized the importance of making a clear distinction between colloquial terms and Latin names of definable zoological meaning, which follow the rules of the International Code of Zoological Nomenclature. Le Gros Clark offered a definition of *Homo sapiens* that makes particular use of specific anatomical criteria that relate to the skeleton, because the principal use of such a definition is in the evaluation of fossil bones. Thus, *Homo sapiens* was provisionally defined by him in precise anatomical terms:

> A species of the genus *Homo* characterized by a mean cranial capacity of about 1,350 cc.; muscular ridges on the cranium not strongly marked; a rounded and approximately vertical forehead; supra-orbital ridges usually moderately developed and in any case not forming a continuous and uninterrupted torus; rounded occipital region with a nuchal area of relatively small extent; foramen magnum facing directly downward; the consistent presence of a prominent mastoid process of pyramidal shape (in juveniles as well as adults), associated with a well-marked digastric fossa and occipital groove; maximum width of the calvaria usually in the parietal region and the axis of glabello-maximal length well above the level of the external occipital protuberance; marked flexion of the sphenoidal angle, with a mean value of about 110°; jaws and teeth of relatively small size, with retrogressive features in the last molars; maxilla having a concave facial surface, including a canine fossa; distinct mental eminence; eruption of permanent canine commonly preceding that of the second molar; spines of cervical vertebrae (with the exception of the seventh) usually rudimentary; appendicular skeleton adapted for a fully upright posture and gait; limb bones relatively slender and straight (from W.E. Le Gros Clark, *The Fossil Evidence for Human Evolution;* The University of Chicago Press, © 1955, © 1964).

A definition of this kind attempts to take account of the total morphological (structural) pattern of the species and pinpoints anatomical features of supposed taxonomic relevance or groups of features that form a functionally significant complex; *e.g.*, a prehensile hand or a bipedal foot. Functional affinity may be deducible from structural similarity, and degrees of functional affinity are quantifiable. Such estimates have been shown to be of particular value when calculated from numbers of measurements that are subjected to multivariate statistical analysis. It must be added, however, that, although the aim of such studies is to try to provide an objective measure of taxonomic (classificatory) affinity, the choice of measurements for analysis is subjective; but with that proviso the choice of functionally significant dimensions will lead to a functional discrimination that, in turn, may be taxonomically significant. Nevertheless, the mere demonstration of significant "morphological distance" between populations need not mean that they must inevitably belong to different species.

The term *Homo sapiens* is a proper Linnean biological term the use of which is governed by rules of priority as set forth by the International Code of Zoological Nomenclature. Attribution of a new fossil find to the group must take account of the definition of both the genus and the species. But the weakness of typological taxonomy of this kind (where a narrow definition of human characteristics is given) is obvious when samples of fossils are small and specimens are incomplete; furthermore, the problem is compounded when closely related or transitional forms are being considered. Characteristics may be found that are typical of less evolved specimens; yet, at the same time, more advanced features may be present in the same specimen. This situation is the result of mosaic evolution—the idea that not only do different populations evolve at different rates but also that different individual parts may evolve differentially. It follows that mosaic evolution will cloud the issues of classical Linnean taxonomy, because the Linnean system depends upon the identification of a set of clearly defined characters of taxonomic relevance occurring together according to an agreed definition.

Mosaic evolution

It is becoming increasingly clear that early *Homo sapiens*, evolving relatively rapidly during the late middle Pleistocene Epoch (about 250,000–100,000 years ago), was subject to this mosaic process.

FOSSIL REMAINS OF HOMO SAPIENS

The sapient skeletal system. Fossil remains of early *Homo sapiens* are known from sites in both Europe and Africa, but later examples come from a wide range of sites in the Old World as a whole. It is of particular interest to look at the skulls of some of the early specimens, for it is in their functional morphology that the combination of features that results in attribution to *Homo sapiens* is often found.

Evolution of the human skull. The human skull is composed of both cranial and facial portions. The cranium consists of the skull vault and base, while the facial skeleton consists of the region of the eye sockets, nose, cheekbones, upper jaw (orbital and maxillary region), and the region of the lower jaw (mandibular region). During the evolution of the hominid skull from its apelike precursors, there are a number of general trends that can readily be discerned. The principal trends are the gradual increase in brain size (as measured by cranial capacity), the rounding of the cranial vault, and the gradual reduction of the size of the whole masticatory complex, including both the upper and lower jaws and the teeth. These trends lead to an overall change in skull shape and proportions, so that, while the vault expands, the "muzzle" tends to retract from a protruding (prognathic) form to a straighter-faced (orthognathic) appearance. At the same time, the whole skull tends to become lighter and more delicate in its structures. If this process is examined a little more closely, it can be at least partly explained in mechanical terms. A comparison of the skull of *Homo erectus* (the species of man immediately preceding *Homo sapiens* in the course of human evolution) and a modern *Homo sapiens* skull illustrates the point (Figure 1).

The skull of *Homo erectus* (the species lived from approximately 1,000,000 years ago until about 200,000 years ago) has a low skull vault, a sharply receding forehead, a low cranial capacity (800–1,100 cubic centimetres [cc]), a large face with big teeth and jaws, and a jutting occipital region (the back of the head). Muscles in the occipital region relate principally to the balance of the head on the vertebral column (spine); those of the temporal region (at the sides of the head) relate to the working of the jaw apparatus. In a skull with a large face and a heavy jaw, the muscles of the occipital region, which balance the head on the neck (nuchal musculature), must be strong in order to counterbalance the heavy face and jaw. This results in marked occipital

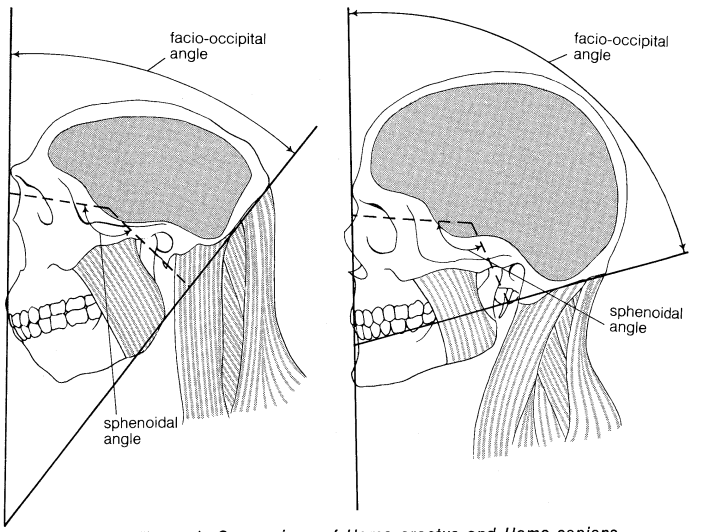


Figure 1: *Comparison of Homo erectus and Homo sapiens skulls, showing differences in cranial capacity, sphenoidal angle, and facio-occipital angle.*
(Left) Small brain, large muscles, and flattened skull vault of *Homo erectus* skull. (Right) Large brain, small muscles, and rounded skull vault of *Homo sapiens* skull.

ridges in *Homo erectus*. Similarly, large jaws and teeth demand well-developed muscles of mastication with strong attachments to the cranium. In order to dissipate the considerable forces produced by these muscles when the teeth meet during chewing, there is in *Homo erectus* a stout maxilla (upper jaw) and a heavy supraorbital ridge, or torus (brow ridge), extending the whole width of the cranium, a ridge rendered even more necessary because *Homo erectus* had a skull of low cranial capacity and therefore lacked a well-developed forehead. Such skulls are found in the examples of *Homo erectus* known from both Peking and Java (see HOMO ERECTUS).

The skull of *Homo sapiens*

By comparison, the skulls of *Homo sapiens* show an expanded cranial vault with a high maximum breadth and a well-developed vertical forehead, resulting from expansion of the frontal region of the brain, so that the supraorbital ridge is, as it were, "overgrown" and buried. At the same time, the face is shortened by a reduction in size of the jaws, which also bear smaller teeth. The masticatory muscles, therefore, need not be so large; and the forces that they produce are less, so that the need to neutralize them through the face is reduced. Thus, the facial structure can be more delicate, and the need for a heavy supraorbital ridge is removed. Finally, the lighter face does not require such powerful muscles behind the point of balance of the skull; as a result, the neck muscles can be reduced in size and their areas of attachment to the skull brought more underneath the back of the head.

The decrease in the size of the teeth of *Homo sapiens* tends to leave the nose and point of the jaw as prominent features of the face. Thus, the presence of a mental protuberance (chin) is an obvious character of the jaw of *Homo sapiens* and provides external support for the symphysial region (the point at which the two sides of the jaw have grown together).

Dental characteristics. The dental characteristics of *Homo sapiens* revolve around the basic fact of reduction of the masticatory apparatus. Thus, the dentition as a whole shows tooth crowding (dento-alveolar disproportion), accompanied by smallness of the individual teeth and marked reduction in size of the third molar. In modern populations the third molar tends to be genetically unstable; *i.e.*, it is frequently absent or malpositioned (impacted wisdom teeth). Similarly, but less frequently, the lateral incisors (the cutting teeth on either side of the "front teeth") may be absent. In other respects the dentition of *Homo sapiens* is best contrasted with that of *Homo erectus* in that it lacks some of the special features known from this species. For example, the teeth of *Homo sapiens* are less likely to show secondary enamel wrinkling of the occlusal (chewing) surface of the teeth, and less taurodontism (pulp space enlargement), although this is still well-known from Neanderthal man. The eruption sequence of teeth seems to be an unreliable criterion because variability is common in modern populations and may well have been so in the past. The teeth of *Homo sapiens* are small, unspecialized, crowded, and liable to be genetically unstable, particularly the wisdom tooth.

Postcranial skeleton. The form of the skeleton of the trunk and limbs of *Homo sapiens* (postcranial skeleton) is characterized by its adaptation for a fully upright posture and a striding bipedal gait. This remarkable locomotor capability is the final expression of an evolutionary process that has taken at least 2,000,000 years to achieve, and so some aspects of the process are well-known from earlier members of the genus *Homo* and also from the genus *Australopithecus* (which may have been ancestral to the genus *Homo*).

Bipedal posture and gait

In terms of posture, the bipedal vertebral column is held upright and shows two secondarily developed curves when viewed from the side, one in the lumbar region of the back (small of the back) and the other in the neck region. From the front the column should appear straight. These curves allow the weight to be evenly disposed about the line of gravity, which passes vertically through the second sacral vertebra (at the base of the spine) and behind the rotation centres of the two hip joints. This permits the pelvis to tip backward just beyond the vertical and rest upon a straplike ligament across the front of the hip joint, a sophisticated effort-saving mechanism that allows most of the muscles around the hip to relax so that the upright stance is an economical posture. Associated with this is the ability to lock the knees back, which also relaxes some surrounding muscles. To rise from the squatting or seated position requires considerable power of extension of the hip joints, and this is provided by the large buttock muscle (gluteus maximus) and a backward extension (posterior superior iliac spine) of the bony pelvic flange (blade of the ilium) for its attachment.

An alternating bipedal gait, to be fully efficient, must allow each leg to swing clear of the ground during walking; this is provided for by a pelvic-tilt mechanism that raises the side of the swinging leg. In addition, such a gait must avoid wild side-to-side movements of the centre of gravity, and this is achieved by inclining the thigh bones toward the midline and thus bringing the feet closer together. Finally the bipedal adaptations of the modern human foot are such that both weight and force are transmitted to the ground through a propulsive system of short levers that permits a heel–toe stride.

The upper limb adaptations to bipedalism are fewer and are concerned with the dynamic balance of the body while moving. Arm swinging is a normal part of bipedalism and compensates for the twist of the body toward the side opposite to the advancing foot. The selective advantages of bipedalism, in terms of the upper limb, are immense in that they free the hands for the carriage of infants, food, tools, or weapons, as well as permitting the development of the hands for a manipulative role such as toolmaking. Although hominids below the human level of evolutionary advance could make tools, the refinement and exploitation of tools demanded hands capable of both power and precision grips. The power grip involves primarily the inner, or ulnar, side of the hand and permits a firm grip on a branch, a rock, or hammer handle. The precision grip involves the outer, or radial, fingers and thumb, as in using a small stone for engraving, a small brush for painting, or a pen for writing (Figure 2). This requires the bringing together of the tips of the opposed thumb and the next two fingers in order to grip a small object, a grip that demands that the lengths of the index finger and thumb be proportionate and that the joint at the base of the thumb be of a special saddle-shaped variety. It seems likely that the precision grip evolved later than the power grip and that its perfection may even have been a specialization in *Homo sapiens*. Only when human hands had evolved to this level, concomitantly with brain expansion, could manipulative

Freeing the hands for tool use

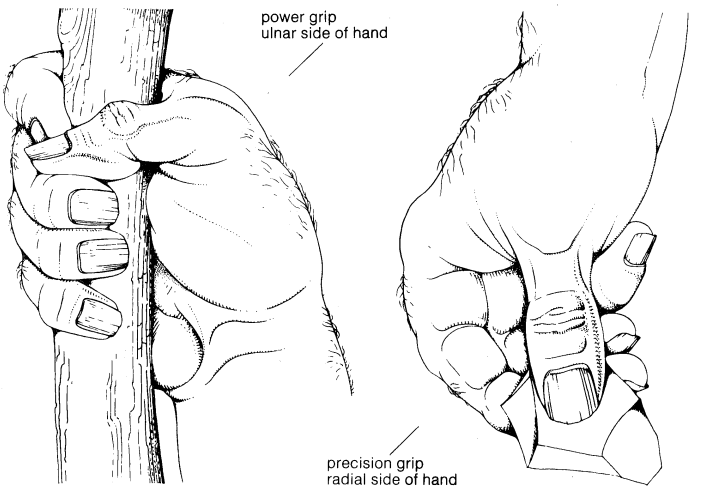


Figure 2: Hand grips.

skills give expression to the artistic impulse in terms of cave painting, bas-relief, and sculpture in the round, all of which are sophisticated behavioral correlates of a highly evolved individual, in terms of both locomotor and intellectual skills.

Having looked at the problems of the identification of human characteristics and also at the difficulties of interpreting the significance of these characteristics singly or in groups, it is appropriate to consider the early fossil materials that might provide reasonable candidates for attribution to *Homo sapiens*. These are likely to be found during the late middle Pleistocene to upper Pleistocene (about 200,000 to about 15,000 years ago)—in archaeological terms, the Paleolithic, or Old Stone Age. Naturally the quantity of fossil material recoverable tends to increase markedly as more recent periods are dealt with, so that the total number of human specimens amounts to some hundreds. On the other hand, toward the earlier end of this time range, the fossils available for study are far fewer in number. Advances in dating methods in recent years have eliminated a number of specimens formerly given an antiquity that they do not deserve, so that the picture has been somewhat clarified. A number of specimens remain that have often in the past been divided into the Neanderthals and the sapients (*i.e.*, members of *Homo sapiens sapiens*, the same subspecies as modern man).

The problem of Neanderthal man. This group was originally recognized by a combination of cranial, dental, and postcranial features that were generally considered distinct enough to classify Neanderthal man as a separate species. Thus *Homo neanderthalensis* was described on morphological grounds and regarded as a specialized group of the genus *Homo* that lived during the last glaciation (Würm Glacial Period) in Eurasia. The sites from which examples of the Neanderthalers were recovered have commonly produced tools of the Mousterian culture, a stone-tool industry dating from about 90,000 to about 40,000 years ago. Quite suddenly these people disappeared from the fossil record, and various theories have been put forward to account for their disappearance. In addition to these "classic" Neanderthalers, exemplified by remains from Neanderthal (Germany); La Chapelle-aux-Saints, Le Moustier, La Ferrassie, La Quina (France); Gibraltar; Monte Circeo (Italy); Shanidar (Iraq); Kiik-Koba (Russia), and many other sites, there is a second group sometimes known as the generalized Neanderthalers. These have included remains from Krapina (Yugoslavia), Saccopastore (Italy), Ehringsdorf (Germany), and Skhūl (Mt. Carmel, Israel). Some of these preceded or at least were contemporaneous with the classic Neanderthalers, whereas others of this second group appear to have existed rather later, since reappraisals of the dates first given to some of the sites have questioned their antiquity; *e.g.*, Krapina, Mt. Carmel.

Many have considered that the classic Neanderthalers were a cold-adapted, specialized side branch from the human line that became isolated in Europe and then became extinct as the climate improved. The generalized Neanderthal group is considered to have avoided this specialization and continued to give rise to the later modern sapient populations. This involves invoking the so-called catastrophic demise of classic Neanderthal man. An alternative view placed classic Neanderthal man in the modern human evolutionary line, his disappearance both anatomically and culturally being due to an absorption process involving some contribution of Neanderthal genes to the succeeding populations. Most modern systematists tend to include Neanderthal man within the species *Homo sapiens* and only accord subspecific status to the combination of morphological characters that make up the anatomy of the classic Neanderthalers of the last (Würm) glaciation; they are thus named *Homo sapiens neanderthalensis* rather than *Homo neanderthalensis*.

Neanderthal man's place in human evolution

In addition to the European and Near Eastern evidence of Neanderthal man, a number of specimens from Africa and Asia must be considered. These have been termed "Neanderthaloid" in the sense that, although they show some cranial characters that parallel the better known European type of classic Neanderthal structure, they also show a number of other (sometimes more advanced) features of the skeleton. Examples of this group would include Rhodesian man (Zambia, formerly Northern Rhodesia), Solo man (Java), and Saldanha man (South Africa), all from upper Pleistocene deposits (100,000 to 10,000 years ago). Their very presence shows that the whole problem of defining *Homo sapiens* must be considered in a wider context and cannot be reasonably considered in European terms alone. Clearly, the problem of the origin of *Homo sapiens* from his middle Pleistocene forebears is complex; hence, it is valuable to examine in detail those specimens that come from the earliest well-dated sites, in order to try to discern the centre of sapient evolution (if such a centre exists) and to discover which are the earliest specimens that show incipient sapient characters. Bearing in mind the principle of mosaic evolution, which has been demonstrated as an important part of the evolutionary process, it is not reasonable to suppose that all of the features that have been mentioned as characteristic of *Homo sapiens* will appear together in these early specimens. Thus, a mixture of advanced and less advanced features may be expected in early forms, some perhaps relating to *Homo erectus* (the species from which *Homo sapiens* descended), to Neanderthal man, or to modern man as he is today.

Vértesszőllős man. In 1965 some fossil remains were recovered from a site in Hungary at the foot of the Gerecse Mountains. In a quarry cut into the fourth terrace of the Danube River system, a number of occupation layers were recognized. The third of these layers contained some human remains as well as a tool culture and some fossil mammalian bones. The site is known as Vértesszőllős, and the first finds were some fragments of milk teeth from the lower dentition of a child (Vértesszőllős I). Rather more important was the second find (Vértesszőllős II), a fine adult occipital bone (the bone at the back of the skull) that was broken into two fragments. The bone is stoutly constructed, having thick walls in the ear region, whereas the floor of the depression for each cerebellar hemisphere (at the base of the skull) is relatively thin. In profile, the bone can be divided into two parts, an upper, curved occipital portion and a lower, flattened neck portion. The flattened area is incomplete and does not include the margin of the foramen magnum (the opening at the base of the skull through which the medulla passes), the borders of which are broken away. The attachments of the neck muscles are well marked on the outer surface of the flattened area, which also has signs of an occipitomastoid crest (a crest of the occipital bone near the mastoid process, which is a projection of bone behind the ear), a primitive feature. The occipital ridge, or torus, that divides the upper, curved portion from the lower, flattened area is prominent and continuous across the bone; indeed, it is so prominent that

Features of cranial remains

in its central portion it is even undercut. Internally, the cerebellar fossae (depressions in the skull where the lobes of the cerebellum lay) are rather small by comparison with the impressions above for the occipital lobes of the brain. The internal occipital protuberance lies well below the external occipital protuberance (inion), and the impressions for the venous sinuses (marks where blood vessels lay against the skull) are distinct. The cranial volume of this individual has been estimated at 1,400 cubic centimetres.

The remains have been attributed to a male under 30 years of age. The thickness and the breadth of the bone, as well as the undivided occipital ridge, are relatively primitive features, whereas the height and curvatures of the upper segment are modern features. Similarly, the configuration of the brain is primitive despite its large size. The morphological comparisons and the statistical analysis of this specimen suggest that, although this population took its origin with *Homo erectus*, it had differentiated from this group and perhaps ought to be classified as at the beginning of a progressive evolutionary line.

Perhaps the principal significance of this specimen, apart from its structure, is that it is dated to a warm phase within the second (Mindel) glaciation, 500,000 to 400,000 years ago. The specimen has been assigned to an unusual new category *Homo (erectus seu sapiens) palaeohungaricus;* whatever the precise taxonomic meaning of this name, it is clear that it is regarded as an intermediate form, the mixed *Homo sapiens* and *Homo erectus* features of which are the result of evolutionary mosaicism.

Swanscombe man. A more widely known example of early *Homo sapiens* is derived from the Thames Middle Gravels at Swanscombe in North Kent, England, which are gravels attributed to the Mindel-Riss Interglacial Stage, 400,000 to 200,000 years ago. Here, at a site well-known for its Paleolithic tools, two parts of a human skull were recovered in 1935 and 1936. About 20 years later, a third piece was recovered that fits with the other two pieces to form the back half of a cranial vault. All three bones are virtually complete, and it is believed that they belonged to a young adult. Generally speaking, the bones are modern; however, the bones of the skull vault are rather thick, and the skull is broad at the back. The occipital bone shows no sign of the heavy occipital torus, or "chignon," characteristic of Neanderthal man, and the occipital ridge is modest, although a little more prominent at its ends than in the middle. This bone is rounded both above and below the ridge, in contrast with the Vértesszőllős occipital. The position of the foramen magnum and the joints for the vertebral column are modern, so that the posture of the head does not appear to differ significantly from that of modern man. In the original report on these bones, it was stated that by measurements the Swanscombe skull could not be distinguished from that of *Homo sapiens*. Despite this, several authorities have been less convinced by the sapient features of Swanscombe man and have emphasized the Neanderthal features of the bones. In a recent reappraisal of the Swanscombe material, making use of multivariate statistics, it has been suggested that it is necessary to emphasize the inter-relatedness of all the early forms of *Homo* and that it is no longer possible to maintain specific status for each form. It is preferred, rather, to regard them all as a spectrum of varieties within one species. It was suggested further that Swanscombe should be regarded as belonging to a "Neanderthaloid Intermediate" group that contains the Ehringsdorf, Skhūl V, Krapina, and Steinheim specimens.

Classifying Swanscombe man

Once again, as with Vértesszőllős, there is a mixture of features that, at this somewhat later date in Europe, combine sapient and Neanderthal characteristics in keeping with the concept of mosaic evolution.

Steinheim man. Another skull of approximately the same age as the Swanscombe skull is known from Steinheim, near Stuttgart, West Germany. This skull was recovered in 1933 from a gravel pit cut into Pleistocene deposits that have been dated to the Mindel-Riss Interglacial Stage. This skull is more complete than either of those mentioned above but suffers from some distortion, perhaps due to the pressure of the deposits. It consists of the cranium and the right side of the face of a young adult; much of the base of the skull is missing, but there is a good deal of the palate and also some teeth present, including a premolar and all the molars of both sides. The face is straight, with little projection of the upper jaw, while the vault of the skull appears to be long and narrow but fairly well rounded in profile, and there is only moderate frontal flattening. The occipital region is marked by a very low occipital ridge, and it does not have a Neanderthal "bun." The frontal region has a pronounced but divided brow ridge. The cranial volume has been estimated variously between 1,070–1,175 cubic centimetres, but the distortion precludes accurate measurement.

The teeth are of particular importance because they are perhaps the earliest sapient teeth known. The premolar crown is symmetrical in shape but has a large outer and a smaller inner cusp. The molars decrease in size order from front to back, but the third molar is markedly smaller than the other two. All of the teeth have some degree of taurodontism (enlarged pulp cavities).

Again, there is a mixture of primitive and advanced features in this skull, the principal resemblances of which are, with the Swanscombe skull, described above. The contour of the occipital region and the rounded vault are advanced sapient features, whereas the broad nasal opening and the brow ridge recall the Neanderthal shape. The teeth are small and sapient both in form and molar size order, while the marked reduction of the size of the third molars is a distinctly modern feature. The enlarged pulp cavities are sometimes regarded as primitive since they are widely known from both *Homo erectus* and Neanderthal teeth. However, this trait is still frequently observed in the teeth of modern populations.

Taxonomic affiliations of Steinheim man

Earlier opinions have suggested that Steinheim man represented an intermediate stage between *Homo erectus* and the later forms of *Homo*, thus being ancestral to both Neanderthal and modern man. Another view does not accept the Steinheim fossil as being ancestral to the classic Neanderthal and modern man, suggesting, instead, that it represents a stage on a separate but parallel line leading to modern man. Both views now appear to be too simplistic and not in harmony with the findings of modern population genetics; the pattern of mixed characters that is becoming apparent follows the general trend of cranial enlargement and rounding, accompanied both by facial and dental reduction.

Fontéchevade man. The fourth early sapient skull known from Europe is less well preserved than the preceding specimens. It consists of some vault fragments that are known from Fontéchevade, Charente, France. The material was recovered during the excavation of a cave site that had produced tools from higher levels. The fossil layer has been attributed to the last (Riss-Würm) interglacial period because of the fauna and the tool culture that it contains. The first specimen (Fontéchevade I) is a small fragment of rather thin frontal bone, but the second specimen is large and includes both parietal bones and part of the frontal bone. French anthropologist H.V. Vallois, who has studied the remains, denies any difference between these remains and those of *Homo sapiens;* he makes it clear that there was no supraorbital ridge and that the cranial volume was over 1,400 cubic centimetres. It is only fair to say that the Fontéchevade remains are fragmentary, and it would be a mistake to place too much reliance on their evidence alone; nonetheless, Vallois has taken the view that Swanscombe and Fontéchevade are related types at the base of a progressive line leading toward modern man, whereas Steinheim would have led to the more specialized Neanderthal group.

Thus, in Europe there is one specimen from the Mindel Glacial Period (500,000–400,000 years ago), with early sapient features; two examples of early *Homo sapiens* from the Mindel-Riss Interglacial Stage (400,000–200,000 years ago); and at least one from the Riss-Würm Interglacial (100,000–70,000 years ago). If this is re-

garded as scanty evidence, then that from Africa has been even more sparse until the most recent finds.

Omo man. In 1967 the Kenyan group of the International Omo Expedition to southern Ethiopia, led by Kenyan paleoanthropologist Richard Leakey, recovered large parts of two skulls and a large number of limb bones from two sites that have been dated to the East African upper middle Pleistocene or early upper Pleistocene Epoch (about 200,000 to 70,000 years ago). The most complete skull (Omo II) is long and narrow with a receding forehead, a rounded cranial vault, and a prominent occipital ridge, below which lies a flattened neck region. The bones of the skull vault are thick, and the maximum breadth of the skull is low in the temporal region, which is marked by the presence of large mastoid processes (bony projections below and behind the ear). The completeness of this skull has allowed accurate estimations of the cranial volume at 1,430 cubic centimetres. Preliminary assessment of the taxonomic affinities of this skull has suggested that, while it shows a number of specialized features, some of which are reminiscent of the earlier *Homo erectus* skulls, it also has some advanced features, such as the rounding of the vault, the large mastoid processes, and the high cranial capacity. In view of this, it has been regarded as an early example of the African segment of evolving *Homo sapiens.*

Features of the Omo skulls

The other skull (Omo I) recovered from the same general area and attributed to the same geological layer was accompanied by a large number of limb bones and is in many ways quite different from Omo II. Although the forehead still slopes from a prominent supraorbital (brow) ridge, the vault is very well rounded down to the occipital region. There is a prominent downturned mastoid process. The vault of this skull is robust by modern standards, but it is less so than that of Omo II.

Viewed from the rear, the two skulls show striking differences—the Omo I skull is far more modern in outline. The limb bones of Omo I are rugged and well marked by muscle impressions, but in general they have few, if any, features that would distinguish them from the limb bones of *Homo sapiens.* A preliminary assessment of Omo I therefore suggests that, although it has some specialized features, they are far fewer than those of the Omo II skull; at the same time, its advanced features are perhaps more marked.

Perhaps the most striking feature of the whole Omo assemblage is once again the mixed character of each of the two principal specimens that are said to be of the same geological age. The Omo II skull undeniably has features that can be paralleled in earlier forms of *Homo,* and at the same time the Omo I specimen seems to foreshadow the later morphology of modern *Homo sapiens,* particularly in the occipital region of the skull and in the limb bones. It is possible that the mosaic evolutionary process is seen here in microcosm. Not only is differential evolution occurring within individual representatives of contemporary populations but also between populations in one area of East Africa.

Obviously, the specimens considered here do not comprise all of the available material that has been or could be attributed to fossil *Homo sapiens;* much of the remainder is of a rather later date, however, such as the Rhodesian remains, the Solo skulls, the Wadjak skulls, the Upper Cave remains from Peking, and materials from a number of other sites. Only a few taxonomists would still say that these remains merit species status on their own (*e.g., Homo rhodesiensis, Homo soloensis, Homo wadjakensis*); the majority accept that they all belong to the same species as modern man, *Homo sapiens.* Some would permit classification according to subspecies or variety, mostly on a combined geographic, temporal, and morphological basis. This implies that *Homo sapiens soloensis* or *Homo sapiens steinheimensis* are representatives of a local population from a given time period in a given area, similar to modern racial groups. The aggregate of the populations represented by fossil remains and all modern populations is a polytypic species (one that cannot be represented by a single typical individual because of physical variation such as skin colour and size of brow ridges within the species), valid both "horizontally and vertically"; *i.e.,* both geographically and through time. Classification by subspecies is no doubt useful in hominid taxonomy, at least for the purposes of communication and description, but whether or not one or two specimens, distinguished by a few morphological characteristics of the bones alone, truly form a valid basis for attribution to a separate subspecies in zoological terms is less certain.

Classifying the later human remains

In examining the earliest specimens that have been attributed to the species *Homo sapiens,* it is apparent that they comprise a very small sample of the populations that must have existed. Their diversity may perhaps be seen in perspective by considering the wide differences that exist in mankind today, differences that are certainly all accommodated within one species.

It seems that the evolution of man must be considered in terms of successive evolving populations, each showing a range of structural and geographic diversity at any given time, a diversity due to groups that have evolved similar but not identical characteristics in different places. Not all of these evolutionary lines necessarily developed into forms that exist today, but neither was the modern line necessarily genetically isolated from the past populations represented in the fossil record.

CULTURAL ASSOCIATIONS OF FOSSIL HOMO SAPIENS

While bones and teeth provide the most direct evidence of the form and locomotor capabilities of fossil man, there are other sources of information from which to obtain some concept of both his surroundings and his activities. This information comes from the careful excavation of the geological deposits from which the fossil remains are taken. Modern methods in archaeology make use of an increasing number of scientific techniques, and through these studies it is sometimes possible to reconstruct both the environmental situation and the behavioral activities of fossil man at a given site and time period.

Cultural remains preserved in the fossil record. In considering early *Homo sapiens* from a cultural standpoint, there are a few general considerations to be taken into account. The evidence that can be found must be inherently capable of preservation; thus, the types of materials that are most likely to be found are stone tools or stone structures and often fossilized animal bones and teeth. On the other hand, wood and plant remains are much rarer because such materials are more subject to decay. It must not be assumed, therefore, that the absence of wooden materials in the fossil record implies that they were not used by early man. Materials not normally preserved may be present in some situations and provide valuable clues; hence, fossil pollens, dealt with by the science of palynology, can give enormous insight into the vegetational structure and thus the climate of the time. Other general considerations relating to a fossil site include the likelihood of its location being close to a source of water, both for man and the animals that he may have hunted. Similarly, it is likely that sources of stone for the manufacture of tools will be found near living sites. Given these basic requirements, the living sites of early man can clearly be either in temporary structures in the open, in caves, or in rock shelters. In many circumstances it is likely early man had little choice about where to shelter, but under glacial conditions the ability to find, acquire, and defend a cave or shelter could have been crucial and might have overridden other considerations. During interglacial periods, when warmer conditions prevailed in summer in Europe and possibly all the year round nearer the Equator, a temporary open-air site with a makeshift windbreak or skin tent may have sufficed.

Homo sapiens' early distribution

The known occupation sites of early *Homo sapiens* are so few that it is unlikely that any valid distribution pattern can be described at this stage. The principal early sites are from Europe, but they vary in age to such an extent that it would be unwise to draw conclusions as to possible migrations or to population density at any given time period. Little more can be said than that early *Homo*

sapiens was possibly present in southeastern Europe 350,000 years ago, in western Europe 200,000–250,000 years ago, and in East Africa 130,000 years ago. After those times, evidence of *Homo sapiens* becomes widespread and can be found in the rest of Europe, the Mediterranean, southern Africa, China, and the Far East. Finally, America was occupied from north to south via the Bering Strait at least 20,000 years ago, and Australia was occupied from the Southeast Asian peninsula a little earlier.

The Vértesszőllős site. The deposits that produced the fossils from this site show four layers of dried mud that may have come from thermal springs. The lowest of these layers was five centimetres thick and contained the fossils. A number of fossil mammalian bones were recovered from the site including those of a beaver (*Trogontherium schmerlingi*), Etruscan rhinoceros (*Didermocerus etruscus*), wild dog (*Canis etruscus*), a large sabre-toothed carnivore (*Epimachairodus*), and a large number of the bones of smaller mammals. The lower two of the four occupation layers correspond with a period of mild climate, and the upper two layers correspond to a period of colder conditions. The imprint of beech leaves in the lower two layers suggests that the remains should most probably be dated to an interval of temperate climate within the Mindel Glacial Period, roughly 350,000 years ago.

In addition to the fossil bones a large number of stone tools of small size were found. They consist of pebble tools, chopper tools, and numerous stone flakes. The stone-tool industry has been termed the Buda industry, and it contains no hand axes, significant because early pebble-tool industries usually also contained hand axes, which later ceased to be used in parts of Europe. Traces of fire were confirmed at this living site; thus, Vértesszőllős man must rival Peking man (*Homo erectus*) as the first known fire user, and he is without doubt the first sapient fire user, if the attribution of the Vértesszőllős remains to early *Homo sapiens* is fully confirmed.

Although from such scanty environmental and cultural evidence it is possible to build only an incomplete picture of the situation of Vértesszőllős man at this site, some of the general considerations seem to apply. The site is in the fourth terrace of the Danube River system, which supplied water; the tools are made of flint and quartz from nearby sources; and the mammalian fauna suggests the availability of quarry for hunters. If the climate was temperate, it may not have been necessary to seek shelter all the year round, so that an open site such as this could have been acceptable. The evidence of fire could suggest the cooking of food, but perhaps it is more likely to have been a necessity in the occupation of an open site as a means of warding off predators, such as the sabre-toothed carnivore, at night. Finally, the concentration on the site of the bones of large mammals could indicate the need for cooperative hunting or trapping.

The Swanscombe site. The environmental and cultural information concerning Swanscombe man has been obtained from the river gravels in which the skull was found and from the layers immediately below. These gravels form part of the 100-foot terrace of the Thames River; the skull was found in the Upper Middle Gravels. Twenty-six species of fossil mammals are known from this layer, and the character of the assemblage suggests that the climatic conditions were relatively warm. Among the species recorded are Merck's rhinoceros (*Didermocerus kirchbergensis-mercki*), straight-tusked elephant (*Palaeoloxodon antiquus*), cave lion (*Panthera spelaea*), wolf (*Canis lupus*), horse (*Equus caballus*), fallow deer (*Dama clactoniana*), giant ox (*Bos primigenius*), red deer (*Cervus elaphus*), and hare (species of *Lepus*). A number of these species would have provided suitable food for Swanscombe man. Although there is no direct evidence that the animals were hunted, it can be inferred that this was the case.

Swanscombe fauna

During the Mindel-Riss Interglacial Stage, the river was very wide and at times slow flowing, with marshy ponded areas along its banks. It is likely that herds would have gathered to drink and browse near the waterside and in this situation would have been very vulnerable to hunting by early sapient man. Some recent evidence from the boundary between the lower loam and lower gravel deposits that precede the layer in which Swanscombe man occurs suggests that kills were butchered at the river edge. It is not unlikely that temporary encampments would have been made after a successful hunt, but so far no true living floor has been found at this site.

Many nonmarine mollusks found at the site have contributed information concerning the environment. These include land snails of species known from European Mindel-Riss Interglacial Stage sites and aquatic snails that had probably spread to the Thames from the Rhine, which by this stage had become connected with the Thames. The types of land snails found suggest a temperate climate with summers that were warmer than those of today and indicate the nearness of dry grassland, marshy riverbank, and scrub and woodland. The bird remains from Swanscombe are mainly those of aquatic birds such as geese (species of *Anser* and *Branta*), ducks (*Anatidae*), and cormorants (species of *Phalacrocorax*), while the scanty fish remains are confined to a large bone from the lower jaw of a common pike (*Esox lucius*) from the lower gravel.

The Swanscombe Middle Gravels are rich in stone tools, many of which are in fine condition, suggesting that they have not been moved very far by the action of water. This tends to confirm that the stream was relatively sluggish and that early *Homo sapiens* discarded these tools nearby. The implements recovered include numerous hand axes of the Middle Acheulean stone-tool culture (*c.* 200,000 years ago) and a number of flake tools (made by chipping stone flakes from a prepared core of stone). Recent studies of new cultural material from the earlier lower loam deposits has shown that some of these implements were made on the spot. A number of flakes have been found, in a circumscribed area of the deposit, that fit together neatly to reconstruct the original nodule of flint, which must have been brought to the site from some nearby source. A rubber mold of the cavity within the reconstructed nodule has shown the shape of the core that was produced by the flaking process. To find a group of flakes that fit in this way is convincing evidence of activity at this site and also shows that there was previously a land surface at this point. This was confirmed not far away by the presence of a preserved deer slot, or footprint.

The Swanscombe site is a good example of how specialists in a variety of fields can each contribute information that in conjunction enormously enhances the value of the whole. Fauna, flora, and climate are always closely linked, and with cooperative scientific efforts it is sometimes possible to build a convincing reconstruction of the ecological situation, as in this case.

The Steinheim site. The Steinheim site is on the Murr River, near Stuttgart, in West Germany, and is situated in a gravel pit excavated into the Pleistocene deposits washed down by this river. There are present a number of layers of sands and gravels that are overlain by a layer of loess—windborne deposits laid down during a subsequent cold phase. Four layers of gravels have been characterized, the third of which contained the Steinheim skull and numerous mammalian bones but no implements. The third-layer mammals include straight-tusked elephant (*Palaeoloxodon antiquus*), giant ox (*Bos primigenius*), bear (*Ursus arctos*), and lion (species of *Felis*). It is generally agreed that this constitutes an interglacial woodland fauna and should be attributed to the Mindel-Riss Interglacial Stage. Further speculations about the environmental situation of Steinheim man lack substance, as no evidence of a land surface or a living floor, in the shape of tools or evidence of fire, has been discovered. It can only be presumed that Steinheim man lived and hunted near a water source, but knowledge of his cultural attainments is entirely lacking.

The Fontéchevade site. The Fontéchevade cave is cut into the side of the valley of the Tardoire River near Montbron, Charente, France. When the cave floor was

Cave dwellers

excavated for the second time, a thick layer of stalagmite was exposed, beneath which was found the layer that contained the human remains. The thickness of the layer of stalagmite has separated the deposits into two principal layers and suggests a considerable time difference between them. The first excavation had removed a lot of later cultural material—including Magdalenian, Aurignacian, and Mousterian implements—but no human remains. The layer beneath the stalagmite contained a number of fossil mammalian bones and stone tools. The fossil fauna includes rhinoceros (species of *Dicerorhinus mercki*), fallow deer (*Dama*), tortoise (*Testudo graeca*), and bear (*Ursus*), fauna that suggest warm or temperate interglacial conditions. The tools recovered belong to two aspects of the flake industry that existed contemporaneously but distinct from a hand-ax culture during the Paleolithic, or the Old Stone Age, a term covering the time period from the oldest stone tools—*c.* 2,000,000 years ago—up to the use of polished stone tools—*c.* 10,000 years ago. The earlier flake tools at this site are termed Clactonian tools (type site Clacton, England), while the later are known as Tayacian tools (type site Tayac, Dordogne, France). On the grounds of the fauna and the tools, the fossil layer has been attributed to the Riss-Würm Interglacial (100,000 to 70,000 years ago).

In many ways the Fontéchevade site fits the general picture of early sapient sites: once again it is near to water; there is an apparent availability of game; there is possibly some evidence of fire use; and early stone tools were used in the cave. Although the flake cultures are relatively unsophisticated in their workmanship and finish, the tools need not be much less effective for general purposes on that account. Fontéchevade may be the earliest European site from which early sapient man is known to have led a relatively settled existence.

The Omo site. This site is situated to the north of Lake Rudolf in the valley of the Omo River in southern Ethiopia. The early sapient remains were found in member I of the Kibish Formation (Site KHS), a group of middle to upper Pleistocene layers (*c.* 100,000 years old) that overlie the older Nkalabong Beds. The Omo River drains southward into Lake Rudolf and is subject to considerable seasonal variation in level; this appears to have been the case over a long period of time. Frequent dry-land emergence characterizes the deposits and indicates the local appearance of brackish pools and marshy reeded areas in a deltaic region.

The Omo I skull and postcranial remains (remains of the skeleton other than the head) were found *in situ* near a minor unconformity, and they were associated with a few worked stone flakes, although not enough to characterize a stone-tool industry. In addition to the stone flakes some fossil mammalian bones and teeth were found from an early example of the modern African elephant (*Loxodonta africana*), from an advanced "archaic" elephant (*Elephas recki*), and from both the white and black rhinoceros (*Ceratotherium simum, Diceros bicornis*). The complete skeleton of a buffalo (*Syncerus caffer*) was collected *in situ* from the deposit in which the skeleton of Omo I was found.

Omo II discovery and dating

The Omo II skull was recovered from the other side of the valley from a site (Site PHS) spread over the side of a small clay deposit next to a hill representing the upper half of Member I and the lower third of Member II of the Kibish Formation strata. While the stratigraphic position (position in the layered geological deposits) of Omo II is not certain, the sedimentary sequence of PHS is analogous to that at KHS, and thus there seems little reason to question the assignment of both sites to the same geological level. No faunal or cultural remains were recovered with the Omo II skull. The dating of these remains was at first given as late middle Pleistocene or early upper Pleistocene (200,000–80,000 years old) on the grounds of geological inference, but later radiometric dating has indicated an age of 100,000–130,000 years.

To build an environmental and behavioral picture from this evidence is again difficult, but the association of early sapient sites with free water, the presence of suitable stone for flaking, and large mammals for hunting is confirmed once more. It can be assumed that the climate was not harsh at this latitude, so that an open-air riverside occupation site would have been possible. The presence of a group of waste flakes or flake tools could indicate local activity on a temporary land surface, much as has been found at Swanscombe, but there is no evidence of the remains of a stone structure or of fire at the places where Omo I or Omo II were found.

Stone tools associated with Homo sapiens. Having considered the fossil remains of early *Homo sapiens* and the environmental and cultural situation of the group by means of the analysis of remains associated with the fossils, it is possible to look further by considering other sites that have provided tools but no human remains. Sites of this kind are very numerous and it is easy, but unwise, to speculate as to the kind of man that occupied them on the theory that human types and tool types will invariably go hand in hand. Experience has shown that such a correspondence cannot be safely inferred in most cases because overlap both of differing human populations and differing tool cultures is known to exist.

Stone tool cultures

In general terms, the earliest known stone-tool culture from the lower Pleistocene (*c.* 1,000,000 years ago) seems to have split into two basically parallel cultural systems, the hand-ax cultures and the flake-tool cultures (Figure 3). Within each of these types there is evidence of late diversification, so that the variety of tools increases; each presumably was used for more specific purposes.

From (top) C. Ovey, *The Swanscombe Skull*, Royal Anthropological Institute of Great Britain and Ireland; (centre, bottom) K. Oakley, *Man the Toolmaker*; used with permission of the Trustees of the British Museum (Natural History)

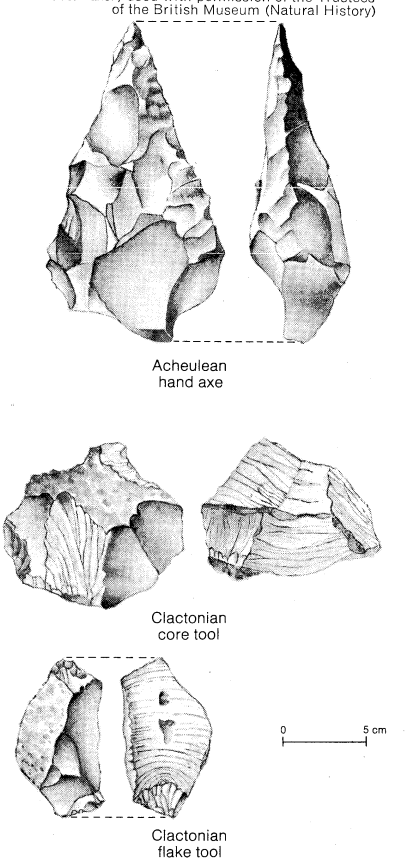


Figure 3: Early stone tools.

The distribution of tool sites shows some general features, such as the restriction of hand-ax cultures to Africa, western Europe, Arabia, and India, whereas the flake cultures overlap with the hand-ax cultures in western Europe but extend through the Balkans and into Southeast Asia. It seems clear that in Africa hand axes were developed from chopping tools, such as those found at Olduvai Gorge in Kenya, by the gradual extension of the flaking process around the edges until the flaked areas

met on the other side of the stone. The earlier stages of this process have been shown at Olduvai, but it is uncertain whether all hand-ax cultures are derived from a single African dispersal centre. The early Oldowan chopper tools lead eventually to the principal group of hand axes known as the Acheulean culture, and it seems very likely that early Acheulean hand axes were made by *Homo erectus*. This does not mean, however, that the later products of this industry must also be attributed to *Homo erectus*, as has been shown by the fine Acheulean tools recovered from Swanscombe and many other sites of much later date.

Flake-tool cultures

The parallel, but not mutually exclusive, flake cultures seem to have originated in the large chopper-tool industries, such as those made by *Homo erectus* at Peking (Choukoutienian tool industry). Similar early flake cultures are known from Burma (Anyathian), Malaysia (Tampanian), and from the North Punjab (Soan). The early "Buda" industry of Vértesszőllős has no hand axes but has many choppers and flake tools made from pebbles, perhaps the most primitive tools yet associated with *Homo sapiens*. The principal early development of the flake industry is the Clactonian culture, which is known from a number of European sites between the Mindel and Riss glacial periods (400,000 to 200,000 years ago). In this industry the choppers and chopping tools were made from large flint nodules by striking off coarse but useful flakes. In its later and more evolved form, this tool industry has been termed Tayacian (it dates from *c.* 200,000 BC) and is known from sites in the Dordogne, in France. Beyond this, the flake and the hand-ax cultures seem to merge into the Mousterian phase, while the more advanced Levallois technique leads on the Solutrean culture. (These are all stone-tool industries of the Upper Paleolithic in Europe—80,000–15,000 years ago.) All of the latter industries are well associated with *Homo sapiens* of the upper Pleistocene (after *c.* 100,000 years ago).

Cultural Correlations in the Paleolithic Era

geological periods	years before present (000)	hand-axe cultures	flake cultures	sapient sites
Würm III Glaciation		Magdalenian		Neanderthal and later sapient sites
Paudorf Interstadial		Solutrean		
Würm II Glaciation		Aurignacian		
Gottweiger Interstadial		Perigordian		
Würm I Glaciation		Mousterian		
Third Interglacial (Riss-Würm)	100–130		Tayacian	Fontéchevade Omo
Riss II Glaciation			Levalloisian	
Inter-Riss Interstadial				
Riss I Glaciation				Steinheim Swanscombe
	250	Acheulean	Clactonian	
Second Interglacial (Mindel-Riss)				
Mindel II Glaciation				
Inter-Mindel Interstadial	350 (?)		Buda	Vértesszőllős
Mindel I Glaciation				
First Interglacial (Günz-Mindel)		Abdevillian		

DATING HOMO SAPIENS SITES

One of the most important aspects of the investigation of a fossil is the determination of its age. Not only does accurate dating set the remains in a proper stratigraphic context but it also has a strong bearing on the phylogenetic (natural evolutionary ordering) interpretation of fossil populations, since their order of succession cannot be safely determined on morphological grounds alone.

Methods of dating fossils. The methods of determining the antiquity of buried objects have improved immensely in recent years with the development of both chemical and radiometric techniques; nonetheless, the traditional means of dating deposits are still valid. At any fossil site it is of the utmost importance that the geology of the area be well understood in terms of both geomorphology and stratigraphy, for herein lies the basis of geochronology, the science of dating. Sequences of deposits usually conform to the superposition principle—that younger layers overlie older layers and, a priori, that objects contained in a layer will be of the same age as that layer. The problem is complicated, however, by the possibility of younger objects finding their way into older layers by burial or older objects finding their way into younger layers by erosion and redeposition.

Relative and absolute dating

In dating fossils it is first necessary to identify the geological layer in which they are found and then to establish that they are of the same age as that layer. This gives the so-called relative dating of the object, and once this is established it is reasonable to attempt to fix the absolute, or chronometric, date of the object in years before the present (BP). In practice it is rarely possible to use all of the many dating methods available, since the materials upon which they depend are not present at every site; thus some sites and some fossils are better dated than others (see the Table). For a discussion of the methods of geological and archaeological dating, see DATING, RELATIVE AND ABSOLUTE.

Dating the remains of Homo sapiens. *The dating of Vértesszőllős.* The geology and stratigraphy of the Vértesszőllős site has been investigated and the fossil layer confirmed as being in the fourth terrace of the Danube system. Its contemporaneity with that layer has been established by the estimation of the amount of fluorine in the fossils and associated remains. This test depends on the principle that buried bone gradually accumulates fluorine from the groundwaters of the deposit. The total quantity of fluorine in the fossil will depend finally upon the concentration of fluorine in the groundwaters, the time spent by the bone in the ground, the nature of the deposit, and the permeability of the bone. Since these are all variable factors, the method is of little value as a means of chronometric dating but is of very great value in determining the contemporaneity of a group of fossil bones. For example, it may show that a human skull is of the same age as the bones of an extinct animal with which it is associated, or it may reveal that the human remains were introduced into the layer at a later date by showing their fluorine level to be much lower than that of the associated remains. The contemporaneity of the remains at Vértesszőllős having been established, the faunal and archaeological evidence from the site can be considered by allowing comparisons with other sites. This comparative evidence has established the relative dating of the site to the Inter-Mindel Interstadial Period. Direct chronometric dating by the uranium–thorium method has given the site an age of 350,000 years BP.

The dating of Swanscombe. The recovery of the Swanscombe remains *in situ* from the 100-foot terrace of the Thames has strongly suggested that these remains became buried during the Mindel-Riss Interglacial Stage. The contemporaneity of the human bones with other mammalian remains has been confirmed by the fluorine method, by the determination of the organic nitrogen content of the bones, and by radiometric assay of their uranium content. The satisfactory results of these relative dating techniques also confirm the contemporaneity of the Swanscombe skull with the tool culture found in the Upper Middle Gravels of the site; this, in turn, correlates well with other Acheulean sites. No direct chronometric date has been given for Swanscombe, either on the specimens or the source deposit, but its age—about 250,000 years BP—can be inferred by correlation of this bed with others that have been determined.

The dating of Steinheim. In the past there has been some difference of opinion regarding the dating of the Steinheim find. The gravels have been assigned by some to the Mindel-Riss Interglacial and by others to the Riss-Würm Interglacial. The contemporaneity of the find with the associated remains has been established by the fluorine method, and recent paleontological opinion is united in attributing the layer to the earlier of these two interglacial periods—*i.e.*, the Mindel-Riss Interglacial. No direct chronometric date has been determined either on the

specimen or the source deposit, but its age in years can be inferred at about 200,000 years BP by correlation of this bed with others that have been determined to be of that age.

The dating of Fontéchevade. The Fontéchevade site is a cave deposit stratified by archaeological layers and a stalagmite layer. The fossil layer has been attributed to the Riss-Würm Interglacial, in view of the fauna and the tool culture that it contains. Confirmation of the contemporaneity of the specimen with the other remains in the layer has been obtained by the fluorine test. No direct absolute date has been determined on either the specimen or the source deposit, but its age in years can be inferred by correlation with other sites at somewhere between 70,000–150,000 years BP.

The dating of Omo. The Omo I human remains were partly *in situ* in Member I of the Kibish Formation, a layer of undeformed Pleistocene lake and river sediments to the north of Lake Rudolf. The Kibish Formation consists of four members (major strata), the lowest of which (Member I) is over 40 metres thick and has seven stratigraphic sub-units. It was from Member I that the Omo I skull was recovered, while the Omo II skull has been referred to the same level in an analogous stratigraphic sequence.

The relative dating of the site has been established by means of the fluorine, nitrogen, and uranium tests, and the fauna recovered is not inconsistent with the upper middle Pleistocene date (*c.* 100,000–200,000 BP), but it cannot be positively dated to this time level. The stone flakes are also undiagnostic in terms of relative archaeological dating. Direct chronometric dating has been attempted on material from the deposit, making use of both the carbon-14 and the uranium–thorium radiometric methods. The carbon-14 test has shown that the upper part of Member III, Member II, and Member I are all too old to be dated by this method and, therefore, older than 35,000–40,000 years BP. The uranium–thorium method depends upon the decay of uranium with the formation of thorium. Since the rate of decay of uranium is known, a chronometric date can be calculated from the proportion of thorium to uranium in the specimen. Member I of the Kibish Formation has been dated to 130,000 years BP by this method, and thus both Omo I and Omo II are regarded as being of this age because their sites are stratigraphically equivalent.

BIBLIOGRAPHY. M.H. DAY, *Guide to Fossil Man* (1965), a basic reference book that brings together the most significant findings in human paleontology and related disciplines, including extensive references and a glossary of terms; W.E. LE GROS CLARK, *The Fossil Evidence for Human Evolution*, 2nd ed. rev. (1964), a classic account of human paleontology; CAMERON D. OVEY (ed.), *The Swanscombe Skull* (1964), a monograph that brings together the information that is available concerning this fossil, its morphology, classification, dating, and site context; KENNETH OAKLEY, *Frameworks for Dating Fossil Man* (1964), an authoritative handbook on the dating of human fossil remains; *Man the Tool-maker*, 5th ed. (1961), a simple but accurate account of the history of stone tool cultures; FRANCOIS BORDES, *Typologie du Paléolithique ancien et moyen* (1961; Eng. trans., *The Old Stone Age*, 1968), an authoritative and readable account of the tool cultures of the Paleolithic period; R.E.F. LEAKEY, KARL W. BUTZER, and M.H. DAY, "Early *Homo sapiens* Remains from the Omo River Region of South-west Ethiopia," *Nature*, 222:1132–1138 (1969), source reference for the new material attributed to *Homo sapiens*.

(M.H.D.)

Honan

Honan (Ho-nan in Pin-yin romanization), a small province in the north central part of the People's Republic of China, has an area of 64,400 square miles (166,800 square kilometres). The province stretches some 300 miles from north to south and 350 miles east to west at its widest point. It is bounded on the north by Shansi and Hopeh; on the east by Shantung and Anhwei; on the west by Shensi; and on the south by Hupeh. The Huang Ho (Yellow River) divides the province into two unequal parts—one-sixth north and five-sixths south of the river —and thus to some extent belies the name Honan, which means South of the River. In 1970 the population was estimated to be 50,300,000. K'ai-feng, the former capital, has been superseded by Cheng-chou, where the Peking–Hankow railway crosses the Huang Ho and meets the Lunghai Railway running from east to west.

History. Honan abounds in prehistorical and early historical interest. Some of the most important evidences of the Neolithic beginnings of Chinese civilization are found in the northern part of the province. It was at Yang-shao-ts'un in north Honan that a Swedish geologist and archaeologist, Johan Gunnar Andersson, in 1921 discovered an assemblage of Neolithic painted pottery, that, together with many later finds, marked the presence of a well-established primitive farming culture, which has been named Yang-shao. The early farmers occupied the lands at the confluence of the Huang Ho, Wei Ho, and Fen Ho, and it was here that Chinese civilization was cradled. The other main Honan sites of the culture are at Miao-ti-kou, Lin-shan-chai, P'an Nan, and Hsi Yin. These earliest of farmers, who were also part-time hunters and fishermen, lived in sunken circular or rectangular dwellings, sometimes of considerable dimensions. They grew foxtail millet, broomcorn millet, and kaoliang and had domesticated dog and pig. Cultivation with their primitive stone tools was comparatively easy in the easily worked loess (wind-borne) soil.

Cradle of Chinese civilization

Immediately to the east, at Lungshan in Shantung, a different culture was discovered, known as the Black Pottery culture, as distinct from the slightly earlier Yang-shao Painted Pottery culture.

It was on these Yang-shao–Lung-shan foundations that the early civilization of the Shang (Yin) dynasty arose (*c.* 1766–*c.* 1122 BC) in north and west Honan, south Hopeh, and west Shantung. Excavations near An-yang and, more recently (1950), in Cheng-chou and Hsing-t'ai, Hopeh, revealed an advanced culture, having a hierarchical class structure, advanced buildings, and elaborate ritual in which beautiful bronze vessels were used. Tung Tso-pin, the leading authority on oracle-bone inscriptions, has calculated from them that the Shang king Pan Keng moved his capital to An-yang in 1384 BC.

When the Shang kingdom fell to the Chou dynasty (1122–221 BC), An-yang lost its status as a capital. When the Chou capital, Hao (Kuoking), was destroyed in 771 BC by western tribes, Lo-yang (then known as Lo-i) took its place. During the period 771 BC to AD 938, the distinction of being the capital was shared alternately by Lo-yang and Chang-an (modern Sian in Shensi Province). Lo-yang was the capital during the following dynasties—the Eastern Chou (771 to 221 BC), Later (Eastern) Han (23 BC to AD 220), Wei (220 to 264), Western Chin (265 to 317), Northern Wei (386 to 534), and Later T'ang (*c.* 924 to *c.* 936). With the fall of the T'ang dynasty, K'ai-feng took over the role of the nation's capital and remained so until the Northern Sung dynasty was overthrown by the Mongol invaders in 1126. From that time on the national capital was never again in Honan. Both Lo-yang and K'ai-feng remained important during that period because of their strategic locations in the gateway leading from the North China Plain into the Huai Ho Basin, thence into the Yangtze Basin.

The natural environment. *Relief.* Honan can be divided topographically into two parts, the western highlands and the eastern plain. In the northwest the rugged T'ai-hang Shan and Chung-t'iao Shan form the steep eastern edge of the Shansi Plateau, rising in places above 5,000 feet. They are part of the T'ai-hang fold system of Permian times and have a general northeast to southwest trend. They mark the northern border of the province.

South of the Huang Ho there is a broad stretch of upland comprised of a number of moderately high mountain basins, the main ranges being the Hsiung-erh Shan and Fu-niu Shan. These mountains, which have an east–west trend, are the eastern extension of the great Nan Shan–Tsinling Shan axis that divides China geologically and geographically into north and south. The T'ung-pai Shan and Ta-pieh Shan form a further extension of this axis,

The western highlands



Prince Siddhārtha, the Gautama Buddha, in meditation, stone relief from the Lung-men caves near Lo-yang, early 6th century. In the Fogg Art Museum, Harvard University.
By courtesy of the Fogg Art Museum, Harvard University, Grenville L. Winthrop Bequest

running in a southeasterly direction and marking the border between Honan and Hupeh. The T'ung-pai Shan is separated from the Fu-niu Shan by a gap of some 75 to 100 miles cut by the T'ang Ho and T'ao Ho, which are tributaries of the Han Shui. This gap gives easy access from the Honan Plain to the central basin of the Yangtze, a route much used from Han times onward in Chinese expansion southward.

To the east lie the plains. Until fairly recent geological times the mountains in the west (the western extension of the present Gulf of Chihli and Yellow Sea) formed the coast of a sea. That sea, now filled with silt brought down by the rivers and by the wind from the loess plateau, forms the North China Plain and the Huai Basin. It is part of the great Neo-Cathaysian Geosyncline (downward-sloping part of the Earth's crust), which extends from Heilungkiang to Kiangsi provinces. The floor of this geosyncline is sinking at a rate equal to that of deposition; it is estimated that the sediment of the plain is now 2,800 feet deep in places.

Drainage. Honan has three river systems: the Huang Ho in the north and northeast, the Huai Ho in the east and southeast, and the T'ang Ho and T'ao Ho in the southwest. The Huang Ho—known in Chinese literature simply as the Ho (or the River)—immediately after its confluence with the Wei Ho, at the Shansi border, turns eastward at T'ung-kuan to flow right across the north of Honan. Near T'ung-kuan it enters the San-men Hsia (San-men Gorge) of some 80 miles, thence issuing onto the plain. It is remarkable that from T'ung-kuan to the sea, a distance of some 600 miles, the Huang Ho receives only two comparatively small tributaries: the right-bank Lo Ho, on which Lo-yang stands, and the left-bank Ch'in Ho.

Enormous silt-carrying capacity of the Huang Ho

The Ho is subject to very great changes in summer and winter flow. The maximum recorded summer flow is 883,000 cubic feet per second, and the minimum recorded winter flow is 8,650 cubic feet per second. In time of maximum flow it carries an enormous load of silt, estimated at 1,600,000,000 tons per year, gathered mainly in its course through the loess plateau of Shensi and Shansi. In this area it carries an average silt concentration of 1.6 pounds per cubic foot, in comparison with the Colorado's 0.5 pound and the Nile's 0.06 pound per cubic foot. There is a Chinese saying that "If you fall into the Huang Ho you never get clean again." While the river is fast flowing in the San-men Hsia, it is able to carry its load of silt, but when it issues onto the plain its pace is checked —it can no longer carry the silt, and flooding occurs. Throughout historical times this menace has been met by building levees to contain the waters. Generally these dikes were built five to eight miles apart to give the river plenty of room in time of spate (flooding), but instead the load of silt has been spread, building up the riverbed through the centuries, until today it lies above the surrounding countryside. Dikes have been built higher and higher, and when they fail to hold—as happened in some part of the province almost every year—the river descended onto the plain, causing disastrous floods, the waters of which could not return to the high stream bed when the river's flow slackened. The result was waterlogging of the soil, crop destruction, and famine. Because the watershed between the Huang Ho and Huai Ho is almost imperceptible, the Huang Ho has radically changed its course several times in the last three millennia, flowing to the sea, first south, then north, of the Shantung Peninsula. The diversion has always been in north Honan between Cheng-chou and K'ai-feng. In 1938, in an attempt to arrest the advance of the invading Japanese army, the Huang Ho was deliberately diverted by blowing up the dikes near Cheng-chou and flooding 21,000 square miles of land, at an estimated cost of 900,000 lives. The river was restored to its former northern course in 1947 by United Nations action.

To meet a continual flood menace, the government published in 1955 a scheme known as the "Staircase Plan," calling for the building of 44 retention dams in the main river and 79,000 silt-precipitation dams in the tributaries. The key to the whole conservation plan was to be the construction of the San-men Dam in the San-men Hsia at a point where two rock islands divide the river into three channels. The dam, 295 feet high, was designed to retain a lake with a capacity of 1,270,000,000,000 cubic feet. The fourfold purpose of the dam was flood control, irrigation, hydroelectric power, and navigation. The hydroelectric plant was intended to supply the rapidly growing industries of K'ai-feng, Lo-yang, and Cheng-chou in Honan and Sian in Shensi Province. The dam was due for completion in 1962; in 1961, before this part of the work was completed, Russian advisers and technicians were withdrawn; there is no report of its final completion.

The Huai Ho itself and all its major left-bank tributaries have their sources in the mountains of western Honan. They flow eastward onto the Anhwei Plain, subjecting it to disastrous floods. In 1949 the Huai Basin was chosen by Mao Tse-tung for the first big water-conservancy program. Six dams were quickly built in the upper reaches of Huai tributaries in Honan, four on left-bank tributaries and two on right-bank tributaries. Since 1957 three very large dams at Lung Shan, Mei-Shan, and Fo-tzu-ling have been built. Dikes were strengthened, with the result that no serious disaster has since occurred.

Huai Basin water-conservancy program

Climate. Climatically, Honan lies in a transitional zone between the North China Plain and the Yangtze Valley. Although protected in some degree from the Mongolian winds by the T'ai-hang Shan, Honan has very cold winters; summers are hot and humid. Average January temperature in the north is 28° F (−2° C) and in the south 36° F (2° C). Average July temperature over the lowlands is 82° F (28° C), while in the western mountains it is a degree or two lower. The north enjoys 210 frostless days in the year and the south 250.

Rainfall is distributed more evenly throughout the year than it is in the rest of North China, although there is a marked spring–summer maximum. K'ai-feng has an average rainfall of 23 inches, of which only 3 inches fall in the autumn and winter months. There is a steady decrease in total rainfall from southeast to northwest and a marked increase in variability. Honan is therefore more subject to years of alternating heavy rain and drought than the provinces of the Yangtze Valley. In the past it has suffered from severe famine. It also experiences spring cloudbursts and occasional hailstorms, both of which can be very destructive. In times of drought, summer dust storms are worse even than those of winter.

Soils. Honan's soils are made up mainly of calcium carbonate (lime) in hardened layers of alluvium. Because of the comparatively low rainfall, there is little leaching. The higher land of the west is mainly mountain yellow-brown earth, better drained than the plains. The more fertile areas fringing the plain were the sites of early civilization. Alluvium is spread throughout the plain; it is yellowish and gray, porous, granular, and poor in organic matter. Since the bed of the Huang Ho lies above the surrounding land, much of the low-lying land on either side is waterlogged. Consequently, soil salinity and alkalinity affect the whole area. There are large areas of bleak, white saline sands. Since 1949 there has been much experimentation aimed at bringing these alkaline lands into production. It is reported that between 1954 and 1964 a fourth of the saline land between K'ai-feng and Cheng-chou was transformed into fertile farmland.

Salinity of the Huang Ho flood basin

Vegetation. The natural vegetation of Honan is deciduous forest and woodland over the plains, and deciduous and coniferous forest in the western highlands. Intensive settlement of the plains has long since led to the clearance of the trees to make way for cultivation. The mountains, however, retain some of their woodland. Since 1949 serious efforts have been made to educate the peasantry in the importance of reforestation. It has been reported that more than 50,000 acres of trees have been planted specifically for sand anchorage and as windbreaks.

Population. According to the census figures published by the State Statistical Bureau in November 1954, Honan had a population of over 44,000,000, making it the third highest in the nation to Shantung and Szechwan, and an average overall population density of 670 per square mile. There is no clear and precise demarcation between rural and urban population in that census. Generally, places of 2,000 or less whose population was engaged essentially in agriculture were classified as rural, but, if mining or some other industrial activity prevailed, the settlement was recorded as urban. Before 1949 Honan's urban population was very small indeed, and even in 1958 it was still only about 8 percent of the whole. There were only seven cities of over 100,000 in 1958: Cheng-chou (capital), 790,000; K'ai-feng, 320,000; Lo-yang, 500,000; Hsin-hsiang, 200,000; Chiao-tso, 250,000; Shang-ch'iu, 165,000; and An-yang, 153,000. There were 300 townships of under 5,000. Since 1958 there have been few official estimates of growth, but it is safe to say that the populations of cities such as Cheng-chou, Lo-yang, and K'ai-feng, with their great industrial development, have increased considerably.

The greatest concentration of rural population is in the eastern plain, where densities of between 800 and 1,000 per square mile are found. Nearly as great densities are found in the I Ho and Lo Ho basins and in the plain around Nan-yang, but in the more mountainous west and south they are considerably less. On the eastern plain, villages average about 400 inhabitants and are fairly close together, usually about one mile apart. In the mountains they are smaller and more widely dispersed. Houses are made mainly of mud-plastered walls and thatched roofs. There was considerable movement of rural people from Honan and other provinces of the plains to towns in the west in 1958–59, during agricultural collectivization and the Great Leap Forward.

Ethnic homogeneity of Honan

The ethnic composition of the population of Honan is remarkably homogeneous. The people are essentially Han Jen (northern Chinese). There are no autonomous minority groups such as are found in the western provinces. Mongol and Manchu invaders alike have been absorbed and Sinicized. In Later Han times a considerable influx of Jewish refugees into the province occurred and in the 10th and 11th centuries, when K'ai-feng was the imperial capital of the Sung dynasty, they formed an interesting and important part of the community. They retained their identity until the early 19th century but have since been absorbed.

Administration. On the victory of the Communists in 1949, the first administrative act of the People's government was to divide the country into six administrative areas. Honan, together with Hupeh, Hunan, Kiangsi, Kwangtung, and Kwangsi formed the Central South Administrative Area. In 1954 provincial government was established, and for local governmental purposes Honan has been subsequently divided into three municipalities (*shih*), ten special districts (*chuan-ch'ü*), 11 county-level municipalities (*shih*), and 110 counties (*hsien*).

This pattern of local government underwent considerable change in 1958–59 with the nationwide formation of communes. The first commune in China was formed in Honan in April 1958 as a result of experiments by a number of advanced agricultural-producer cooperatives who banded together into a much larger unit, largely to increase production. Many nearby localities immediately followed suit. By November 1958, 99 percent of the rural population was grouped in communes, with a marked effect on local administration.

Although there was no stereotyped blueprint, communes generally gathered a number of advanced cooperatives into one large unit. Their responsibility extended to every activity of local life—the propagation of Communist ideology, the organization of agriculture, the promotion of local industry, primary- and middle-school education, public health, recreation, and the training of local militia. The communes had considerable autonomy and differed much in actual practice. The outbreak of the Cultural Revolution in 1966 led to disorganization and confusion in local government. This was not resolved until the appointment of provincial Revolutionary Committees. The committee for Honan was established in January 1968, and consisted of 11 members, nine of whom were selected from the People's Liberation Army, one from the "revolutionary cadres," and one from the "revolutionary masses."

Effect of the Cultural Revolution

Social conditions. *Education and culture.* Since its imperial days K'ai-feng has remained the cultural and educational centre, although it has come to share that role with Cheng-chou. The first impact of Western learning came, as in the rest of China, through the primary and middle schools of Christian missions. Little real progress was made in the disturbed years between 1911 and 1949, and the vast mass of the people remained illiterate. Successful efforts were made in the first years of the People's government to overcome illiteracy, and a real attempt at universal primary education was launched.

In 1959 responsibility for primary- and middle-school education was placed with the communes, and stress was laid on part-time study and part-time work schools. The Cultural Revolution seriously interfered with education between 1966 and 1969. At its conclusion many agricultural middle schools were set up in Honan by local poor and middle peasants, giving prominence to proletarian politics and basic agriculture. The curriculum includes class struggle, farming knowledge, military training, and physical culture. The use of the abacus, stock-breeding, and the installation, operation, and maintenance of machinery are taught. The fields, machine shed, and animal sheds serve as classrooms.

In an essentially agricultural province, such as Honan, cultural life is centred in the rural commune. While families, for the most part, retain and own their own homes, cultural life is focussed in the community centre, with its reading room, library, and teahouses, in which the old tradition of storytelling has been developed and is very popular. The better lighting of the centre makes it attractive. Loudspeaker radio is used to ensure communication, usually on a basis of one loudspeaker to three families. Indoor and outdoor games are fostered, table tennis being very popular. Cheng-chou was one of the first to establish city communes, which are much more complex in their organization. One, which includes 20 streets and a population of 18,000 adults and 11,000 children under 16, has 56 factories and workshops and three agricultural brigades raising food.

Health. Modern Western medicine, like education, was introduced by Christian missions but made very little impact on the vast area of Honan. On attaining power in 1949, the People's government concentrated attention on public hygiene and preventive medicine. A doctor's training was cut from six to three years, and

Emphasis on the "barefoot doctor"

teams were dispersed throughout the province to teach hygiene, vaccinate, inoculate, and advise. Although a doctor's training has now reverted to six years, the old pattern and emphasis remain, and the "barefoot doctor" is the order of the day. In Honan, with great development in coal mining, attention has been focussed on the prevention of silicosis, which is being achieved mainly by improving working conditions. Kala azar, the debilitating disease carried by sand flies, is also receiving special attention.

The economy. *Agriculture.* In 1949 Honan was considered one of the more backward provinces. Its economy is essentially agricultural. It lies in the region of the wheat eaters, as distinct from the rice-eating region of the south.

From 60 to 70 percent of the total cultivated area lies in the plains to the east of the Peking–Hankow railway, and 20 to 30 percent in the upland regions. The only idle land is found in the mountains and in the saline lands of the northeast. The main food crops are winter wheat, millet, kaoliang, soybeans, barley, corn, sweet potatoes, rice, and green lentils. These account for 89 percent of the total crop acreage. Wheat is by far the most important, in both acreage and production, Honan ranking first in China's output. Millet, kaoliang (Chinese sorghum), corn (maize), and sweet potatoes each occupy about 6 percent of the crop area. Millet is widely distributed; corn is grown mainly north of the Lunghai Railway and kaoliang to the east. Soybeans are produced on the eastern side of the province, with a marked concentration in the south, near the Anhwei border. Kaoliang is a hardy crop—more resistant to waterlogging and aridity than other grains, growing well on infertile soil, and requiring little attention. Rice occupies only some 3 percent of the crop area. Its yield per acre, however, is almost three times as great as wheat. With rapidly improving irrigation, rice cropping is increasing. Double-cropping is common in Honan. The wheat crop is usually followed by barley or rapeseed. Fruit growing has received considerable impetus in recent years, partly for its own sake and partly for soil conservation, particularly in the idle sandy lands of the northeast and on mountain slopes. Dates, persimmons, and pears are the main fruits, with apples, walnuts, and chestnuts on the increase.

Chief centre of draft animals

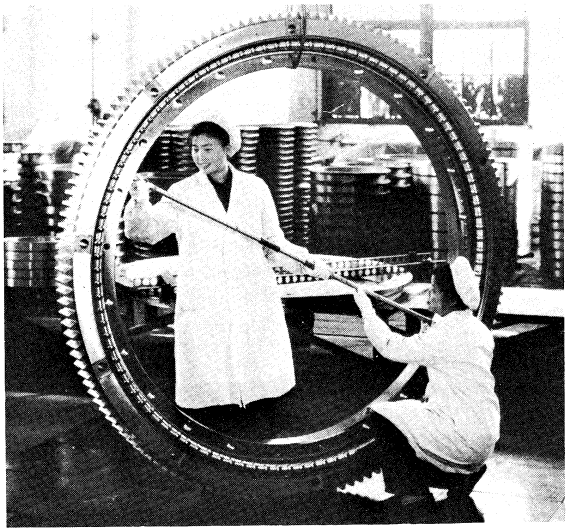
Honan produces more draft animals of good quality than any other province. The most important are yellow oxen and donkeys. Hogs are by far the most important food animals. Goats and sheep are raised in the western mountains and produce good mutton and wool.

The chief industrial crops, on about 10 percent of the total crop acreage, are cotton, tobacco, vegetable oils, and silk. Cotton is widely grown on about half the acreage, with its main concentration north of the Huang Ho around An-yang and Hsin-hsiang. Weak techniques of the past are being improved. Tobacco growing, introduced in Honan in 1916 by the British-American Tobacco Company, Ltd., has increased enormously since 1950. Vegetable oils are important, with Honan the largest producer of sesame in China, grown mainly in the east and south. Ramie, the most important of the leafy fibres, is grown in east Honan in the Huai Valley.

Honan is one of the oldest centres of sericulture (silkworm raising) in China. The industry dates back to the Later Han dynasty (AD 23–220). Both mulberry-leaf culture and oak-tree culture for silkworms were flourishing between the two world wars but suffered severely during the Sino-Japanese War. Since 1950 there has been a big revival on the slopes of the Fu-niu Shan, and the province has become an important exporter of silk.

Honan suffers very severely at times from locusts, which winter in the arid, sandy alkaline soils beside the Huang Ho. Between 1941 and 1943 they caused an 80 percent crop failure in the region. Extended and improved cultivation in these areas has helped control the pest.

Since 1959 considerable irrigation work has been carried out in the northwest, almost entirely by labour from the communes. Rural areas of Honan are now covered with a network of 40,000 miles of high- and low-tension electric lines that power pumps for 90,000 wells in the province. The total area now irrigated by pumps is reported to be more than 1,630,000 acres.



Workers in a ball bearing plant in the industrial centre of Lo-yang.
Eastfoto

Mining. Before 1949 there was very little industrial development in Honan, despite its very rich coal seams in the northwest. Both bituminous and anthracite coal are found along the slopes of the T'ai-hang Shan, and big reserves of good coking coal in thick, easily mined seams are found in the Fu-niu Shan between Hsü-ch'ang and P'ing-ting-shan. Iron ore is found at Ju-yang on the Ju Ho in the Hsiung-erh Shan, as well as some pyrite, bauxite, and mica. In recent years development has been rapid. Large coal mines at Chiao-tso supply the fast-growing industries of Lo-yang, Cheng-chou, K'ai-feng, and Hsin-hsiang but are still inadequate. The vast coalfield at P'ing-ting-shan was surveyed in 1953. Three mechanized mines, with a total annual production capacity of nearly 4,000,000 tons, were opened in 1964.

Site of China's first tractor factory

Industry. Lo-yang was chosen as the site for China's first tractor factory. No. 1 Tractor Plant was opened in 1958 with a designed capacity of 15,000 tractors a year. Since then, its output has burgeoned, and it is producing many models, ranging from tractors of 28 horsepower to 160-horsepower bulldozers. Lo-yang has become a heavy-industry centre. Cheng-chou lies in the heart of the cotton-growing area and is now the centre of the textile industry. It also has locomotive and tractor-repair workshops, food-processing factories, and the main thermal-power plant of the region. Its rapid growth is indicated by an increase in population from 100,000 in 1949 to 785,000 in 1958. K'ai-feng, imperial capital of the Sung emperors, declined when the river dikes were broken after the Yüan dynasty and the region ruined. Recent construction of large chemical-fertilizer works and a tractor-accessories plant have led to its revival. Hsin-hsiang, the most important city of north Honan, is the centre of the railway network of the area.

Transport and communications. Although the Huang Ho flows through north Honan, it serves it poorly as a line of communication. Within the province it was navigable only in the San-men Hsia until the construction of the San-men Dam. Even now it is useful over the plain only for small rivercraft. The Huai Ho and its tributaries flowing down from the western mountains are rapid in their upper courses and silted in their lower, so that they, too, serve only small craft. The Wei Ho, flowing north into the Hai Ho system, has been joined by canal to the Huang Ho. In 1964–65 it was successfully dredged in an experiment aimed at deepening the riverbed and so increasing flow and reducing waterlogging.

Cheng-chou is the junction of China's two greatest trunk railways, the Peking–Hankow–Canton line and the Lunghai line, which runs from the east coast via K'ai-feng, Cheng-chou, and Lo-yang to Sinkiang, Uighur

Autonomous Region, in the far west. The Peking–Canton line is now double-tracked throughout its entire length and carries a heavy load of both freight and passengers. The coalfields to the west are linked to it by branch lines.

The first modern roads in Honan date from the famine of 1920–21, when the American Red Cross built 850 miles of earth track to bring relief to the stricken provinces. Even in 1936 Honan had but 1,786 miles of road, of which only 167 miles were surfaced. By 1953 approximately 3,800 miles were open to traffic, mainly in the north centre and east. The most important and sole all-weather highway at that time ran from K'ai-feng via Hsu-ch'ang, P'ing-ting-shan, and Nan-yang to Hsiang-fan on the Han Shui in Hupeh. Since 1959 the great bulk of road building has been done with little modern technology by commune labour, some of it penetrating the more remote mountain region, as, for example, a road with a 240-foot tunnel in the T'ai-hang Shan between Hui-hsien and Ling-ch'uan.

BIBLIOGRAPHY. SUN CHING-CHIH (ed.), *Excerpts from Economic Geography of North China* (Eng. trans. 1958), one of many volumes by Chinese geographers covering the whole country; P.M. ROXBY (ed.), *China Proper*, 3 vol. (1944), a survey made during World War II; A. DONNITHORNE, *China's Economic System* (1967), a standard work drawing extensively on Chinese sources; S. CHANDRASEKHAR, *China's Population*, 2nd ed. (1960), an analysis of the census of 1953–54; K.C. CHANG, *The Archaeology of Ancient China* (1963), a standard work; T.R. TREGEAR, *A Geography of China* (1966), a general, college-level text; and *Economic Geography of China* (1970), more advanced, amplifying the economic and social section.

(T.R.T.)

Honduras

Honduras is a small republic of Central America situated between Guatemala and El Salvador, to the west, and Nicaragua to the south and east. The Caribbean washes its northern coast, the Pacific its narrow coast to the south.

With an area of 43,277 square miles (112,088 square kilometres) and a population (1970) of only 2,508,700, Honduras is one of the smallest countries of Latin America. The bulk of the population lives a rather isolated existence in the mountainous interior, a fact that may help explain the rather insular policy of the country in relation to Latin and Central American affairs. Its economy is primarily agricultural, bananas for export being particularly important, although maize, the chief staple, is by far the most important product in terms of value of annual output. The capital is Tegucigalpa, but—unlike most other Central American countries—there are other cities, such as San Pedro Sula, that are equally important industrially and commercially.

The landscape. Over 75 percent of the land area is mountainous, lowland being found only along the coasts and in the several river valleys that penetrate toward the centre of the interior. No clear-cut pattern of mountain trends is apparent, the interior taking the form of a dissected upland with numerous small peaks. Geological formations, however, have a general east–west orientation. Except for a narrow plain of alluvium bordering the Gulf of Fonseca in the south, the southwestern mountains consist of alternating layers of rock composed of dark, volcanic detritus and lava flows, both of Tertiary age (from 25,000,000 to 65,000,000 years old). The mountains in other regions are more ancient, granite and crystalline rocks predominating.

The four regions

Geographical regions. Four distinct geographical regions may be discerned: (1) The eastern lowlands and mountain slopes embrace about 20 percent of the total land area. Hot and humid, this area is densely forested, lumbering being an important economic activity. Subsistence agriculture and fishing are the main support of the scattered population of some 500,000 persons. (2) The northern coastal and alluvial plains and coastal sierras comprise about 13 percent of the land area and about 20 percent of the population. This is an economically important area, the clayey and sandy loam soils producing rich crops of bananas, rice, cassava, corn, and beans. Cattle, poultry, and swine are raised in considerable numbers. The nation's railroads are confined to this northern area, which has four of the five important ports of entry. (3) The central highlands take up 65 percent of the national territory and contain 70 percent of the population. The mountains are rugged and the valleys situated at altitudes of from 2,000–5,000 feet. The generally fertile soils, derived from lava and volcanic ash, produce coffee, tobacco, wheat, corn, sorghum, beans, fruits, and vegetables. Cattle, poultry, and swine are raised. (4) The Pacific lowlands and lower mountain slopes comprise only 2 percent of the land area and contain 5 percent of the population. The soils are fertile—composed of alluvium or volcanic detritus—and produce sesame seed, cotton, and small amounts of corn and sorghum.

Climate. The climate is generally hot, with high humidity in the tropical coastal lowlands becoming modified by altitude toward the interior. Lowlands below 1,500 feet have mean annual temperatures between 79° and 82° F (26° and 28° C). The north coast is occasionally affected from October to April by cool northern winds of continental origin. Mountain basins and valleys, from 2,000 to 4,000 feet, have mean annual temperatures 66° and 73° F (19° and 23° C). At Tegucigalpa, located on hilly terrain at an altitude of 3,200 feet, the rainy season starts in May and continues until the middle of November, with temperatures sometimes reaching 90° F (32° C) in May and dropping to 50° F (10° C) in December, the coolest month. Around 7,000 feet, mean annual temperatures are about 58° F (14° C). In the northern and eastern coastal and alluvial plains and on adjacent mountains, mean annual precipitation ranges from 70 to 110 inches or more, with a less rainy season from March to June; these areas occasionally have summer hurricanes with strong winds and heavy rains. Pacific plains and mountain slopes get 60 to 80 inches of rain annually, but there the months December to April receive little or no rain. Interior sheltered mountain basins and valleys get 40 to 70 inches annually, with December to April having little rain. Interior higher mountains have a longer wet season and receive considerably more rain.

Vegetation and animal life. In eastern Honduras the coastal and lagoon swamps have mangrove and palm forests, and west of these are low, rainy, sandy plains with pine-savanna (*Pinus caribaea*) forests, extending inland for 40 miles or more. West of the pine-savanna forests, in low valleys and on lower mountains, rainy all year, and on the low rainy northern mountains, are broad belts of dense evergreen broadleaf forests with many species of large trees, including mahogany, lignum vitae, Spanish cedar, balsa, rosewood, ceiba, sapota (yielding chicle used for the base of chewing gum), and Castilla rubber. The high rainy mountain slopes of highland Honduras support excellent oak–pine forests. The interior highland basins and valleys have open, dry deciduous woodlands and temperate grasslands. The Pacific plains and adjacent mountain slopes have deciduous tropical forest and savanna. Mangrove occupies the low coastal swamps.

Insects, birds, and reptiles are the most conspicuous animal forms. There are many species of butterflies, moths, beetles, bees, wasps, spiders, ants, flies, and mosquitoes, many of them beautifully coloured. Waterfowl in large numbers inhabit the coastal areas. Crocodiles, snakes, lizards (giant iguana and others), and turtles are found in the tropical forest areas. Bears, pumas, leopards, and panthers are more rare but may be found in the hills. Fish and mollusks are abundant in lagoons and coastal waters; by 1969 more than 1,300,000 pounds were exported annually.

The people. Honduras has been inhabited since well before the Christian Era: the ruins at Copán in western Honduras indicate that the area was the centre of Mayan civilization before the Mayas migrated to the Yucatán Peninsula. Although small, isolated groups of non-Spanish-speaking Indians—such as the Jicaques, Mosquito (Miskito, Mísquito), Zambos, and Payas—continue to exist, they are declining in numbers; the population is

Ethnic composition

Honduras, Area and Population

	area		population	
	sq mi	sq km	1961 census	1970 estimate
Central district (Distrito Central)*	636	1,648	165,000	263,000
Departments (*departamentos*)				
Atlántida	1,641	4,251	93,000	125,000
Choluteca	1,626	4,211	149,000	204,000
Colón	3,427	8,875	42,000	57,000
Comayagua	2,006	5,196	96,000	130,000
Copán	1,237	3,203	126,000	162,000
Cortés	1,527	3,954	200,000	274,000
El Paraíso	2,787	7,218	107,000	138,000
Francisco Morazán	2,432	6,298	119,000	157,000
Gracias a Dios	6,421	16,630	11,000	16,000
Intibucá	1,186	3,072	73,000	89,000
Islas de la Bahía	101	261	9,000	10,000
La Paz	900	2,331	61,000	70,000
Lempira	1,656	4,290	112,000	135,000
Ocotopeque	649	1,680	53,000	61,000
Olancho	9,402	24,350	111,000	140,000
Santa Bárbara	1,975	5,115	147,000	213,000
Valle	604	1,565	81,000	99,000
Yoro	3,065	7,939	131,000	164,000
Total Honduras	43,277‡	112,088	1,885,000‡	2,509,000§

*The Distrito Central of Tegucigalpa is not completely autonomous, as it is subordinate to the department of Francisco Morazan in selected administrative affairs. †Converted area figures do not add to total given because of rounding. ‡Census result excludes adjustment for underenumeration estimated at 5.3%. §Figures do not add to total given because of rounding.
Source: Official government figures.

predominantly mestizo (a mixture of Spanish and Indian). According to a census of 1950 the ethnic distribution was as follows: mestizos, 91 percent; Indians, 6 percent; blacks, 2 percent; and whites, 1 percent. The mestizos of Honduras are generally darker and more Indian in appearance than those of certain other Latin American countries such as Chile, Uruguay, and Argentina. Most of the Indians are found in the southwest, near the Guatemala border—near the most important Indian centres of the pre-Columbian period. A pronounced shift in population took place during the early part of the 20th century from the interior to the hot, humid north coast, where employment opportunities were provided by the United Fruit Company. There has been some reversal of this trend in later decades. The annual rate of increase of population is about four percent.

Almost 77 percent of the population is rural, living in small villages or isolated settlements. The mountainous, forested terrain and rather poor road transportation are conducive to local isolation, and, as a result, most Hondurans have never travelled more than a few miles from their home.

The economy. *Production.* Honduras possesses an agricultural economy, with only small-scale manufacture of such items as textiles and tobacco, chiefly for local consumption. In terms of contribution to the gross domestic product, agriculture, forestry, hunting, and fishing are followed by manufacturing, trade, services, and transport and communications.

Two large United States corporations—the United Fruit and Standard Fruit and Steamship companies—hold about 5 percent of the country's agricultural land and produce about 20 percent of the national income. Important export crops other than bananas include coffee beans, abaca fibre (manila hemp), tobacco, and coconuts. The extensive pine forests were attacked by blight in the 1960s, and mahogany—the major timber export—was declining in importance.

Mineral resources are considerable, though largely unexploited, and include antimony, iron, coal, mercury, and copper. Production of silver and gold has remained of some importance, and lead and zinc are produced in commercial quantities.

With a per capita gross national product of only $260, Honduras is a poor country, and the lot of the majority of Hondurans, who work on the land, is a hard one. The government has in recent years pursued more active economic policies, carrying out investment programs and granting incentives to private industry.

The Labour Code of 1954

In 1954 the trade-union movement obtained one of its most resounding triumphs, resulting in the promulgation of a labour code that has been considered one of the most complete instruments of its kind in Latin America. Labour–management relationships are regulated under the different articles of the code, which protects not only the worker but also the businessman. As a general rule, labour–management relationships are defined by means of collective contracts subscribed to by the labour unions and the management representatives of private enterprise. In some autonomous and semi-autonomous state agencies, there is the same labour–management relationship. The code has generally resulted in a higher standard of living for the worker and better operating conditions for the businessman.

Transportation. Honduras has less than 120 miles of public-service railroads. The two fruit companies own an additional 520 miles, which are confined to the north coast banana region and provide some passenger transport facilities.

A part of the Pan-American Highway, running from El Salvador to Nicaragua, cuts across southern Honduras for about 100 miles; a branch highway leaves it at Nacaome and runs north through Tegucigalpa to the north coast.

Air transportation is of great importance within Honduras and is frequently the normal means of conveyance for passengers and freight.

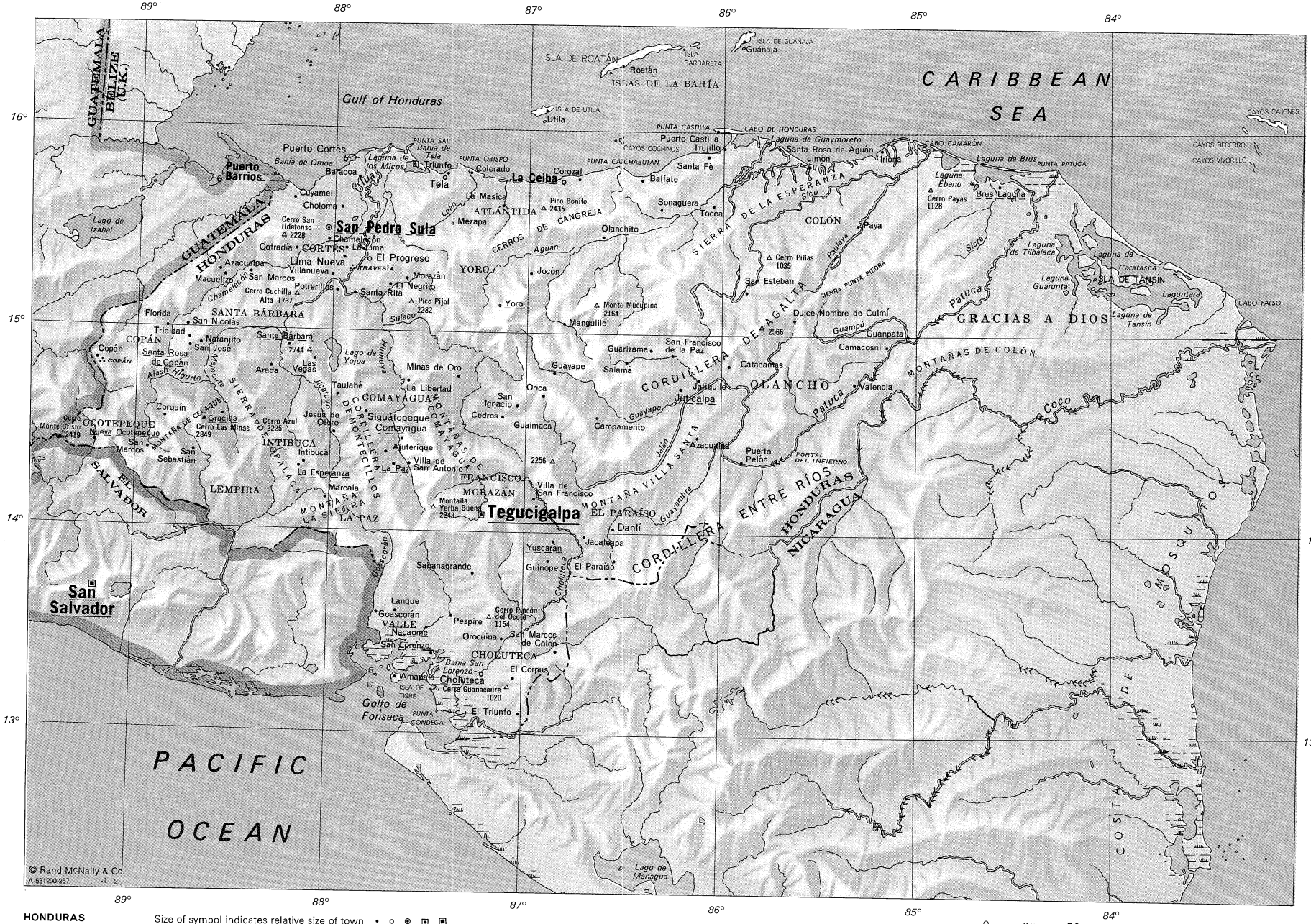
Administration and social conditions. Since acquiring independence in 1821, Honduras has been described constitutionally as a democratic, representative, unitary state with power divided among legislative, executive, and judicial branches. The country has had 16 constitutions, the latest having gone into effect in June 1965. Power, however, has frequently been mobilized and changed by violent, undemocratic means. Peaceful methods have included *imposición* ("rigged election"), *candidato único* ("one candidate"), and *continuismo* (continuing the president in power beyond his legal tenure). Although the legislature is given the power to pass laws, practically all important legislation is drafted by the executive and his assistants. The congress in theory has great authority to check the administrative activities of the president, but only in the period 1925–31, when several cabinet ministers appointed by the president were forced to resign through censure, was such authority effective. Strong, centralized dictatorial government was re-established in 1932. A theoretical counterweight to excessive executive power was devised in the 1950s and is embodied in the constitution of 1965: the chief of the armed forces, an almost autonomous official, is given the power of interposition against presidential orders, subject to final decision by congress.

Power of the executive

Honduras has a single-house legislature with single-member districts based on a ratio of one deputy for each 30,000 inhabitants. This numerical base may be changed by congress in accordance with increases in population. The congress convenes each year from May 26 to October 26, and its session may be prolonged by resolution initiated by a deputy or by the executive. Special sessions may be convoked by a permanent committee that represents congress during adjournment or upon petition by other deputies. All men and women who are citizens and over 18 years of age are permitted to vote, and a national electoral council supervises elections. The president is elected directly by popular vote for a period of six years with a second term prohibited.

For purposes of local administration, Honduras is divided into 18 departments and a central district.

The justices of the supreme court are elected by the congress for six years. The supreme court exercises centralized control over the lower courts, including the appointment of justices, and has original and exclusive jurisdiction to declare acts of the legislature unconstitutional. The departments are subject to the control of the central government, but the constitution establishes autonomy for municipalities.



HONDURAS

Size of symbol indicates relative size of town • ○ ◉ ▣ ■

Elevations in metres

MAP INDEX

Political subdivisions

Atlántida....15·30n 87·00w
Choluteca....13·20n 87·10w
Colón....15·20n 84·30w
Comayagua....14·30n 87·40w
Copán....14·50n 89·00w
Cortés....15·30n 88·00w
El Paraíso....14·10n 86·30w
Francisco Morazán....14·15n 87·15w
Gracias a Dios...15·10n 84·20w
Intibucá....14·20n 88·15w
Islas de la Bahía.16·20n 86·30w
La Paz....14·15n 87·50w
Lempira....14·15n 88·35w
Ocotepeque....14·30n 89·00w
Olancho....14·45n 86·00w
Santa Bárbara...15·10n 88·20w
Valle....13·30n 87·35w
Yoro....15·15n 87·15w

Cities and towns

Ajuterique....14·20n 87·43w
Amapala....13·17n 87·40w
Arada....14·48n 88·18w
Azacualpa....14·27n 86·09w
Azacualpa....15·19n 88·33w
Balfate....15·48n 86·25w
Baracoa....15·43n 87·52w
Brus Laguna....15·47n 84·35w
Camacosni....14·57n 85·08w
Campamento....14·33n 86·42w
Catacamas....14·54n 85·56w
Cedros....14·35n 87·08w
Chamelecón....15·24n 88·01w
Choloma....15·34n 87·56w
Choluteca....13·18n 87·12w
Cofradía....15·24n 88·09w
Colorado....15·47n 87·19w
Comayagua....14·25n 87·37w
Copán....14·50n 89·09w
Corozal....15·48n 86·43w
Corquín....14·34n 88·52w
Cuyamel....15·38n 88·12w
Danlí....14·00n 86·35w
Dulce Nombre de Culmí....15·09n 85·37w
El Corpus....13·16n 87·03w
El Negrito....15·16n 87·41w
El Paraíso....13·51n 86·34w
El Progreso....15·21n 87·49w
El Triunfo....13·06n 87·00w
El Triunfo....15·46n 87·26w
Florida....15·01n 88·50w
Goascorán....13·36n 87·45w
Gracias....14·35n 88·35w
Guaimaca....14·32n 86·51w
Guanaja....16·27n 85·54w
Guanpata....15·01n 85·02w
Guarizama....14·55n 86·20w
Guayape....14·45n 86·52w
Güinope....13·51n 86·55w
Intibucá....14·16n 88·10w
Iriona....15·57n 85·11w
Jacaleapa....14·00n 86·40w
Jesús de Otoro..14·26n 87·59w
Jocón....15·17n 86·58w
Juticalpa....14·42n 86·15w
Jutiquile....14·45n 86·08w
La Ceiba....15·47n 86·50w
La Esperanza....14·15n 88·10w
La Libertad....14·43n 87·36w
La Lima....15·24n 87·56w
La Masica....15·37n 87·07w
Langue....13·37n 87·39w
La Paz....14·16n 87·40w
Las Vegas....14·49n 88·06w
Lima Nueva....15·23n 87·56w
Limón....15·52n 85·33w
Macuelizo....15·18n 88·31w
Mangulile....15·03n 86·49w
Marcala....14·07n 88·00w
Mezapa....15·33n 87·23w
Minas de Oro....14·46n 87·20w
Morazán....15·17n 87·34w
Nacaome....13·31n 87·30w
Naranjito....14·57n 88·41w
Nueva Ocotepeque....14·24n 89·13w
Olanchito....15·30n 86·35w
Orica....14·41n 86·56w
Orocuina....13·26n 87·06w
Paya....15·37n 85·17w
Pespire....13·35n 87·22w
Potrerillos....15·11n 87·58w
Puerto Castilla...16·01n 86·01w
Puerto Cortés....15·48n 87·56w
Puerto Pelón....14·22n 85·53w
Roatán....16·18n 86·35w
Sabanagrande...13·50n 87·15w
Salamá....14·50n 86·36w
San Esteban....15·17n 85·52w
San Francisco de la Paz....14·55n 86·14w
San Ignacio....14·38n 87·02w
San José....14·54n 88·44w
San Lorenzo....13·25n 87·27w
San Marcos....14·24n 88·56w
San Marcos....15·17n 88·23w
San Marcos de Colón....13·26n 86·48w
San Nicolás....15·00n 88·45w
San Pedro Sula..15·27n 88·02w
San Sebastián...14·24n 88·42w
Santa Bárbara...14·53n 88·14w
Santa Fé....15·55n 86·05w
Santa Rita....15·09n 87·53w
Santa Rosa de Aguán....15·57n 85·43w
Santa Rosa de Copán....14·47n 88·46w
Siguatepeque....14·32n 87·49w
Sonaguera....15·38n 86·20w
Taulabé....14·38n 87·59w
Tegucigalpa....14·06n 87·13w
Tela....15·44n 87·27w
Tocoa....15·41n 86·03w
Trinidad....14·57n 88·45w
Trujillo....15·55n 86·00w
Utila....16·06n 86·54w
Valencia....14·47n 85·18w
Villa de San Antonio....14·16n 87·36w
Villa de San Francisco....14·10n 86·58w
Villanueva....15·17n 88·00w
Yoro....15·09n 87·07w
Yuscarán....13·55n 86·51w

Physical features and points of interest

Agalta, Cordillera de, *mountains*...15·00n 85·53w
Aguán, *river*....15·58n 85·44w
Alash Higuito, *river*....14·43n 88·40w
Azul, Cerro, *mountain*....14·32n 88·23w
Bahía, Islas de la, *islands*....16·20n 86·30w
Barbareta, Isla, *island*....16·26n 86·10w
Bonito, Pico, *peak*....15·38n 86·55w
Brus, Laguna de, *lagoon*....15·50n 84·35w
Camarón, Cabo, *cape*....16·00n 85·04w
Cangreja, Cerros de, *mountains*...15·35n 86·55w
Caratasca, Laguna de, *lagoon*....15·20n 83·50w
Caribbean Sea...16·20n 84·30w
Castilla, Punta, *point*....16·01n 86·02w
Catchabutan, Punta, *point*....15·50n 86·32w
Celaque, Montaña de, *mountains*....14·32n 88·43w
Chamelecón, *river*....15·54n 87·48w
Choluteca, *river*..13·05n 87·20w
Cochinos, Cayos, *islands*....16·00n 86·30w
Coco, *river*....15·00n 83·08w
Colón, Montañas de, *mountains*...14·55n 84·45w
Comayagua, Montaña de, *mountains*....14·23n 87·26w
Condega, Punta, *point*....13·06n 87·25w
Copán, *ruins*....14·50n 89·09w
Cuchilla Alta, Cerro, *mountain*.15·10n 88·12w
Ébano, Laguna, *lagoon*....15·52n 84·50w
Entre Ríos, Cordillera, *mountains*....14·00n 86·00w
Esperanza, Sierra de la, *mountains*....15·40n 85·45w
Falso, Cabo, *cape*....15·12n 83·21w
Fonseca, Golfo de, *gulf*....13·08n 87·40w
Goascorán, *river*..13·25n 87·48w
Guampú, *river*....14·59n 85·03w
Guanacaure, Cerro, *mountain*.13·14n 87·04w
Guanaja, Isla de, *island*....16·28n 85·54w
Guarunta, Laguna, *lake*....15·22n 84·11w
Guayambre, *river*....14·26n 85·58w
Guayape, *river*...14·26n 85·58w
Guaymoreto, Laguna de, *lagoon*....15·58n 85·55w
Honduras, Cabo de, *cape*....16·00n 85·55w
Honduras, Gulf of....16·10n 87·50w
Humuya, *river*....15·13n 87·57w
Jalán, *river*....14·42n 86·11w
Jicatuyo, *river*....15·00n 88·16w
Laguntara, *lagoon*....15·12n 83·30w
La Sierra, Montaña, *mountains*....14·04n 87·54w
Las Minas, Cerro, *mountain*....14·33n 88·39w
Leán, *river*....15·49n 87·19w
Mejocote, *river*...14·43n 88·39w
Micos, Laguna de los, *lagoon*....15·48n 87·36w
Montecillos, Cordillera de, *mountains*....14·25n 87·51w
Monte Cristo, Cerro, *mountain*....14·25n 89·21w
Mucupina, Monte, *mountain*....15·08n 86·38w
Obispo, Punta, *point*....15·50n 87·21w
Omoa, Bahía de..15·40n 88·08w
Opalaca, Sierra de, *mountains*....14·30n 88·20w
Patuca, *river*....15·50n 84·18w
Patuca, Punta, *point*....15·50n 84·16w
Paulaya, *river*....15·48n 85·06w
Payas, Cerro, *mountain*....15·45n 84·56w
Pijol, Pico, *peak*..15·06n 87·35w
Piñas, Cerro, *mountain*....15·25n 85·47w
Portal del Infierno, *waterfall*....14·22n 85·38w
Punta Piedra, Sierra, *mountains*....15·21n 85·16w
Rincón del Ocote, Cerro, *mountain*.13·36n 87·10w
Roatán, Isla de, *island*....16·23n 86·26w
Sal, Punta, *point*.15·55n 87·36w
San Lorenzo, Bahía, *bay*....13·20n 87·28w
San Ildefonso, Cerro, *mountain*.15·31n 88·17w
Sico, *river*....15·58n 84·58w
Sicre, *river*....15·49n 84·38w
Sulaco, *river*....15·02n 87·44w
Tansín, Isla de, *island*....15·20n 84·00w
Tansín, Laguna de, *lagoon*....15·13n 83·50w
Tela, Bahía de, *bay*....15·52n 87·30w
Tigre, Isla del, *island*....13·17n 87·38w
Tilbalaca, Laguna de, *lagoon*....15·28n 84·15w
Travesía, *ruins*...15·20n 87·53w
Ulúa, *river*....15·56n 87·43w
Utila, Isla de, *island*....16·05n 86·55w
Villa Santa, Montaña, *ridge*..14·10n 86·25w
Yerba Buena, Montaña de, *ridge*....14·08n 87·27w
Yojoa, Lago de, *lake*....14·53n 88·00w

By the 1970s Honduras, like most other Latin-American countries, had moved more and more in the direction of the interventionist, or welfare, state. A number of provisions in the constitutions of 1936, 1957, and 1965 gave the central government great power to direct and regulate social relations and the economic system. In February 1955 the Honduran basic labour code came into effect, granting the right to work, minimum wages, an eight-hour day and 48-hour week (reduced to 44 hours by the constitution of 1965), freedom to form labour unions, collective bargaining, conciliation, and the right to strike.

Education and cultural life. The Honduran educational system follows the European model of centralized control through the Ministry of Education. According to law, education is free and compulsory for all children. Efforts have been made to combat illiteracy, which in 1974 was still almost 40 percent for those 10 years of age and over. Related national efforts were reflected in the constitution of 1957, which dedicated 2 percent of the budgeted national income to the National Autonomous University of Honduras in Tegucigalpa (founded 1847), and in the constitution of 1965, which dedicated 3 percent. The university had an enrollment of about 8,000 in the mid-1970s.

The press

There is complete freedom of the press in Honduras, and daily newspapers are published in the principal cities of the country. Those of Tegucigalpa and San Pedro Sula are outstanding.

The progressive and rapid development of radio and television has provided the country with excellent facilities for speedy and effective communication. There are two private radio and television networks that cover the entire country. The state, private enterprise, and international organizations take advantage of these means for communicating to the Honduran people.

Cultural institutions in Honduras include the National School of Music, the House of Culture, and the San Pedro Cultural Centre, as well as many other state and private institutions. Among these others, especially noteworthy are the organizations dedicated to the production of theatrical works in both Spanish and English.

BIBLIOGRAPHY

Geography and topography: PAN AMERICAN UNION, DEPARTMENT OF ECONOMIC AFFAIRS, *Honduras* (1965), is a general study; V.W. VON HAGEN, *Jungle in the Clouds* (1945), tells of

archaeological remains, fauna, and Indians; while P. KEENAGH, *Mosquito Coast: An Account of a Journey Through the Jungles of Honduras* (1937), is another full exploration account. MARY LESTER, *A Lady's Ride Across Spanish Honduras, by Maria Soltera* (1964), a facsimile of the 1884 edition, is interesting. Also recommended is K. WEDDLE, *Honduras in Pictures* (1969).

History and politics: R.S. CHAMBERLAIN, *The Conquest and Colonization of Honduras, 1502–1550* (1966), deals with the period of the country's acquisition by Spain, while T.E. WRIGHT, *Into the Maya World* (1969); and DORIS STONE, *The Archaeology of Central and Southern Honduras* (1957), deal with archaeology. W.S. STOKES, *Honduras: An Area Study in Government* (1950), examines the political framework closely, as well as the PAN AMERICAN UNION, DEPARTMENT OF LEGAL AFFAIRS, *Constitution of the Republic of Honduras, 1965* (1966), and the INTER-AMERICAN PEACE COMMITTEE, *Report to the Council of the Organization of American States on the Termination of the Activities of the Honduras-Nicaragua Mixed Commission* (1963).

Economy: CONTINENTAL-ALLIED COMPANY, INC., *Helping Honduran Industry: a Diagnostic Study* (1961), is useful, while V. CHECCHI, *Honduras: A Problem in Economic Development* (1959); J.P. COGHILL, *Honduras, June 1954: Economic and Commercial Conditions in Honduras* (1954); A.C. MCKECHNIE, *Notes on a Visit to Panama, Honduras, El Salvador and Guatemala* (1953), brief and factual; and G.E. STOCKLEY, *Honduras: Economic and Commercial Conditions in Honduras, May, 1951* (HMSO 1951), point to the historical problems.

Cultural: R.N. ADAMS, *Cultural Surveys of Panama—Nicaragua—Guatemala—El Salvador—Honduras* (1957).

(J.R.M.R.)

Hōnen

The founder of the Jōdo-shū, or Pure Land sect of Japanese Buddhism, Hōnen is one of the outstanding figures in the history of Japanese religion. Through his personal influence and the direct simplicity of his teaching, he played a seminal role in making Pure Land pietism one of the central forms of Buddhism in Japan. He was also called Genkū, Enkō Daishi, Ganso (Founder), and Hōnen Shōnin (Saint Hōnen).

Hōnen was born in 1133 at Inaoka, in western Honshu, the only son of Uruma Tokikuni, the regional military chief. His childhood name was Seishi-Maru. When he was nine years old, his father was attacked and killed by the headman of the village. The dying father instructed the son not to spend his life trying to avenge him but to enter the priesthood. After a period of local instruction in which he displayed marked ability, at 15 Hōnen was sent to Mt. Hiei, the monastic centre of the Tendai sect of Buddhism. The centre prospered externally, in wealth and prestige, but suffered from the internal power struggles of ambitious abbots and the moral and spiritual corruption of the priests.

Hōnen, along with other serious-minded young priests, however, came under the influence of the Pure Land doctrine, which taught salvation by the mercy of Amida (or Amitābha) Buddha, who dwells in the Western Paradise, to which he brings his devotees. Hōnen was haunted by the sense of the human futility of man "with blind eyes, capable of doing nothing" or, in the words of his disciple Shinran, "destined for hell, with nothing in his reach." He was greatly inspired by the *Ōjōyōshū* ("Essentials of Salvation"), by a 10th–11th-century Japanese Amida Buddhist, Genshin, and the *Kuan-ching-su* ("Commentary on the Meditation *sūtra*,") by a 7th-century Chinese Pure Land master, Shan-tao (Japanese Zendō). In 1175 Hōnen, then 43 years old, proclaimed his message that the one and only thing needed for salvation is *nembutsu* (calling the name of Amida Buddha).

Hōnen's basic teaching: the *nembutsu*

In his main work, the *Senchaku hongan nembutsu-shu* ("Book on the Choice of *nembutsu*"), written in 1198, Hōnen classified all the teachings of Buddhism under two headings: Shōdō ("Sacred Way") and Jōdo ("Pure Land"). According to him, Buddha, confident of man's inner character, had shown men the Sacred Way to Enlightenment by means of precepts, meditation, and knowledge, thus enabling them to be emancipated from this world of lust and delusion and to attain the other world of ultimate peace. Hōnen, convinced of his own "sinful and avaricious" nature, however, came to the conclusion that, while it was theoretically possible, it was practically impossible for him and others like him to follow the Sacred Way. The only alternative open for them was the way of Jōdo, which is based on implicit trust in the original vow of Amida Buddha, the lord of the Sukhāvatī (Sanskrit: "Pure Land"), who in his mercy assures salvation to the believer who clings to Amida's merciful hand and calls upon Amida's holy name—*namu Amida Butsu* ("Homage to Amida Buddha")—with all his heart. This is the *nembutsu* ("calling the name") of Hōnen's book.

Hōnen established his headquarters in the midst of the secular city of Kyōto, away from ecclesiastical establishments, and gathered together devoted disciples, including Shinran, who was to become the founder of the Jōdo Shinshū ("True Pure Land sect"). Hōnen and his followers accepted the legendary periodization of Buddhist history, according to which the first 1,000 years following the demise of the Buddha was the period of the "perfect law" (*shōbō*), in which the true teaching prospers; the second 1,000 years was the period of the "copied law" (*zōbō*), in which piety continues but true teaching declines, and the last 1,000 years is the period of the "latter end of law" (*mappō*), in which Buddhism declines and the world is destined to be overwhelmed by vice and strife. It is to be noted that, according to the accepted calculation of the Japanese Buddhists, the last period began in AD 1051. And, as though to substantiate this view of history, Japanese society during the 12th century suffered from political instability and social disintegration that resulted in the establishment of feudal government under the leadership of the warrior class. Understandably, Hōnen's simple teaching found eager followers among the various levels of Japanese society of that time.

The popularity of the faith in the Pure Land of Amida Buddha aroused jealousy from the established schools of Buddhism, and it so happened that two court ladies were converted to Hōnen's teaching and gave up the court form of Buddhism. Hōnen was thereupon charged with their seduction and was banished, together with his immediate disciples, from the capital in 1207 (some of his disciples were beheaded). Compelled to use a nonclerical name, he called himself Fujii Motohiko and proved to be an effective evangelist even during his exile to Shikoku island. He was permitted to leave Shikoku at the end of the year, but not to return to Kyōto until 1211, when he received a warm popular welcome. He died early the next year at the age of 80.

Banishment to island of Shikoku

Although he insisted on faith in Amida and the recitation of the name as the best way to salvation, Hōnen, an intrepid but nonaggressive person, was markedly tolerant and nonpolemical, urging his followers to respect the other Buddhas and other Buddhist ways of faith and practice. Observance of other forms of discipline, indeed, according to him, might be occasions leading toward Amida and his Paradise. Hōnen was also especially careful to warn against the temptation of accompanying the *nembutsu* with an immoral life or of believing that its recitation removes the stain of violations of the Buddhist life discipline or other immoral acts. He combined the cultured heritage of the established Buddhism of the past with the pioneering spirit of the new Buddhism of the 13th century. The Jōdo-shū founded by him continues to be one of the most influential schools of Japanese Buddhism; and the far more numerous Jōdo Shinshū founded by his disciple Shinran adds still more to the Pure Land influence initiated by him.

BIBLIOGRAPHY. The only significant work on Hōnen available in English is H.H. COATES and R. ISHIZUKA, *Hōnen, the Buddhist Saint: His Life and Teaching* (1925).

(Ma.Fu.)

Hong Kong

Hong Kong is a British crown colony situated on the coast of China's Kwangtung Province—China's "South Gate"—at the mouth of the Hsi Chiang Estuary. It consists of the island of Hong Kong and adjacent islets,



Hong Kong Harbour, seen from Victoria Peak, showing the business district of Victoria and Kowloon Peninsula across the bay.
J. Allan Cash—Rapho Guillumette

Stonecutters Island, Kowloon Peninsula on the mainland, and the New Territories—acquired on a 99-year lease from China, due to expire in 1997—which partly consist of mainland territory, and partly of more than 230 surrounding islands. About 90 miles to the northwest is the Chinese port city of Canton (*q.v.*) and about 40 miles to the west across the estuary is the old Portuguese enclave of Macau (*q.v.*).

The total land area of Hong Kong is 404 square miles (1,046 square kilometres), but this figure is not final, since land reclamation schemes occasionally add small bits of land to the total area. Hong Kong Island and its adjacent islets have a total area of 29 square miles; the Kowloon Peninsula and Stonecutters Island together have an area of three and three-quarter square miles; and the New Territories have a total area of about 366 square miles. There are extensive port facilities on the Kowloon Peninsula, while Victoria, the capital city, situated on Hong Kong Island, is built overlooking one of the world's finest natural harbours.

Between the end of World War II and 1971 the population of Hong Kong increased from about 1,500,000 (1947) to about 4,000,000—an increase indirectly attributable to the dramatic changes that occurred during this period in China itself. The increase has caused a serious housing problem but has also led to a modification of Hong Kong's historic function, which was to serve as an entrepôt for trade between China and Western countries. Because in recent years the economic policy of the People's Republic of China has made it increasingly difficult for Hong Kong to survive on trade alone, the city has been obliged to develop industry instead; by 1971 the percentage of Hong Kong's exports locally produced, which was 73 percent in 1960, had increased to 80 percent.

The development of industry

That Hong Kong has been able to maintain its present political status is due to the interplay of national interests. China wishes to keep Hong Kong as a point of contact with Western countries, as well as for economic reasons; yet Western countries, which also contain economic advantages from Hong Kong, use the colony as a station from which they can monitor information emanating from China. (For an asssociated physical feature, see HSI CHIANG; for historical background, see CHINA, HISTORY OF.)

The landscape. *Physiography.* The mainland and islands of Hong Kong form part of a drowned mountain range composed of volcanic and sedimentary materials atop a vast, subterranean expanse of granitic rock extending from southeastern China southwestward into the Indochinese Peninsula. The submersion is comparatively recent in geological terms since the overlying rocks date only from the Jurassic Period (136,000,000 to 190,000,000 years ago); hills rise directly from the sea with no intervening coastal flats. The numerous mountain peaks are mostly composed of weather-resistant volcanic rock. The highest peak, Tai Mo Shan, on the mainland, reaches an altitude of 3,140 feet (957 metres).

Geological structure

The colony has a wide variety of landform features. Erosion along the coast is marked and has produced many sea cliffs that were once steep hillsides, as well as many sea caves, rock tunnels and pinnacles, and other features. River courses are strongly etched into the landscape; hillsides are characterized by scars from landslides, gullies, and small terraces with steep sides.

The terrain of Hong Kong is mostly hilly, and in places very steep. In general, the colony can be divided into two distinctly different parts, with the Kowloon ridge as the dividing line. The harbour area and the commercial, industrial, and residential districts are to the south, while the agricultural area is restricted to the north and occupies some of the islands.

Drainage. Most of the longer streams in the New Territories flow to the northwest, while a large proportion of the surface water flows across the alluvial plain to drain into the Sham Chun River and Deep Bay in the northwest. The Sham Chun River, which forms the boundary between China and Hong Kong, is the longest river in the colony; being tidal, its water is brackish quite a long way up its course.

Soils. Around Deep Bay, which receives most of the rivers, the richest farmland in the colony lies amid fertile deposits that form part of the Chu Chiang (Pearl River) delta. Soils of this alluvial plain, much of which has been reclaimed during the present century, exhibit marine characteristics as much as three miles from the coast. Elsewhere, rapid weathering of granites and volcanic rocks, because of heavy rains and high temperatures, has either produced less fertile soils, which tend to be clayey and lateritic (leached and iron bearing), or has resulted in an altogether denuded terrain. In general, the natural residual soils are acid and of low fertility, needing the addition of lime, potash, and superphosphates for cultivation.

Climate. Hong Kong, lying at the northern limit of the tropical zone, has hot, humid summers, and cool, dry winters. Mean January and July temperatures are, respectively, about 60° F (16° C) and 82° F (28° C); the yearly average is 70° F (21° C). Normal annual rainfall is about 85 inches, of which more than half falls during June, July, and August and 90 percent between April and October. Typhoons occur mostly between July and October and occasionally in November; they bring torrential downpours and strong winds that often cause damage. In the winter, frost occurs occasionally in higher altitudes.

The typhoon season

Vegetation and animal life. Hong Kong has a year-long growing period. Centuries of cutting and burning have destroyed Hong Kong's original cover, leaving only about 11 percent of the land forested; bush fires consume large parts of this forested area during the annual dry season. Strict rules against cutting and intensive reafforestation programs have restored some of the stands of pine, eucalyptus, banyan, casuarina, and palm trees. Most of the land, however, remains under tropical herbaceous growth, including mangrove and other swamp cover. The heavily eroded badlands (barren areas) are an exception. Altogether, about 50 square miles of land are under cultivation.

Hong Kong's animal life consists of a mixture of mammals adapted to the tropics, and those adapted to a temperate or subtropical climate; there are at least 36 species of wild mammals, of which the most numerous are bats and various types of rodents. There are a few wild deer and monkeys, but since 1950 tigers and leopards are no longer to be seen.

Local cattle and buffalo are kept mainly for farm work.

Pigs and poultry are the principal animals reared for food.

Patterns of human settlement. As mentioned, the Kowloon mountain range at the northern end of the peninsula is the dividing line between urban and rural Hong Kong.

Hong Kong's urban development has been limited by a shortage of flat land. The principal urban areas are on Hong Kong Island and on Kowloon Peninsula, both of which are hilly. The built-up area on the island generally lies below a height of 600 feet; in Kowloon the built-up area has been extending northward towards the hills in recent years. As the total utilizable land of urban Hong Kong amounts to no more than 19 square miles, land reclamation has proved a profitable enterprise.

Land reclamation

Long continued and large-scale land reclamation has resulted in changes in the coastline. The filling material is usually stone and earth quarried from the adjacent hills; such quarrying has resulted in changes in local land forms. Because the coastal waters are relatively deep, the reclaiming of land from the sea is costly; but since the demand for land is great, and land prices are consequently extremely high, investment yields handsome returns. Many major streets, especially those on Hong Kong Island, are built on reclaimed land, as is much of Victoria's central business district; the industrial district of Kwun Tong, developed since 1958, is entirely built on reclaimed land.

Urban development on Hong Kong Island, especially in the western part, appears to have reached the saturation point. More vacant space, however, is available in Kowloon, where urban development has been proceeding rapidly in recent years. As Kowloon suffers less restriction from topography, its streets are comparatively wider, flatter, and straighter. Both cities are basically Chinese in character, with business sections having shops on the ground floor and living quarters above. The jumbling together of shops, small industries, and dwellings occurs in only a few scattered areas.

In the New Territories, rural settlements vary from hamlets to small towns. Most of the villages are compactly built, in the form of a square—a pattern which reflects past defense needs. The oldest villages, which have populations of Cantonese Chinese, have a history of continuous settlement from the 11th century; some are fortified with walls and moats. The Cantonese villages are mainly in the flat alluvial regions, and since floods are a constant threat, many houses are raised on earthen foundations; the pits left by the excavations are used as ponds. The villages of the Hakka people from N.E. Kwantung and Fukien provinces are usually in narrow valleys and on foothills; as a result the typical Hakka village, instead of being square, is long or round.

Features of the landscape

The *feng-shui* grove and the pond are characteristic features of both Cantonese and Hakka villages. The grove is generally planted above or around one side of the village as a kind of amulet to ward off evil, appealing to the *feng-shui*, the spirit of wind and water residing in the hills. The pond receives drainage and sewage from the village, which is usually on slightly raised ground. The water is used to irrigate fields, and its mud is used to fertilize the soil. Fish are also bred in the ponds.

The building of new roads and the advent of cooperative marketing created a number of large market towns in the New Territories. With the growing prosperity of vegetable farming, there has been a trend toward the establishment of scattered farmsteads so that farmers can live closer to their crops. A great number of recent immi-

MAP INDEX

Cities and towns

Aberdeen.....22·15n 114·09e
Cha Kwo Ling..22·18n 114·13e
Cheung Shui Tan.....22·26n 114·12e
Chik Kang.....22·26n 114·21e
Chuen Lung....22·24n 114·06e
Hang Hau Town.....22·19n 114·16e
Ho Chung.....22·22n 114·14e
Hok So Wan....22·13n 114·14e
Ho Poi.....22·25n 114·03e
Hung Hom.....22·18n 114·11e
I Pak.....22·19n 114·00e
Kam Tin Hui....22·27n 114·03e
Kowloon.....22·18n 114·10e
Kowloon City...22·19n 114·11e
Lam Uk Wei....22·26n 114·22e
Lan Nai Wan....22·24n 114·20e
Long Ke.....22·23n 114·22e
Luk Kang.....22·27n 114·02e
Ma-liu Shui.....22·25n 114·12e
Ma On Shan Tsuen.....22·25n 114·14e
Mui Wo.....22·16n 113·59e
New Kowloon ..22·20n 114·10e
Ngo Ki.....22·18n 113·58e
North Point.....22·17n 114·12e
Pak Kong Tsun.22·23n 114·15e
Ping Shan.....22·27n 114·00e
Sai Kang.....22·26n 114·16e
Sai Kung.....22·23n 114·16e
Sek Kong.....22·26n 114·06e
Sha Tin New Town.....22·23n 114·11e
Siu Lik Yuen....22·23n 114·12e
Stanley.....22·13n 114·12e
Tai Hang Village.....22·17n 114·11e
Tai Lam Chung.22·22n 114·01e
Tai Long.....22·13n 113·59e
Tai Long.....22·25n 114·23e
Tai Po Tsai.....22·21n 114·15e
Tai Shui Hang..22·25n 113·56e
Tai Tong.....22·25n 114·01e
Tai Wan Tau....22·18n 114·17e
Tai Wan Tsun...22·19n 114·12e
Ta Mong Tsai...22·24n 114·18e
Ting Kau.....22·23n 114·04e
Tin Sam.....22·22n 114·11e
Tiu Keng Wan..22·18n 114·15e
Tong O.....22·12n 114·08e
Tsim Sha Tsui..22·18n 114·10e
Tsun Wan Wai..22·22n 114·07e
Victoria.....22·17n 114·09e
Wong Ka Wai...22·24n 113·58e
Yuen Long Kau Hui.....22·26n 114·02e
Yung Shu Wan.....22·14n 114·06e

Physical features and points of interest

A Peak.....22·27n 114·18e
Aplichau, *island*.....22·15n 114·09e
Basalt Island....22·19n 114·22e
Beacon Hill.....22·21n 114·09e
Bluff Island.....22·19n 114·21e
Bluff Point.....22·11n 114·12e
Botanic Gardens.....22·17n 114·09e
Cheung Chau Island.....22·12n 114·01e
Cheung Kwan O, *bay*.....22·17n 114·15e
Chik Chu Wan, *bay*.....22·12n 114·12e
Clear Water Bay.....22·17n 114·18e
Cross Harbour Tunnel.....22·18n 114·10e
East Brother, *island*.....22·20n 113·58e
East Lamma Channel.....22·15n 114·07e
Fu Tau Pun Chau, *island*...22·21n 114·22e
Grassy Hill.....22·25n 114·09e
Government House.....22·17n 114·09e
Government Stadium.....22·16n 114·11e
Hak Kok Tau, *cape*.....22·16n 114·15e
Happy Valley Race Course.. 22·16n 114·10e
Hong Kong, *island*.....22·15n 114·11e
Hong Kong, University of...22·17n 114·08e
Hong Kong Harbour (Victoria Harbour).....22·17n 114·10e
Jubilee Reservoir.....22·23n 114·08e
Kai Tak Airfield.....22·20n 114·12e
Kau Lung Peak.....22·21n 114·13e
Kau Sai Chau, *island*.....22·22n 114·18e
King's Park.....22·19n 114·10e
Kiu Tsui Chau, *island*.....22·22n 114·17e
Kowloon Peninsula.....22·19n 114·10e
Lamma Island, see Pok Liu Chau
Lan Tao, *island*.....22·17n 113·59e
Lei U Mun, *channel*.....22·16n 114·14e
Lion Rock Tunnel.....22·21n 114·09e
Lo Chau, *island*.....22·11n 114·15e
Long Harbour...22·27n 114·20e
Lower Shing Mun Reservoir.....22·23n 114·09e
Lung Shun Wan Chau, *island*...22·22n 114·21e
Ma On Shan, *mountain*.....22·25n 114·15e
Ma Wan Island..22·21n 114·03e
New Territories, *historical region*.....22·24n 114·10e
Ngang Kong, *bay*.....22·16n 114·00e
Ngon Shun Chau (Stonecutters Island).....22·19n 114·08e
Ninepin Group, *islands*.....22·16n 114·21e
North Ninepin Island.....22·16n 114·20e
Papai, *island*....22·15n 114·02e
Ping Chau Island.....22·17n 114·02e
Pok Liu Chau (Lamma Island).....22·12n 114·07e
Port Shelter, *bay*.....22·21n 114·17e
Po Toi Group, *islands*.....22·11n 114·16e
Po Toi Island...22·10n 114·15e
Rocky Harbour..22·20n 114·19e
Royal Observatory....22·18n 114·10e
Sam Kong, *island*.....22·11n 114·17e
Sek Kong Airfield.....22·27n 114·05e
Sha Tin Hoi, *bay*.....22·24n 114·12e
Shek Ku Chau, *island*.....22·12n 113·59e
Shelter Island...22·20n 114·17e
Shing Mun, *river*.....22·23n 114·10e
South China Sea.....22·13n 114·20e
South Ninepin Island.....22·15n 114·21e
Stanley Mound, *hill*.....22·14n 114·12e
Stonecutters Island, see Shun Chau
Sui Kau Island.....22·16n 114·03e
Tai Lam Chung Reservoir.....22·23n 114·01e
Tai Long Head.....22·12n 114·15e
Tai Long Wan, *bay*.....22·24n 114·24e
Tai Mo Shan, *mountain*.....22·25n 114·07e
Tai Po Hoi, see Tolo Harbour
Tathong Channel.....22·15n 114·15e
Tathong Point.....22·14n 114·17e
Technical College.....22·19n 114·10e
Tiu Chung Chau, *island*.....22·20n 114·19e
Tolo Harbour (Tai Po Hoi)...22·26n 114·12e
Tsing Island....22·21n 114·05e
Tsin Shui Wan, *bay*.....22·13n 114·10e
Tung Lung, *island*.....22·15n 114·17e
Unicorn Ridge..22·22n 114·11e
Victoria Harbour, see Hong Kong Harbour
Victoria Park....22·17n 114·07e
Victoria Peak...22·17n 114·08e
Wang Lan, *island*.....22·11n 114·18e
West Lamma Channel .....22·13n 114·04e

Hong Kong, Area and Population

	area		population	
	sq mi	sq km	1961 census	1971 census*
Total Hong Kong†	404	1,046	3,133,000	3,948,000

*Provisional. †No first-order subdivisions are given because of inconsistencies of their boundaries and populations in various country sources.
Source: Official government figures.

grants from China have settled along the major highways.

The people. *Ethnic composition.* About 99 percent of Hong Kong's people are Chinese, many of them from neighbouring Kwangtung and Fukien provinces. About half the population, despite this immigration, was born in Hong Kong. As in China itself, a great variety of regional peoples and linguistic groups are represented, nearly all of them maintaining separate communities even in the urban areas. The most numerous, by far, are the Cantonese, whose dialect is universally understood and who are both rural and urban in their way of life. Other Chinese peoples include the Hakka, who are mainly hill farmers, and the Hoklo and Tanka. The latter two groups traditionally have comprised Hong Kongs's boat-dwelling population. In recent years, however, many thousands have moved ashore; some Hoklos have been landsmen for generations. Many of the current boat dwellers belong to neither group. The major non-Chinese elements in the population are from the British Commonwealth countries, the United States, Portugal, and Japan.

Religious groups. The religious persuasions of the people of Hong Kong are as complex as their political ideas. About 10 percent of the population are communicants of various Christian sects, but Islām claims only a few thousand adherents. Buddhism and Taoism have far outstripped other groups during the population growth of recent decades, and their many temples and monasteries play an important part in the life of the average Chinese. In each of the temples there is usually a principal deity after whom the temple is named, as well as other deities such as the sea gods and goddesses who were worshipped when Hong Kong was a major fishing port. The best known of the major deities is T'ien Hou, the Goddess of Heaven and patroness of sailors. A temple to her stands at nearly every entrance to a fishing harbour. Each year a celebration honouring her birthday is held with fanfare in the temple on the northern bank at the eastern entrance to Hong Kong Harbour.

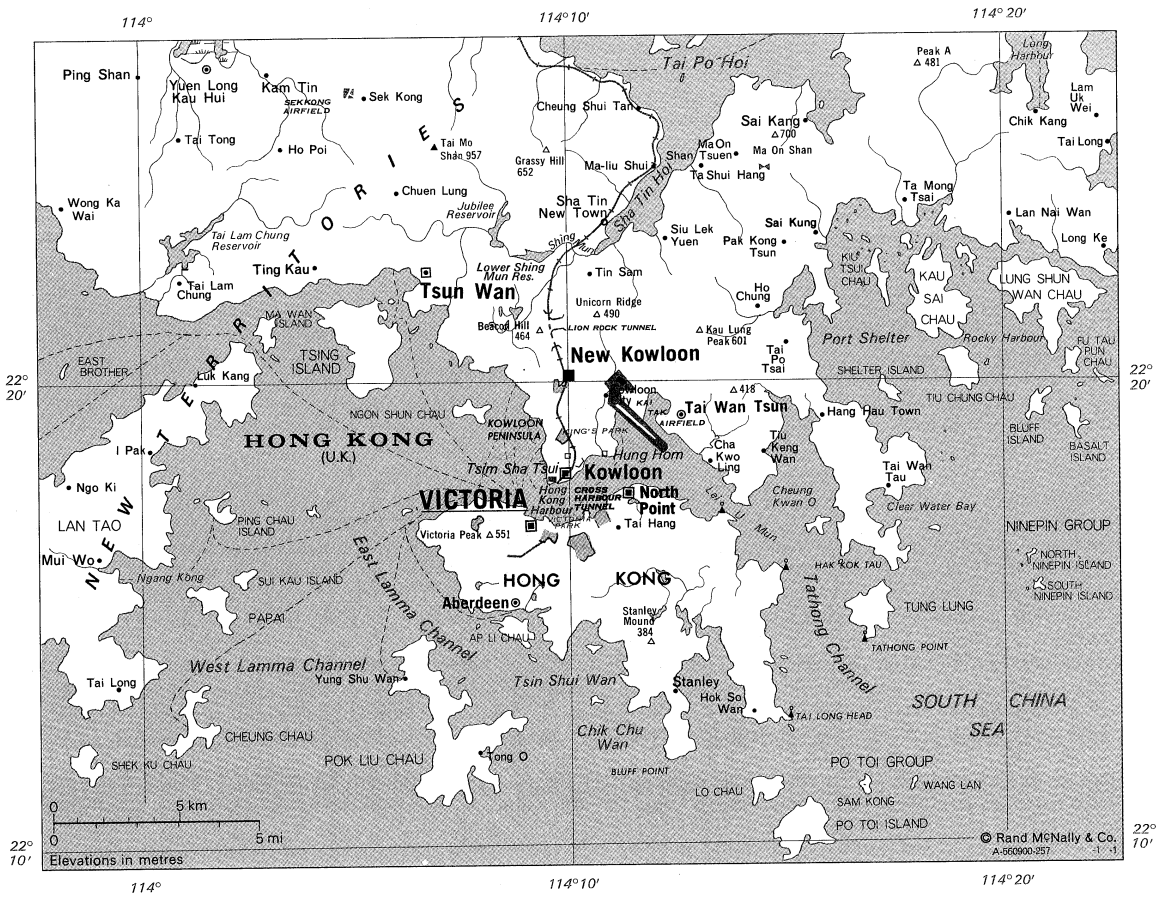
Buddhist and Taoist groups

Demography. Because of the influence of birth control and family planning programs, the birthrate, which in 1960 stood at 36 per 1,000 people, began to decline in 1961. By 1971 the figure had dropped to 19 per 1,000 but was still far ahead of the death rate, which was about 5 per 1,000. Infant mortality remains relatively high, however, standing at 18 per 1,000 live births. Nearly half the population is under 20 years of age.

About 87 percent of Hong Kong's people live in the urban areas of the colony, where industry has continued to attract the youth from the farming areas. Only a minority of the rural people are engaged in farming. In 1971, the total floating population of Hong Kong amounted to 80,000 people who were living in 10,000 boats.

The economy. The decisive factor forcing the change from a trading to an industrial economy was the adherence of Hong Kong in 1950 and 1951 to a United Nations' embargo placed on trade with China and North Korea during the Korean War. The transition was so rapid that few persons at first realized its significance. The needs of the free port are now met by a complex of light industry that flourishes in international trade competition without governmental subsidy or protection. Not only does the colonial government not protect local industry from competing imports but neither does it re-

Governmental economic policy



HONG KONG

taliate against countries that impose trade restrictions on Hong Kong's exports.

Resources. As the colony is not rich in natural resources, industry depends almost entirely on imports of raw material. A small iron mine is active at Ma On Shan, situated about 8 miles northeast of Kowloon in the New Territories, and there are deposits elsewhere of kaolin clay for pottery, graphite, lead, and other minerals. Most ore is shipped to Japan. The dependence on surface water has created serious shortages in the past, but agreement with China on the provision of a supply and the building of many new reservoirs and storage facilities are helping to overcome this problem. Electrical power is supplied to Hong Kong Island and the neighbouring islands of Ap Lei Chau and Lamma by plants at North Point and Ap Lei Chau; Kowloon and the New Territories are supplied by generating plants at Hok Yuen and on Tsing Yi Island. Fish is one of Hong Kong's major resources, used not only for domestic consumption but for export as well.

Sources of national income. The rapid industrial development in the 1950s was made possible almost entirely by money and technology contributed by Chinese industrialists mainly from Shanghai. Overseas investments, however, especially from Japan and the United States during the 1960s, began to have an effect upon the development of sections of the textile, garment, and electronic industries. One of Hong Kong's major economic assets has been its huge supply of labour, although this economic advantage is associated with serious social problems of overcrowding.

Manufacturing. The garment and textile industries contribute 45 percent of Hong Kong's exports and employ 45 percent of the industrial workers. Most of the yarn production is used for weaving fabrics on local looms and making garments for the world market. Other major light industries produce electronic equipment, toys, wigs, plastic flowers, and rubber products. Heavy industry includes shipbuilding and repairing, steel rolling and aluminum extrusion, and cement manufacture. Most of the industry is confined to Kowloon and Hong Kong Island, but some of the towns in the New Territories have a brisk industrial activity as well.

The decline in farming

Agriculture and fishing. Farming acreage continues to decline, even in the production of rice, the largest single crop. The majority of vegetables grown are the Chinese leaf types, and the local farms supply more than 40 percent of the colony's needs. As the growing season is year round, from four to ten crops per year are possible, depending on the crop. Many farmers also keep pigs and poultry. Large livestock farms are able to operate in areas unsuitable for crops. Dairy cattle are usually stall fed, and rarely leave their stalls.

The catching, processing, and exporting of marine fish is a major industry, forming an important element in Hong Kong's economy. Pond fish are also marketed. In Deep Bay, edible oysters are grown for export.

Financial services. The unit of currency is the Hong Kong dollar. At the end of 1971 there were 70 incorporated banks in Hong Kong with 431 branch offices; these included banks and other financial institutions from the Western nations. In the same year there were six stock exchanges which had ties with stock exchanges in Europe and the Americas.

Exports and imports

Foreign trade. The crowded harbour has long provided visual proof of Hong Kong's position in world trade. The United States is the main market for the colony's products, receiving 35 percent of Hong Kong's exports in 1971, compared to approximately 12 percent to the United Kingdom, the second largest purchaser. The major imports are foodstuffs and raw material for local industry. Alcoholic beverages and tobacco are the only major imports on which a tariff is levied. Imports come primarily from Japan, China, and the United States, in that order. Re-exported goods are important, accounting for some 20 percent of total exports. Much of this transshipment is of products from China, a factor that contributes to the relative harmony characterizing relations between the colony and its powerful neighbour.

The government seeks to create favourable conditions for private enterprise. Most semiskilled and unskilled workers in the manufacturing industries are usually paid for piece-work (*i.e.*, at a standard rate for each unit produced), although daily rates of pay are also common. Men and women receive the same rates for piece-work, but women are generally paid less when engaged on a time basis. There are no legal restrictions on hours of work for men, although most men employed in industry work ten hours a day or less. The restrictions on the hours of work for women were introduced in 1959.

With the exception of a small independent segment, workers' unions are either affiliated to, or associated with, one of two local federations. At the end of 1971 there were 276 workers' unions with a declared membership of about 198,000.

Transportation. The 22-mile Hong Kong section of the Kowloon–Canton Railway runs from the southern end of the Kowloon Peninsula to the Chinese border at Lo Wu, and thence into mainland China. Since 1949, passengers have had to change trains at the frontier and walk about 300 yards between the two terminals, although no transshipment is required for mail and cargo traffic. Traffic, especially of passengers, is frequent, and is heavier at weekends and on public holidays. A new railway terminus at Hung Hom (at the northern tip of Hung Hom Bay, Kowloon), is planned to replace the present one at Tsim Sha Tsui on the southern tip of the Kowloon Peninsula.

There are over 600 miles of roads in the colony, of which 200 miles are on Hong Kong Island, 170 miles in Kowloon, and the remainder in the New Territories. Nearly 150,000 vehicles are registered.

With the exception of the Kowloon–Canton Railway, public transport, which includes bus and tramway services, is operated by private interests. Of all the major metropolitan centres in the world, the cities of Hong Kong may doubtless lay claim to having the most crowded, noisy, and filthy buses.

Ferry services

A fleet of diesel-powered ferries, including several car ferries, connect Hong Kong Island, Kowloon, and the New Territories, as well as the outlying islands; another ferry service runs regularly between Victoria on Hong Kong Island and Tsim Sha Tsui on the southern tip of the Kowloon Peninsula. Both carry many millions of passengers each year.

Vessels arriving and leaving Hong Kong each year include thousands of oceangoing ships, river steamers, junks, and mechanized vessels of various kinds. An ocean-passenger terminal, containing more than 100 shops, together with restaurants and night clubs, was completed in 1966. The repair and maintenance of shipping is carried on at two major dockyards. Hong Kong is also a centre for the training of thousands of merchant seamen.

Major international airlines use the international airport at Kai Tak, about three miles from downtown Kowloon.

Administration and social conditions. *Governmental structure.* The governor, appointed by the British crown, is the Queen's representative and head of the executive branch of the colony. He presides over an executive council of five *ex officio* members (*i.e.*, persons who are members by virtue of their official positions) and eight unofficial members appointed by the governor. Of the eight unofficial members holding office in 1971, four were Chinese. Four of the five *ex officio* members also serve on the Legislative Council, of which the governor is president; in addition, the council consists of eight other official members and 13 unofficial members nominated by the governor. None of the members of the Legislative Council are elected. Laws are enacted by the governor with the advice and consent of the council, which controls finances and expenditure.

An urban council, partly appointed and partly elected, administers Victoria and Kowloon. The New Territories are divided into five administrative districts, each under a district officer. For local representation, villages are grouped into Rural Committees, of which there are 27.

The chairmen and vice chairmen of the Rural Committees, together with justices of the peace and counsellors, form the Full Council of the New Territories, which is called the Heung Yee Kuk ("Rural Consultative Council").

The courts consist of the Full Court, the Supreme Court, the District Court, the Magistrates' Court, the Tenancy Tribunal, and the Marine Court. The laws of England, with modifications, are in force. The judiciary, headed by the chief justice, is independent of the administration.

Social conditions. A large population crowded into a tightly restricted area is the primary conditioning factor in Hong Kong's way of life.

Although from 1955 to 1965 the number of persons accommodated in Hong Kong's public housing tracts, known as resettlement estates, increased from about 125,000 to more than 755,000, the squatter population in 1970 nevertheless remained at its 1955 level of about 360,000.

Elsewhere, entire households live in cubicles, bed spaces, attics, roofs, verandas, and similar quarters; well over half of all households share accommodations with others. The government is the largest landlord in the colony, but average rents remain high.

The most serious menace to health in Hong Kong has long been tuberculosis. This and other contagious diseases are spread by overcrowded and often unsanitary living conditions. The government works closely with voluntary health and welfare agencies in promoting large-scale public health programs. It operates a dozen general and special hospitals, and twice as many more are privately operated. Medical service at the government hospitals is not free but is cheap; private clinics are mostly expensive.

With education, as with housing and water supplies, the facilities available are outstripped by the population. All education is voluntary but not free, although even private schools receive public grants. Nearly 1,250,000 students attend schools at all levels. At the lower levels, more than half the schools are privately run. Facilities are particularly strained at college level. The University of Hong Kong, founded in 1911, is private, but receives public subsidies. The Chinese University of Hong Kong, established in 1963, is located at Shatin Shui in the New Territories, and represents the amalgamation of three older colleges. Both institutions are modelled on British systems, and enrollments are limited. Many students seek higher education in the United States, Japan, or Commonwealth countries.

The universities

Cultural facilities. There were 67 daily newspapers in 1971, including 63 Chinese and four English. There have been television services since 1957. Hong Kong is served by three broadcasting organizations, two of which are commercial; the other is a government-operated station. The colony has more than a hundred cinemas.

Prospects for the future. Hong Kong's cities, bustling, vastly overcrowded, and often dirty, are nevertheless actively engaged in commerce and industry. As a result, despite its problems, Hong Kong is experiencing an economic boom. Its future, however, is to a great extent dependent upon the economic and political relations existing between the People's Republic of China and other major powers. The New Territories, for example, are due to be returned to China in June 1997, when the agreement by which they were leased expires; whether the Chinese can wait this long, however, will depend upon how long they are willing to remain deprived of this lost land on the doorway of their South Gate.

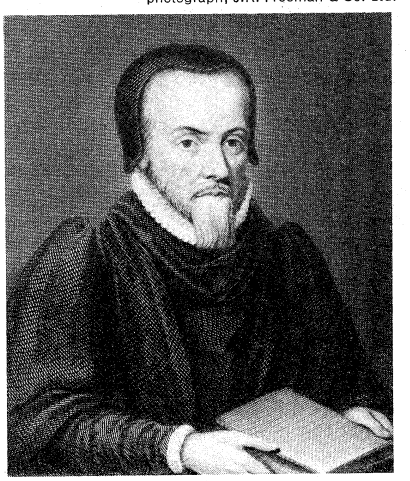
BIBLIOGRAPHY. For geological studies of Hong Kong, see P.M. ALLEN and E.A. STEPHENS, *Interim Report on the Geological Survey of Hong Kong* (1969); and B.P. RUXTON, "The Geology of Hong Kong," *Q. Jl. R. Geol. Soc. Lond.*, 115: 233–260 (1960). A thorough compendium of official statistical data is *Hong Kong Statistics, 1947–1967* (1969), issued by the Hong Kong Department of Census and Statistics. Socio-economic studies include C.S. CHEN, *The Functional Land Use of Urban Areas in Hong Kong*, Research Report, no. 2, Geographical Research Centre (1967); *The Population Distribution and Density of Hong Kong*, Report no. 3 (1967); and *The Socio-Economic Atlas of Hong Kong*, Report no. 20 (1970); E.F. SZCZEPANIK, *The Economic Growth of Hong Kong* (1958); and T.R. TREGEAR, *A Survey of Land Use in Hong Kong and the New Territories* (1958). See also the *Hong Kong Report* (annual); and M.I. BERKOWITZ and EDDIE K.K. POOH (comps.), *Hongkong Studies: A Bibliography* (1969).

(C.-S.Ch.)

Hooker, Richard

Richard Hooker, an Elizabethan and child of the English Renaissance, in *Of the lawes of ecclesiasticall politie*, proved himself to be one of the greatest masters of English prose and of legal philosophy. He created a distinctive Anglican theology and his conceptions of constitutional government were so well founded that they remain congenial to modern thought. Almost nothing is known of Hooker's private life except what is revealed in legal documents about his estate and references to him by his contemporaries.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.



Richard Hooker, engraving by E. Finden (1791–1857) after a print by W. Hollar (1607–77).

Hooker was born toward the end of 1553 or the beginning of 1554 near the city of Exeter. His family did not have the financial means to send young Richard to the university, but, with John Jewel, bishop of Salisbury, as his patron, in 1568 he entered Corpus Christi College at Oxford. At that time in the life of the Church of England, the dominant influence was Calvin's *Institutes of the Christian Religion*, and thus Hooker was trained in the traditions of Genevan Protestantism. Leading scholars at Oxford were, however, loyal to the Anglican *Book of Common Prayer* and used the vestments demanded by the ecclesiastical law of the realm. Hooker, always a staunch Anglican, went far beyond even liberal Calvinism and read widely the best scriptural interpretation of his day, the early Church Fathers, and even Renaissance Thomism (the philosophical school influenced by the thought of St. Thomas Aquinas). He thus avoided the limits of narrow academic Calvinism and became a man of wide Renaissance learning. Hooker said that he grew in his opinions and gave up narrow conceptions previously held.

Years at Oxford

While at Oxford he formed three friendships of lasting importance to him. The first was with John Reynolds, his teacher. Reynolds was a liberal Calvinist, and when, as a fellow of Corpus Christi College, Hooker defended him as a nominee for the presidency of the college against an orthodox Calvinist, both he and Reynolds were temporarily expelled from their fellowships. The second friend was George Cranmer, a lawyer and grandnephew of Thomas Cranmer, Henry VIII's theological reformer and archbishop of Canterbury. Cranmer later helped Hooker with comments on the *Politie*. The third friend was Edwin Sandys, the son of the Archbishop of York, who was knighted and became a member of the Inner Temple, the

great law confederation, and was a member of Parliament. He was very close to Hooker when he composed the *Politie*. Both Cranmer and Sandys were pupils of Hooker.

Master of the Temple

Hooker became a scholar of Corpus Christi College in 1573 and took his M.A. in 1577. In the same year he became a fellow of his college. In 1585 he was elected master of the Temple. The other candidate for this position was Walter Travers, an ardent Calvinist. Travers, an able man, had written a scholarly book, *Ecclesiasticae disciplinae defensio* ("Defense of Ecclesiastical Discipline"). Although he had not received Anglican orders, he was made lecturer (preacher) of the Temple Church. Hooker, a loyal Anglican, preached in the morning, and Travers, a firm Calvinist, in the afternoon. Thus it was said that the Temple Church congregations heard Canterbury in the morning and Geneva in the afternoon.

With the defeat of the Spanish Armada in 1588, the Church of England passed beyond the crisis of the threat of Roman Catholicism. Now the threat was that of Calvinism, not only in doctrine but in ecclesiastical organization as well. Small cells, or conventicals, of reformed worship were formed throughout the realm. Their very radicalism appealed, and their hold on general sympathy was so strong that even the bishops were lukewarm about suppressing them and allowed their growth to increase unchecked. Travers, in fact, set up an organization in the afternoon congregation on the model of the Reformed Church in the Low Countries and chided Hooker for not using the Reformed organization in the Temple Church. The difference between the two men was radical. Hooker did not approve of many of the decisions of the Roman Catholic Council of Trent (1545–63), which attempted to reform the Catholic Church following the Protestant Reformation, but he did approve of many of the medieval Scholastic philosophers and theologians, such as St. Thomas Aquinas, and he used their teaching with approval. This was anathema to Travers, who thought of the teaching of the Scholastics as sheer rubbish. Hooker does not seem to have lived in the parsonage of the Temple, but with John Churchman, a good friend of the Church of England. There were two reasons for this: first, the parsonage was not in good repair, and, second, Travers lived there.

On February 13, 1588, while still master of the Temple, Hooker married Joan Churchman, daughter of his friend and host. Izaak Walton, the English author and biographer, was responsible for the story, accepted for 300 years, that his future father-in-law tricked him into the marriage with his ill-favoured daughter. In 1940 it was proved by examination of the Court of Chancery records about Hooker's estate that the story was a tale devised to explain the incompleted state of the last books of the *Politie*. Joan Churchman brought with her a large dowry. At the time of his marriage Hooker had no known financial means, and yet at his death he left a considerable estate.

His masterpiece: the *Politie*

After he ceased to be master of the Temple in 1591, Hooker resided at his father-in-law's house and wrote his masterpiece, *Of the lawes of ecclesiasticall politie*. The *Politie* was the last link in a long strife, known as the admonition controversy: in June 1572 radical religious reformers had issued from a secret press *An Admonition to the Parliament*, which, though Queen Elizabeth forbade its consideration by Parliament, became the platform of the Puritans—those within the Church of England who wished for radical reforms along the lines developed in Geneva by Calvin. The leading bishops, now alarmed by the influence of the *Admonition*, knew that an answer was needed, and the Archbishop of Canterbury turned to John Whitgift, vice chancellor of the University of Cambridge, to reply to the *Admonition*. This Whitgift did and was answered in turn by Thomas Cartwright, professor at Cambridge and the leading Puritan clergyman. The controversy was continued in a whole series of books.

The *Admonition* was still much in the mind of England when Hooker left the Temple, and he set himself the task of replying to it. The *Politie* was to be a work of eight books, but the fifth book was the last one to appear in Hooker's lifetime. The tradition that his manuscripts were destroyed by Puritan ministers who were assisted by Hooker's wife does not seem to be correct. The incomplete condition of the last books of the *Politie* merely means that Hooker had not yet revised them at the time of his death.

In the *Politie*, Hooker defended the Elizabethan church against Roman Catholics and Puritans alike. He affirmed the Anglican tradition as being that of a threefold cord not quickly broken—Bible, church, and reason. Roman Catholics put Bible and tradition on a parity as the authorities for belief, while Puritans looked to Scripture as sole authority. Hooker avoided both extremes, allowing to Scripture absolute authority when it spoke plainly and unequivocally; where it was silent or ambiguous, wisdom would consult the tradition of the church; but he insisted that a third element lay in man's reason, which should be obeyed whenever both Scripture and tradition needed clarification, or failed to cover some new circumstance. The core of Hooker's thinking on the relations of church and state is unity. In his view, the Puritans adopted an impossible position; they claimed to be loyal to the Queen while repudiating the Queen's church. By law and by reason, an Englishman must be an Anglican, pledged to serve Elizabeth as the supreme magistrate of his country and the supreme governor of his church.

According to tradition Hooker served the churches at Drayton Beauchamp and Boscombe following his term as master of the Temple, but more probably he practiced pluralism, which means he received his salary as a vicar but allowed a lesser clergyman to perform the duties the parish required. In 1595 he accepted an appointment as vicar of Bishopbourne, near Canterbury and in 1597 the fifth book of the *Politie* was published. He died in 1600 and was buried at Bishopbourne.

BIBLIOGRAPHY. C.J. SISSON, *The Judicious Marriage of Mr. Hooker and the Birth of the Laws of Ecclesiastical Polity* (1940), is essential for any biography of Hooker. See also R.A. HOUK's introduction to his edition of book 8 of the *Politie* (1931); *The Works of That Learned and Judicious Divine Mr. Richard Hooker*, ed. by JOHN KEBLE, 3 vol., in the 7th rev. ed. by R.W. CHURCH and F. PAGET (1888), prefaced by the traditional life of Hooker by IZAAK WALTON; P. MUNZ, *The Place of Hooker in the History of Thought* (1952); and JOHN S. MARSHALL, *Hooker and the Anglican Tradition* (1963). For the Admonition Controversy, see A.F.S. PEARSON, *Thomas Cartwright and Elizabethan Puritanism, 1535–1603* (1925).

(J.S.Ma.)

Hooker Family

The Hookers, father and son, were virtually synonymous with English botany for nearly a century. Their many publications, some the product of joint effort, placed them in the forefront of plant taxonomy throughout the world. Although their major contributions involved the technical description and classification of new species, they also had great influence at the popular level. The accounts of their voyages and collecting trips, the introduction of new horticultural varieties, and the founding of The Royal Botanic Gardens at Kew, near London, stimulated the English-speaking world, directly and indirectly, to the most vigorous renewal of botany since the time of the Swedish botanist Carolus Linnaeus in the 18th century.

William Hooker

William Jackson Hooker was born in Norwich, England, on July 6, 1785, son of a merchant's clerk and descendant of Richard Hooker, noted theologian of the 16th century and author of *Of the lawes of ecclesiasticall politie* (1593). A fortuitous discovery in 1805 of a rare moss, which he communicated to James Edward Smith, founder of the prestigious Linnaean Society, redirected his interests from general natural history to botany. His early education at Norwich Grammar School was followed by a voyage to Iceland in 1809, a period of extensive study in England, and a trip to France, Switzerland, and Italy in 1814–15, where he met some of the leading continental botanists. He married Maria Turner, daughter of the botanist Dawson Turner, in 1815. Joseph Dalton Hook-

er, the second of their five children, was born on June 30, 1817, in Halesworth, Suffolk. Three years later, the elder Hooker accepted the chair of Regius Professor of Botany at Glasgow, a position he held until 1841. Until his death at Kew, Surrey, on August 12, 1865, he was actively engaged in promoting the importance of botany.

Beginning with William Hooker's *Journal of a Tour in Iceland in the Summer of 1809*, in 1811, more than 20 major works as well as numerous periodical articles appeared in the following 50 years. His main interest was in cryptogamic botany (*e.g.*, ferns, mosses, fungi), as evidenced by his publications: *British Jungermanniae*, 1816; *Musci Exotici*, 1818–20; *Icones Filicum*, 1829–31; *Genera Filicum*, 1842; and *Species Filicum*, 1846–64. He also published important floristic studies (*Flora Scotica*, 1821; *The British Flora*, 1830; *Flora Borealis Americana: or the Botany of the Northern Parts of British America*, 1840) and was a pioneer in the study of economic botany. The publications, together with his own herbarium, which he generously made available to all scholars, and the journals that he founded and edited, made him the centre of English botany. The climax of his career came in 1841 when he was appointed the first director of Kew Gardens. Under his leadership, Kew Gardens became the world's leading botanical institution. Now a vast complex, including laboratories, a museum, a library, and greenhouses, it is a national showpiece as well as his personal monument. Before his retirement in 1865, he founded the Museum of Economic Botany at Kew (1847).

By courtesy of the Linnean Society of London



William Hooker, painting by S. Gambardella, 1843. In the collection of the Linnean Society of London.

Joseph Dalton Hooker, unlike his father, had the benefit of a formal education and was graduated from the University of Glasgow with an M.D., in 1839. Through his familiarity with his father's herbarium, he was well prepared for the first of his many travels—as surgeon-botanist aboard HMS "Erebus" on the Antarctic expedition of 1839–43. Thereafter, a steady stream of publications followed, punctuated by his own travels: *The Botany of the Antarctic Voyage of H.M. Discovery-Ships Erebus and Terror in 1839–1843*, 1844–60; *Rhododendrons of Sikkim Himalya*, 1849; *The Flora of British India*, 1872–97; *Handbook of the New Zealand Flora*, 1864; and *Journal of a Tour in Marocco and the Great Atlas*, 1878. His last major botanical expedition, to the Rocky Mountains and California (1877), led to the publication of several important papers concerning the relationship of American and Asian floras. His travels resulted in the discovery of species new to science, many of which were soon introduced to horticultural circles. Even more important, however, were the data, which gained him an international reputation as a plant geographer.

Joseph Hooker

In 1851 Joseph Hooker married Frances Henslow, the daughter of a botanist. Six children survived her death in 1874. By his second wife, Hyacinth Symonds Jardine, whom he married in 1876, he had two sons. He became assistant director of Kew in 1855, a position he retained until 1865, when he succeeded his father as director, serving in that capacity until his own retirement in 1885. Many honours came to Hooker, including the presidency of the Royal Society (1872–77) and, like his father, he was knighted for his many services. He remained active until shortly before his death on December 10, 1911, at Sunningdale, Berkshire.



Joseph Hooker, drawing by W. Rothenstein, 1903. In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

One of the most significant results of his travels was an attempt to explain the geographical distribution of plants and their seemingly anomalous variations. As a close friend to Charles Darwin and one well acquainted with the latter's early work, Hooker, along with the geologist Sir Charles Lyell, presided at the historic meeting of the Linnaean Society in July 1858. It was their function to adjudicate the priority claims concerning natural selection as the mechanism for evolution, which had been advanced simultaneously by Darwin and Alfred Russel Wallace. By lending his support to a scientific claim that was soon to be attacked on extra-scientific grounds, Hooker was among the first to demonstrate the importance and applicability of the evolutionary theory to botany in general and to plant geography in particular. The capstone to the younger Hooker's career came in 1883 with the publication of the final volume of the *Genera Plantarum*, written in conjunction with George Bentham. This world flora, describing 7,569 genera and approximately 97,000 species of seed-bearing plants, was based on a personal examination of the specimens cited, the vast majority of which were deposited at Kew. The volume is not only a monument to the Hookers but to the institution they administered for 44 years.

BIBLIOGRAPHY. J.D. HOOKER, "A Sketch of the Life and Labours of Sir William Jackson Hooker (with Portrait)," *Ann. Bot.*, 16:9–120 (1902), a careful study by the son and successor of the biographee containing data not otherwise recorded; MEA ALLEN, *The Hookers of Kew, 1785–1911* (1967), a popular but well written and carefully researched work, based on correspondence and archival materials at Kew; W.B. TURRILL, *Pioneer Plant Geography: The Phytogeographical Researches of Sir Joseph Dalton Hooker* (1953), the fundamental study of the younger Hooker by a leading plant geographer; and *Joseph Dalton Hooker: Botanist, Explorer and Administrator* (1963), popular but important because of the author's long association with Kew and his familiarity with Hooker's herbarium.

(Je.St.)

Hopeh

Hopeh (He-bei in Pin-yin romanization) is a province in northern China, located between the province of Shansi, on the west, and the Gulf of Chihli (Po Hai) and the Northeast (Manchuria), on the east. It is bounded on the north by China's Inner Mongolian Autonomous Region and on the south by the provinces of Shantung and Honan. Hopeh means "north of the (Yellow) river." Hopeh has an area of 74,300 square miles (192,400 square kilometres), or 2.0 percent of the national total. The esti-

mated population in 1970 was 43,000,000. The provincial capital was at Pao-ting until 1958, when it was transferred first to Tientsin and then, in 1967, to Shih-chia-chuang Shih, 160 miles southwest of Peking. The present capital is at the junction of three railways: the Peking–Canton line (the north–south trunk line), the Shih-chia-chuang–T'ai-yüan line to Shansi, and the Shih-chia-chuang–Te-chou line to Shantung. The large municipality of Peking, the national capital, lies within Hopeh Province but is independent of the provincial administration. Culturally and economically, Hopeh is the most advanced province in northern China.

Under the Manchu dynasty (1644–1911), Hopeh was known as Chih-li Sheng ("directly ruled province"), a name that was retained until 1928, when the Nationalist government moved the capital from Peking to Nanking, and the province was renamed Hopeh Sheng (Hopeh Province). From 1937 to 1945 Hopeh was occupied by the Japanese. After Japan's defeat the occupiers surrendered to the Chinese Nationalists in 1945. Chinese Communist forces took the province in January 1949, opening a new chapter in its long history.

Administrative levels

Hopeh Province and the separate municipalities of Peking and Tientsin form three of the 29 primary administrative divisions of the People's Republic of China. The province is subdivided into 11 secondary administrative units, consisting of the provincial capital and ten special districts (*chuan-ch'ü*). The subprovincial administrative areas are further subdivided into 142 counties (*hsien*), eight municipalities (*shih*), and two autonomous counties (*tzu-chih-hsien*). The traditional sub-*hsien* administrative unit was the *hsiang* (civil township, or rural district), which was supplanted in 1958 by the commune. The communes were envisaged as multiple-purpose basic units of rural organization and development: administrative, economic, social, even military. Numerous communes have established labour-intensive rural industries. The profits from these enterprises have helped to pay for primary and secondary education, social welfare, recreation, and medical care, thus achieving a certain degree of self-sufficiency.

LAND AND POPULATION

Relief. Hopeh Province consists of two almost equal sections: the northern part of the North China Plain and the mountain ranges along the northern and western frontiers. The former is sometimes called the Hopeh Plain. It is formed largely by the alluvial deposits of the five principal tributaries of the Hai Ho (Hai River), which flows past Tientsin to the sea. Two of them, the Yung-ting and the Pai, flow down from the northern highlands. The other three have their sources in the western part of Hopeh: the Ta-ch'ing, the Tzu-ya, and the Southern Grand Canal (Nan Yün Ho).

The Hopeh Plain slopes gently from west to east. It is bounded by the Yen Shan (Yen Mountains) on the north, the T'ai-hang Shan to the west, and the Gulf of Chihli to the east. The mountains have at their base a string of alluvial fans. This inner belt of the Hopeh Plain is generally well drained. The groundwater level is usually less than 33 feet from the surface and is easily tapped for domestic water and irrigation.

The Yen Shan range forms the northern rim of the North China Plain, displaying to the traveller an endless sea of rounded hills, with peaks averaging 4,900 feet above sea level. The Great Wall of China zigzags along its crests. Beyond these mountains the Mongolian Plateau stretches from the northernmost part of Hopeh Province to the Mongolian People's Republic. This part of Hopeh has an average elevation of 3,900 to 4,900 feet. The rim of the plateau is rugged and inhospitable to human settlement. Between the Yen Shan are large basin plains, cultivated and well-inhabited. Coal and iron are mined in the northern mountains.

To the west of the North China Plain sprawls the lofty north–south range of T'ai-hang Shan (Grand Run Mountains), separating the Hopeh Plain from the Shansi Plateau. Its general elevation is over 3,280 feet, but the highest peaks are 6,560 feet. The range is pierced by a number of west–east streams whose narrow valleys (the famous "Eight Gorges" of T'ai-hang) are the routes of highways and railroads between the Hopeh Plain and the Shansi Plateau.

Rivers

The major Hopeh rivers flow down from the loess-covered T'ai-hang Shan and the Shansi Plateau. They carry a heavy load of silt after the summer downpours, depositing it in the shallow channels downstream on the plain, gradually silting them up and causing widespread floods in low-lying areas. Under the People's Republic, vigorous measures for water control and soil conservation have been carried out, together with reforestation in the upland areas. Numerous dams, generally small to medium-size, have been built upstream and in the tributaries to conserve the water for irrigation and other uses; flood-retention basins and storage reservoirs have been built downstream. The Tu-liu-chen Canal, connecting the Ta-ch'ing Ho to the sea, helps to drain the extremely low-lying tract around the large Pai-yang Tien (Pai-yang Lake) and the Wen-an Wa (Wen-an Marsh). Water from the streams is utilized to wash away the excess salt in the alkaline soil and make it arable. Similar *chien-ho* ("reducing streams") have been completed for the Southern Grand Canal, and others have been under construction on the Pei Ho.

In the southernmost section of the province, the floodplains have a compact, sticky, waterlogged surface, which makes well digging difficult. The areas between the alluvial fans and the sea in this section are very flat, and there are a number of shallow lakes and marshes of varying degrees of salinity. The groundwater is generally alkaline and unfit for irrigation or domestic use.

The Hai Ho is only 35 miles long, from the city of Tientsin to the sea, but the drainage basin of its five tributaries covers over two-thirds of the province. Another major river is the Luan, draining northeastern Hopeh.

All the major Hopeh rivers empty into the Gulf of Chihli, a shallow sea with an average depth of only 100 feet. The water and nutrient matter brought down by the rivers nourish a rich marine fauna. In winter the surface water along the coast is frozen, but navigation is possible with the use of icebreakers. There are three important ports: Tientsin, which is about 30 miles up the Hai Ho; T'ang-ku, which has a new harbour at the river's mouth; and the major coal-handling port of Ch'in-huang-tao.

Climate and soil. The province has a continental climate. The January mean temperatures range from 25° F (−4° C) in the south to 14° F (−10° C) north of the Great Wall. The average July temperature is about 77° F (25° C) in the North China Plain, and 73° to 77° F (23° to 25° C) in the northern and western highlands. The annual precipitation (rain and snow) is more than 20 inches (500 millimetres) in most parts of the province. The summer months of June, July, and August are the rainy season.

The most common soil in the Hopeh Plain is dark-brown earth developed on loessic alluvium, modified by cultivation over several millennia. It is extremely fertile; the famous "good earth" has yielded crops with little fertilization for thousands of years. New alluvium is distributed in the areas along the rivers by frequent flooding. In the mountains the soils vary: the upland hills have leached dark-brown soils; the more humid mountainous areas of Yen Shan and T'ai-hang Shan have brown forest soils suited to fruit trees; the northernmost Chang-pei plateau has light-chestnut zonal soils.

The natural vegetation of the greater part of the province is broadleaf deciduous forest, but, after many centuries of human settlement, cultivation, and deforestation, little of the original vegetation remains except in the high mountains and other inaccessible areas. The northernmost Chang-pei plateau has steppe grass of the Mongolian Plateau type. The higher mountains have coniferous forests. In the saline areas along the coast and in the low-lying depressions, plants that flourish in a salty environment dominate. There is a conspicuous absence of forests in the lowlands and lower hills. The flora is predominantly of a northern character. It includes the willow, the elm, the poplar, the Chinese scholar tree

(*Sophora japonica*), the tree of heaven (*Ailanthus*), and drought-resistant shrubs.

Animal life

The North China Plain has been inhabited by man for several millennia. The present fauna include elements of the temperate forest (such as the forest cat *Felis euptilus*) and of the cold-winter steppe (such as the camel), as well as some tropical elements from the Indo-Malay region (such as the tiger and the monkey). The domestication of animals such as the dog, sheep, goat, cow, horse, donkey, mule, camel, and cat has led to the extinction or near-extinction of many wild species. The smaller mammals are better preserved, including moles, bats, rabbits and hares, rats, mice, and squirrels. Birds include the Mandarin duck (*Aix galericulata*), native to China. The Hopeh Plain was the home of *Sinanthropus pekinensis* (Peking man), one of the oldest known examples of fossil man, who lived about 400,000 years ago. The Peking man used tools and fire.

The people. The only complete census made in China in the first 70 years of the 20th century was that of June 30, 1953. It recorded 35,984,644 inhabitants in Hopeh Province, not including Peking Municipality (2,768,149), the national capital district. More detailed demographic data were not released. The population estimates of 1954 to 1957 were based on the annual reports of the provinces and municipalities, as collated by the State Statistical Bureau. On December 31, 1957, the registered population of Hopeh Province was estimated at 44,720,000, including Tientsin and Peking municipalities. The estimated population figures for 1970 were 43,000,000 for Hopeh, 4,300,000 for Tientsin Municipality, and 7,600,000 for Peking Municipality. These implied an increase of about 17,920,000 in 17 years. The ethnic composition of the population is almost entirely Chinese. Minority groups include 500,000 Hui (Chinese Muslims) and slightly over 10,000 Mongolians.

The population density of Hopeh in 1970 was approximately 597 inhabitants per square mile. This was less than that for Japan (726 per square mile) but greater than the population densities of the United Kingdom (573) and the state of New York (367). Since nearly one-half of Hopeh Province is mountainous, the density of population is really much higher than the average suggests. Its economic development, moreover, is far below that of New York, the United Kingdom, or Japan. At the start of the 1970s, the highest population densities in Hopeh were found in the Shih-chia-chuang and the Ting-hsien areas at the foot of T'ai-hang Shan, in the belt of alluvial fans, where irrigation by wells is widely practiced; there, densities of 1,036 to 1,300 persons per square mile were attained. This is a district settled since antiquity, on the ancient highway from the middle of the North China Plain (Chung-yüan, or the "Middle Plain") to Peking and on to the north of the Great Wall. The other areas along the alluvial fans of T'ai-hang Shan and Yen Shan also had very high densities, ranging from 777 to 1,036 persons per square mile. These piedmont plains have also been settled since ancient times. The rural settlement pattern is that of huge nucleated villages of more than 1,000 inhabitants. Farther east and south of the alluvial-fan belt are the low-lying districts subject to flood, which had somewhat lower densities (518 to 777 per square mile).

Among the poorer regions is an area along the Gulf of Chihli coast, in the shape of a gigantic crescent, where the population was around 250 per square mile. This section has many saline and alkaline tracts. The northwestern uplands of the province had a comparable population density, except for the arable basin plain (Hsüan-hua and Wei-hsien basins) and the intermontane valley plains (Ch'eng-te and Luan-p'ing), where it increased to between 250 and 500 per square mile. The Chang-pei plateau north of the Great Wall and the remote mountainous areas had the lowest densities, ranging from around 26 to 130 per square mile.

Before the Communist victory there was substantial migration from northwestern Hopeh to Inner Mongolia, but afterward most of this ceased. Peasants in southeastern Hopeh have also migrated in large numbers since the beginning of the 20th century to Manchuria, Inner Mongolia, and China's Northwest. This has apparently continued under the Communists, in a more organized way. With the expansion of large cities such as Peking and Shih-chia-chuang, there has also been an organized migration of peasants from the urban fringes to the Northeast (Manchuria) and the Northwest.

URBAN CENTRES AND ECONOMIES

City dwellers constitute about 20 percent of the total population of Hopeh, the highest proportion in North China, exceeded only by Kiangsu among the provinces south of the Great Wall. The Peking Municipality has an estimated population of 7,600,000 (1970), and Tientsin Municipality has 4,300,000 (1970).

Peking, Tientsin, and T'ang-shan

Major metropolitan complexes. The Peking–Tientsin industrial region, extending from the base of Yen Shan and T'ai-hang Shan to the sea, is the largest and most important in North China. It includes the important industrial cities of T'ang-shan and Ch'in-huan-tao, in eastern Hopeh, Pao-ting, and Shi-chia-chuang, in western Hopeh, and Liu-li-ho, in Peking Municipality. At the present stage of China's industrialization, manufacturing is predominantly concentrated in the large and medium-sized cities.

Heavy industry is closely associated with the major coal mines and railways. Through the foothills of Yen Shan and T'ai-hang Shan, from the Gulf of Chihli to Han-tan and Chiao-tso (in Honan), stretches a great crescent with rich coal deposits. The major coal mines of K'ai-luan, Ching-hsi (in the Western Hills of Peking), Ching-hsin (northwest of Shih-chia-chuang), and Han-tan supply fuel for the electric-power plants of Peking, Tientsin, T'ang-shan, Ch'ing-huan-tao, Pao-ting, Shih-chia-chuang, and Han-tan.

Tientsin, the second-largest city, was the capital of Hopeh Province until February 1967, when it was placed under the direct control of the central government. It is the primary industrial and commercial centre of North China and the second most important trade centre in all China. Tientsin was only a small trading centre in the 17th century, but since 1860 its development has been rapid.

Tientsin is also a major educational centre, with six institutions of higher learning at the start of the present decade. There was a Chinese museum, a museum of science and technology, a sports stadium, a large public library, a theatre, an aquatic park, and numerous public amusement centres (see also TIENTSIN).

T'ang-shan (812,000 inhabitants in 1958), the third city in the region, is a centre of heavy industry. Its development has been associated with that of the K'ai-luan coal mines, initiated by British capital and further developed under the Communists to become one of the largest coal mines in China. T'ang-shan's power plant supplies electricity to the K'ai-luan mines and to Peking and Tientsin. Heavy industry includes steel, cement, and railway-repair shops. Light industry in the city includes textiles, paper, flour mills, and ceramics. The larger factories often have their own workers' clubs, libraries, and clinics. There are literally hundreds of factory-run spare-time workers' schools. The city also has the famous T'ang-shan Institute of Railway Engineering.

Pao-ting and Shih-chia-chuang

Other important cities. Pao-ting (250,000 inhabitants in 1958), in central Hopeh, is an ancient city. During the Liao and Chin dynasties, from the 10th to the 13th centuries, it formed part of Peking's outer defense perimeter. It was the provincial capital from 1912 to 1958 and has long been a cultural and educational centre. Its most famous monument is the Ancient Lotus Pond, around which the Lotus Pond College was built during the reign of Yung-chen (1723–35); the college later became the Hopeh Library. The city has a large cultural museum and a museum of natural history, a medical college, an agricultural college, and more than ten secondary schools. Pao-ting functions as a commercial hub for the central Hopeh subregion and as an administrative and residential city. A period of industrial development began in the 1950s with the establishment of plants pro-

ducing agricultural machinery, flour mills, vegetable oils, leather goods, woolens, and bricks and tiles.

Shih-chia-chuang (623,000 inhabitants in 1958) is the present provincial capital. Its growth followed the completion of the Peking–Hankow railway and the Shih-chia-chuang–T'ai-yü'an line, which have a junction there. By the 1930s, there were a few small-scale industrial plants, a railway-repair shop, a coke oven, small textile mills, and flour mills. Under Japanese occupation (1937–45) a new rail line was built to Shantung Province, which joined the Tientsin–P'u-k'ou railway, the trunk line of eastern China. The population increased to 170,000. Despite this, industrial output declined during the occupation. After 1949 the old plants were reconstructed and expanded, and large, new plants were built, including cotton-textile, woolen, dyeing, food-processing, brick and tile, and fertilizer plants. Between 1953 and 1958 its population grew from 373,000 to 623,000, making Shih-chia-chuang fourth in population and industrial output in Hopeh, after Peking, Tientsin, and T'ang-shan.

Han-tan and Chang-chia-k'ou

The fastest growing urban–industrial centre in the province is the ancient city and new cotton-textile centre of Han-tan, whose population grew from 90,000 in 1953 to 380,000 in 1958. Located 100 miles south of Shih-chia-chuang, on the Peking–Canton railway and the Hai Ho inland navigation system, it serves as a central market for the economic subregion of south Hopeh. It became a rising centre of light industry during the 1950s with the building of cotton-ginning and textile plants and factories for food processing and agricultural machinery. The city is in the midst of a cotton-producing region, linked by railway with the coal-mining complex of Feng-feng, in the T'ai-hang Shan—a mining city of over 100,000 inhabitants. Han-tan was the capital of the state of Chao in the 4th century BC. The area was a Chinese Communist base during the Sino-Japanese War (1937–45) and the civil war that followed. A huge Martyrs' Monument in Han-tan commemorates this episode of its long history.

The former market town of Chang-chia-k'ou (Kalgan) on the Peking–Pao-t'ou railroad in northwestern Hopeh has become a rising industrial–commercial city. In 1953–58 its population grew from 229,300 to 480,000. It is a food-processing centre and also makes mining machinery. Chang-chia-k'ou is 19 miles by rail from Hsüan-hua and the Lung-yen iron mines, the largest south of the Great Wall. Another 12 miles farther south are the coal-mining districts of Hsia-hua-yüan, which supply fuel and electric power to the area. The power is carried over high-voltage lines to the Peking–Tientsin–T'ang-shan electric grid. Chang-chia-k'ou has been a centre for trade between the Chinese and the steppe peoples of Mongolia since time immemorial.

A basic policy of the government since 1957–58 has been the simultaneous development of large-scale industry by the national authorities and, at the local level, of small to medium plants. This is called the policy of "walking on two legs." After a period of economic difficulties in the early 1960s, a new upsurge of local industrial development began in 1966. The emphasis was upon industries supporting agriculture.

The province is served by seven major railways: the Peking–Canton and the Tientsin–P'u-k'ou lines, running north and south; the Peking–Shen-yang (Mukden) line, to the northeast and the sea; the Peking–Pao-t'ou line, to the Inner Mongolian Autonomous Region, northern Shansi, and the Northwest; the Peking–Ch'eng-te line, to Ch'eng-te and the western part of the Northeast; the Shih-chia-chuang–T'ai-yüan line, to Shansi Province; and the Shih-chia-chuang–Te-chou line, to Shantung.

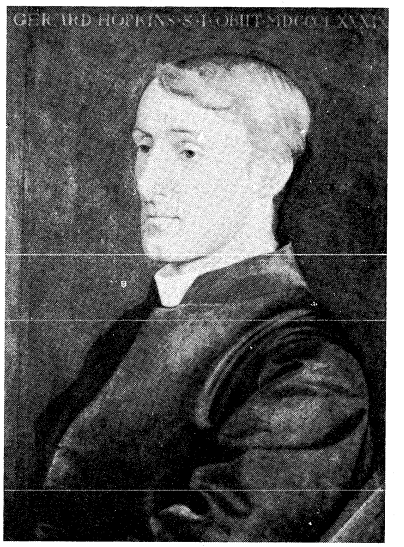
Sea transport moves through Tientsin and the coal port of Ch'in-huang-tao. The province exports coal and salt to Shanghai; products of light industry to Dairen and the Northeast; and coal, metals, and consumer goods to the Shantung ports. From Dairen come petroleum, soybean cakes (fertilizer), paper, chemicals, and iron and steel; from Shanghai, staple foods and consumer goods; and, from Shantung, fruit and native products. Overseas exports consist chiefly of agricultural products in exchange for metals, machinery, and industrial raw materials.

BIBLIOGRAPHY. The best English regional economic monograph is SUN CHING-CHIH (ed.), *Excerpts from Economic Geography of North China* (Eng. trans. 1958). The best physical geography of China in a Western language is V.T. ZAYCHIKOV (ed.), *The Physical Geography of China* (1965; orig. pub. in Russian, 1964). A. HERRMANN, *An Historical Atlas of China*, new ed. (1966), is indispensable, but contains some mistakes. For economic data to 1958, see CHEN NAIRUENN, *Chinese Economic Statistics: A Handbook for Mainland China* (1967). Useful for comparison is JOHN LOSSING BUCK, *Land Utilization in China*, 3 vol. (1937, reprinted in 1 vol., 1964). See also YUAN-LI WU, *The Spatial Economy of Communist China* (1967).

(F.Hu.)

Hopkins, Gerard Manley

One of the most individual of all Victorian writers, Gerard Manley Hopkins created poems that combine subtlety of perception, force of intellect, and intensity of religious feeling, reflecting his vocation as a Jesuit priest. Rich in complex experiments in prosody and diction, his style was so highly original that it was unappreciated, except by a few friends, in his own lifetime. Not until nearly 30 years after his death was the first volume of his poems published, in 1918, and not until the 1930s was he generally recognized as one of the great poets of the English language.



Gerard Manley Hopkins, portrait by Harry Ellis Wooldridge (1845–1917). In a private collection.

Early education and conversion

Gerard Manley Hopkins was born at Stratford, Essex, on July 28, 1844. He was the eldest of the nine children of Manley Hopkins, an Anglican, who had been British consul general in Hawaii and had himself published verse. Hopkins won the poetry prize at the Highgate grammar school and in 1863 was awarded a grant to study at Balliol College, Oxford, where he continued writing poetry while studying classics. In 1866, in the prevailing atmosphere of the Oxford Movement, which renewed interest in the relationships between Anglicanism and Roman Catholicism, he was received into the Roman Catholic Church by John Henry (later Cardinal) Newman. The following year, he left Oxford with such a distinguished academic record that Benjamin Jowett, then a Balliol lecturer and later master of the college, called him "the star of Balliol." Hopkins decided to become a priest. He entered the Jesuit novitiate in 1868 and burned his youthful verses, determining "to write no more, as not belonging to my profession."

Until 1875, however, he kept a journal recording his vivid responses to nature as well as his expression of a philosophy for which he later found support in Duns Scotus, the medieval Franciscan thinker. Hopkins' philos-

ophy emphasized the individuality of every natural thing, which he called "inscape." To Hopkins, each sensuous impression had its own elusive "selfness"; each scene was to him a "sweet especial scene."

In 1874 Hopkins went to St. Beuno's College in North Wales to study theology. Here he learned Welsh and, under the impact of the language itself as well as that of the poetry and encouraged by his superior, he began to write poetry again. Moved by the death of five Franciscan nuns in a shipwreck in 1875, he broke his seven-year silence to write the long poem "The Wreck of the Deutschland," in which he succeeded in realizing "the echo of a new rhythm" that had long been haunting his ear. It was rejected, however, by the Jesuit magazine *The Month*. He also wrote a series of sonnets strikingly original in their richness of language and use of rhythm, including the remarkable "The Windhover," in recent years one of the most frequently analyzed poems in the language. He continued to write poetry, but it was read only in manuscript by his friends and fellow poets, Robert Bridges (later poet laureate), Coventry Patmore, and the Rev. Richard Watson Dixon. Their appreciation of the strangeness of the poems (for the times) was imperfect, but they were, nevertheless, encouraging.

Ordination to the priesthood

Ordained to the priesthood in 1877, Hopkins served as missioner, occasional preacher, and parish priest in various Jesuit churches and institutions in London, Oxford, Liverpool, and Glasgow and taught classics at Stonyhurst College, Lancashire. He was appointed professor of Greek literature at University College, Dublin, in 1884. But Hopkins was not happy in Ireland; he found the environment uncongenial, and he was overworked and in poor health. From 1885 he wrote another series of sonnets, beginning with "Carrion Comfort." They show a sense of desolation produced partly by a sense of spiritual aridity and partly by a feeling of artistic frustration. These poems, known as the "terrible sonnets," reveal strong tensions between his delight in the sensuous world and his urge to express it and his equally powerful sense of religious vocation.

While in Dublin, Hopkins developed another of his talents, musical composition; the little he composed shows the same daring originality as does his poetry. His skill in drawing, too, allowed him to illustrate his journal with meticulously observed details of flowers, trees, and waves.

His friends continually urged him to publish his poems, but Hopkins resisted; all that he saw in print in his lifetime were some immature verses and original Latin poems, in which he took particular pleasure.

Hopkins died of typhoid fever on June 8, 1889, and was buried in the Glasnevin Cemetery, Dublin. Among his unfinished works was a commentary on the *Spiritual Exercises* of St. Ignatius Loyola, founder of the Jesuit order.

After Hopkins' death, Robert Bridges began to publish a few of the Jesuit's most mature poems in anthologies, hoping to prepare the way for wider acceptance of his style. By 1918, Bridges, now poet laureate, judged the time opportune for the first collected edition. It appeared but sold slowly. Not until 1930 was a second edition issued, and thereafter Hopkins' work was recognized as among the most original, powerful, and influential literary accomplishments of his century; it had a marked influence on such leading 20th-century poets as T.S. Eliot, Dylan Thomas, W.H. Auden, Stephen Spender, and C. Day Lewis.

Poetic style

Hopkins sought a stronger "rhetoric of verse." His exploitation of the verbal subtleties and music of English, of the use of echo, alliteration, and repetition, and a highly compressed syntax were all in the interest of projecting deep personal experiences, including his sense of God's mystery, grandeur, and mercy, and his joy in "all things counter, original, spare, strange," as he wrote in "Pied Beauty." He called the energizing prosodic element of his verse "sprung rhythm," in which each foot may consist of one stressed syllable and any number of unstressed syllables, instead of the regular number of syllables used in traditional rhythm. The result is a muscular verse, flexible, intense, vibrant, and organic, that combines accuracy of observation, imaginative daring, deep feeling, and intellectual depth.

Hopkins' letters reveal a brilliant critical faculty, scrupulous self-criticism, generous humanity, and a strong will. His friends paid tribute to his personal integrity and to his rare "chastity of mind." Coventry Patmore wrote of him: "There was something in all his words and manners which were at once a rebuke and an attraction to all who could only aspire to be like him."

BIBLIOGRAPHY. RUTH SEELHAMMER, *Hopkins Collected at Gonzaga* (1970), a bibliography of Hopkins' writings.

Editions: Poems of Gerard Manley Hopkins, ed. by ROBERT BRIDGES (1918); 2nd ed. rev. by CHARLES WILLIAMS (1930); 3rd ed. edited by W.H. GARDNER (1948; rev. 1956); 4th ed. edited by W.H. GARDNER and N.H. MACKENZIE (1967, reprinted 1970).

Letters and papers: C.C. ABBOTT (ed.), *Letters of Gerard Manley Hopkins to Robert Bridges; The Correspondence of Gerard Manley Hopkins and Richard Watson Dixon*, 2 vol. (1935); *Further Letters, Including His Correspondence with Coventry Patmore*, 2nd ed. rev. (1956); H. HOUSE and G. STOREY (eds.), *Journals and Papers* (1959); C. DEVLIN (ed.), *Sermons and Devotional Writings* (1959).

Biographies: W.H. GARDNER, *Gerard Manley Hopkins*, rev. ed., 2 vol. (1948–49), the definitive critical biography; G.F. LAHEY, *Gerard Manley Hopkins* (1930); JOHN PICK, *Gerard Manley Hopkins: Priest and Poet* (1942); ALFRED THOMAS, *Hopkins the Jesuit* (1969); J.G. RITZ, *Robert Bridges and Gerard Manley Hopkins* (1960).

Critical studies: W.A.M. PETERS, *Gerard Manley Hopkins: A Critical Essay Towards the Understanding of His Poetry* (1948); ALAN HEUSER, *The Shaping Vision of Gerard Manley Hopkins* (1958); D.A. DOWNES, *Gerard Manley Hopkins: A Study of His Ignatian Spirit* (1959); R.R. BOYLE, *Metaphor in Hopkins* (1961); W.S. JOHNSON, *Gerard Manley Hopkins: The Poet As Victorian* (1968); D. MCCHESNEY, *A Hopkins Commentary* (1968).

Concordance: R.J. DILLIGAN and T.K. BENDER, *A Concordance to the English Poetry of Gerard Manley Hopkins* (1970).

(J.C.Rd.)

Horace

An outstanding Latin poet, Horace (Quintus Horatius Flaccus) flourished in Rome's Augustan age, an era named after the emperor Augustus (sole ruler 31 BC–AD 14), whose friend and supporter the poet was. Although he wrote other works, his enormous influence upon Western literature rests principally upon his *Odes*, four books of short poems in lyric metres dealing with many themes, and his *Epistles*—hexameter poems (with six metrical feet) in the artificial form of letters, principally concerning conduct, although the two longer poems of the second book and the *Ars poetica* concentrate on literature and poetic criticism. Horace's fame rests chiefly upon the likable person revealed in his works. Autobiographical touches proliferate in his poetry: he seems to tell far more about himself, his character, his development, and his way of life than any other great poet of antiquity, and this, combined with the fact that he lived in one of the best documented periods of Roman history, creates the illusion that a good deal is known about him. Yet his autobiographical "disclosures" are not realistic but literary, a feature of his technique for creating an intimate atmosphere. What he tells about himself is part of the work of art he is creating; even his beliefs are impossible to pin down with any certainty, since apparent declarations of them are often no more than devices to create a mood or theme. Horace claimed that criticism in his poetry was for the good of society. As time went on he became convinced that the good poet must first be a good man and useful to the community as educator and civilizer.

Literary autobiographical "disclosures"

Family background and career. Born at Venusia (the modern Venosa) in Apulia, in southeast Italy, in December 65 BC, Horace was probably of the Sabellian hillman stock of Italy's central highlands. His father had once been a slave but gained freedom before Horace's birth and became an auctioneer's assistant. He also owned a small property and could afford to take his son to Rome



Horace, bronze medal, 4th century. In the Bibliothèque Nationale, Paris.
By courtesy of the Bibliotheque Nationale, Paris

and ensure personally his getting the best available education in the school of a famous fellow Sabellian named Orbilius (a believer, according to Horace, in corporal punishment). In about 46 BC Horace went to Athens, attending lectures at the Academy. After Julius Caesar's murder in March 44 BC, the eastern empire, including Athens, came temporarily into the possession of his assassins Brutus and Cassius, who could scarcely avoid clashing with Caesar's partisans, Mark Antony and Octavian (later Augustus), the young great-nephew whom Caesar, in his will, had appointed as his personal heir. Horace joined Brutus' army and was made *tribunus militum*, an exceptional honour for a freedman's son.

Horace's survival and rise

In November 42 BC, at the two battles of Philippi against Antony and Octavian, Horace and his fellow tribunes (in the unusual absence of a more senior officer) commanded one of Brutus' and Cassius' legions. After their total defeat and death, he fled back to Italy—controlled by Octavian—but his father's farm at Venusia had been confiscated to provide land for veterans. Horace, however, proceeded to Rome, obtaining, either before or after a general amnesty of 39 BC, the minor but quite important post of one of the 36 clerks of the treasury (*scribae quaestorii*). Early in 38 BC he was introduced to Gaius Maecenas, a man of letters from Etruria in central Italy who was one of Octavian's principal political advisers. He now enrolled Horace in the circle of writers with whom he was friendly. Before long, through Maecenas, Horace also came to Octavian's notice.

During these years Horace was working on Book I of the *Satires*, ten poems written in hexameter verse and published in 35 BC. The *Satires* drew on Greek roots, stating Horace's rejection of public life firmly and aiming at wisdom through serenity. He discusses ethical questions: the race for wealth and position, the folly of extremes, the desirability of mutual forbearance, and the evils of ambition. His 17 *Epodes* were also under way. Mockery here is almost fierce, the metre being that traditionally used for personal attacks and ridicule, though Horace attacks social abuses, not individuals. The tone reflects his anxious mood after Philippi.

In the mid-30s he received from Maecenas, as a gift or on lease, a comfortable house and farm in the Sabine hills (identified with considerable probability as one near Licenza, 22 miles [35 kilometres] northeast of Rome), which gave him great pleasure throughout his life. After Octavian had defeated Antony and Cleopatra at Actium, off northwest Greece (31 BC), Horace published his *Epodes* and a second book of eight *Satires* in 30–29 BC. Then, while the victor, styled Augustus in 27 BC, settled down, Horace turned, in the most active period of his poetical life, to the *Odes*, of which he published three books, comprising 88 short poems, in 23 BC. Horace, in the *Odes*, represented himself as heir to earlier Greek lyric poets but displayed a sensitive, economical mastery of words all his own. He sings of love, wine, nature (almost romantically), of friends, of moderation; in short, his favourite topics. Some of the *Odes* are about Maecenas or Augustus: although he praises the ancient Roman virtues the latter was trying to reintroduce, he remains his own master and never confines an ode to a single subject or mood. At some stage Augustus offered Horace the post of his private secretary, but the poet declined on the plea of ill health. Notwithstanding, Augustus did not resent his refusal, and indeed their relationship became closer.

Themes of the *Odes*

The last ode of the first three books suggests that Horace did not propose to write any more such poems. (He was possibly disappointed with their reception following publication in 23 BC.) The last of his epistles (in Book II, published 20–19 BC)—literary "letters" that were more mature and profound versions of the *Satires*—certainly announces an abandonment of "frivolous" lyric poetry for this more moralistic kind of verse. Very shortly afterward, he set to work on three further epistles (much longer than any in the first book), all relating in different ways to poetic activities. In these, Horace abandoned all satirical elements for a sensible, gently ironical stance, though the truisms praising moderation are never dull in his hands. Two epistles comprise a second book, and the third, the *Epistles to the Pisos*, was also known, at least subsequently, as the *Ars poetica*. These last three epistles embody literary criticism in a loose, conversational frame, the "Epistle to Florus" (Book II, Epistle 2) choosing to explain why Horace abandoned lyric poetry for "philosophy." The best poems, Horace thought, edify as well as delight; the secret of good writing is wisdom (implying goodness); the poet needs teaching and training to give of his best. The "Epistle to Florus" may have been written in 19 BC, the *Ars poetica* (consisting of nearly 30 maxims for young poets' guidance) in *c.* 19–18 BC, and the last epistle of Book I in 17–15 BC. This last named is dedicated to Augustus, from whom there survives a letter to Horace in which the Emperor complains of not having received such a dedication hitherto. In the last epistle contemporary poetry is asserted against Rome's earlier literary background, but this was doubtless a defense of Horatian methods.

Poet laureate

By this time Horace was virtually in the position of poet laureate, and in 17 BC he composed the *Secular Hymn* (*Carmen saeculare*) for ancient ceremonies called the Secular Games, which Augustus had revived to provide a solemn, religious sanction for the regime and, in particular, for his moral reforms of the previous year. The hymn was written in a lyric metre, Horace having resumed his compositions in this form; he next completed a fourth book of 15 *Odes*, mainly of a more serious (and political) character than their predecessors. The latest of these poems belongs to 13 BC. In 8 BC Maecenas, who had been less in Augustus' counsels during recent years, died. One of his last requests to the Emperor was: "Remember Horace as you would remember me." A month or two later, however, Horace himself died, after naming Augustus as his heir. He was buried on the Esquiline Hill near Maecenas' grave.

During the latter part of his life, Horace had been accustomed to spend the spring and other short periods in Rome, where he appears to have possessed a house. He would winter sometimes by the southern sea and spent much of the summer and autumn at his Sabine farm or sometimes at Tibur (Tivoli) or Praeneste (Palestrina), both a little east of Rome. A short "Life of Horace," of which the substance apparently goes back to Suetonius, a biographer of the 2nd century AD, quotes a jocular letter he received from Augustus, from which it emerges that the poet was short and fat. He himself confirms his short stature and, describing himself at the age of about 44, states that he was gray before his time, fond of sunshine, and irritable but quickly appeased.

Influences, personality, and impact. To a modern reader, the greatest problem in Horace is posed by his continual echoes of Latin and, more especially, Greek forerunners. The echoes are never slavish or imitative and are very far from precluding originality. For example, in one of his satires Horace wrote what looks at first like a

realistic account of a journey made to Brundisium (Brindisi, on Italy's "heel") in 37 BC. Two of the incidents, however, prove to have been lifted—and cleverly adapted—from a journey by the earlier Latin satirist Lucilius. Often, however, Horace provides echoes that cannot be identified since the works he was echoing have disappeared, though they were recognized by his readers.

Another disconcerting element is provided by Horace's own references to his alleged models. Very often he names as a model some Greek writer of the antique, preclassical, or classical past (8th–5th centuries BC), whom he claims to have adapted to Latin, notably, Alcaeus and Archilochus and Pindar. Yet his style of writing is much nearer to that of the more "modern," refined, and scholarly Greek writers of the Hellenistic, Alexandrian period (3rd and 2nd centuries BC), though to these (as to certain important Latin predecessors) his acknowledgments are selective and inadequate.

If this continuous relationship with the literary tradition is borne in mind, together with certain other factors that preclude wholly direct expression, such as the political autocracy of the time and Horace's own detached and even evasive personality, then it does become possible, after all, to deduce from his poetry certain conclusions about his views, if not about his life. The man who emerges is kindly, tolerant, and mild but capable of strength; consistently humane, realistic, astringent, and detached, he is a gentle but persistent mocker of himself quite as much as of others.

Character and attitudes

His attitude to love, on the whole, is flippant; without telling the reader a single thing about his own amorous life, he likes to picture himself in ridiculous situations within the framework of the appropriate literary tradition—and relating, it should be added, to women of Greek names and easy virtue, not Roman matrons or virgins. To his male friends, however—the men to whom his *Odes* are addressed—he is affectionate and loyal, and such friends were perhaps the principal mainstay of his life. The gods are often on his lips, but, in defiance of much contemporary feeling, he absolutely denied an afterlife. So "gather ye rosebuds while ye may" is an ever-recurrent theme, though Horace insists on a Golden Mean of moderation—deploring excess and always refusing, deprecating, dissuading.

Some of his modern admirers see him as the poet of the lighter side of life; others see him as the poet of Rome and Augustus. Both are equally right, for this balance and diversity were the very essence of his poetical nature. But the second of these roles is, for modern readers, a harder and less palatable conception, since the idea of poetry serving the state is not popular in the West—and still less serving an autocratic regime, which is what Horace does. Yet he does it with a firm, though tactful, assertion of his essential independence. Not only is he unwilling to become Augustus' secretary, but, pleading personal inadequacy, he also gracefully sidesteps various official, grandiose, poetic tasks, such as the celebration of the victories of Augustus' admiral Agrippa. And he refers openly to his own juvenile military service against the future Augustus, under Brutus at Philippi. He himself ran away—he characteristically says—and threw away his shield. But that, equally characteristically, turns out to be copied from a Greek poet, indeed from more than one. It is not autobiography; it is a traditional expression of the unsuitability of poets—and of himself—for war. The whole poem absolves Horace of any possible charge of failing, because of his current Augustan connections, to maintain loyalty to his republican friends.

Attitude toward the Augustan regime

Horace's intellectual formation had to a large extent been completed before the Augustan regime began; yet he came to admire Augustus sincerely and deeply, owing him many practical benefits. But, above all, he deeply admired him for ending a prolonged, nightmare epoch of civil wars. So great was that achievement that Horace, at least, had no eye for any crudities the new imperial regime might possess. This was one of the ages when people wanted order more than liberty, though Augustus was an adept at investing his new order with a sufficient respect for personal freedom and a sufficient facade of republican institutions to set most men's minds at rest. He also restored the temples, and to Horace, though he probably did not believe in the gods whose names he called upon, the religious traditions and rituals of the Roman state seemed an integral, venerable part of Rome's greatness. The Emperor was on more delicate ground when he sought, by social legislation, to purify personal morals and to protect and revive the Roman family. But here, too, Horace, in spite of his own erotic frivolity, was with him, perhaps because of the famous austerity of his Sabine stock. And so the *Secular Hymn* contains a specific allusion (poetically not altogether successful) to these reforms.

Yet, before the hymn, Horace had already written the magnificent *Roman Odes,* numbers one to six of Book III—a great tribute to Augustus' principate, perhaps the greatest political poetry that has ever been written. But these *Odes* are by no means wholly political, for much other material, including abundant Greek and Roman mythology, is woven into their dense, compact, resplendent texture. This cryptic, riddling sonority is the work of a poet who saw himself as a solemn bard (*vates*), a Roman reincarnation of Pindar of Thebes (518–438 BC), a stately Greek lyricist. Pindar increasingly becomes Horace's model in the further state odes of his fourth and last book.

Horace's influence on posterity

After Horace's *Secular Hymn*, his works were known and appreciated by all educated Romans. Already at the time of Horace's death, his *Odes* were suffering the fate he deprecated for them and had become a school textbook. But their excellence was so great that they had few ancient lyrical successors, until some early Christian writers—Ambrose, Prudentius, and Paulinus—occasionally echoed Horace's forms, though with a difference in spirit. Thereafter, the medieval epoch had little use for the *Odes*, which did not appeal to its piety, although his *Satires* and *Epistles* were read because of their predominantly moralistic tones. The *Odes* came into their own again with the Renaissance and, along with the *Ars poetica*, exerted much influence on Western poetry through the 19th century. The English Victorian poet Alfred, Lord Tennyson, hailed the lines of the *Odes* as:

> Jewels five-words-long
> That on the stretch'd forefinger of all Time
> Sparkle for ever.

The many-facetted intricacy of these "jewels" has challenged translators throughout the centuries; in spite, or because, of their not wholly conquerable problems, every ode has been translated hundreds—perhaps thousands—of times. And still new versions, some of them admirable, continue to appear.

MAJOR WORKS

Among Horace's earlier poems, the following may be found particularly interesting: *Satires*, Book 1, fifth, sixth, and ninth satires; Book 2, sixth, seventh, and eighth satires; *Epodes*, ninth and 16th. Most of the *Odes* and *Epistles* are important. For a selection of translations of various epochs, see Michael Grant (ed.), *Roman Readings*, pp. 180–221 (1958, reprinted 1967); and for the *Odes, The Odes of Horace,* trans. by James Michie (1963).

BIBLIOGRAPHY. EDWARD FRAENKEL, *Horace* (1957, paperback 1966), is a fundamental work. GORDON W. WILLIAMS, *Tradition and Originality in Roman Poetry* (1968), and his shorter book *The Nature of Roman Poetry* (1970), deal extensively with Horace in an illuminating fashion. Other recent general studies are KENNETH J. RECKFORD, *Horace* (1969); and ANTONIO LA PENNA, *Orazio e la morale mondana europea* (1968). DAVID A. WEST, *Reading Horace* (1967), analyzes individual poems.

Odes: L.P. WILKINSON, *Horace and His Lyric Poetry*, 2nd ed. rev. (1968); GIORGIO PASQUALI, *Orazio lirico* (1920); R.G.M. NISBET and MARGARET HUBBARD, *A Commentary on Horace: Odes, Book I* (1970); GORDON W. WILLIAMS (ed.), *The Third Book of Horace's Odes* (1969); N.E. COLLINGE, *The Structure of Horace's Odes* (1961).

Satires, Epistles, Ars poetica: C.O. BRINK, *Horace on Poetry*, vol. 1, *Prolegomena to the Literary Epistles*, and vol. 2, *The Ars poetica* (1963–71); NIALL RUDD, *The Satires of Horace* (1966); M.J. MCGANN, *Studies in Horace's First Book of Epistles* (1969).

(M.Gr.)

Hormone

Hormones are organic substances that are secreted by plants and animals and that function in the regulation of physiological activities and in the maintenance of a constant internal environment (homeostasis). They carry out their functions by evoking responses from specific organs or tissues that are adapted to react to minute quantities of them. The classical view of hormones is that they are transmitted to their targets in the bloodstream after discharge from the glands that secrete them. This mode of discharge (directly into the bloodstream) is called endocrine secretion. The meaning of the term hormone has been extended beyond the original definition of a blood-borne secretion, however, to include similar regulatory substances that are distributed by diffusion across cell membranes instead of by a blood system.

The hormones of vertebrate and invertebrate animals and of plants are characterized in this article. Comparative discussions of the hormone-producing (endocrine) glands of animals are found in ENDOCRINE SYSTEMS. The article ENDOCRINE SYSTEM, HUMAN deals with the endocrine system of man. The consequences of abnormal hormonal function in man are dealt with in ENDOCRINE SYSTEM DISEASES AND DISORDERS.

This article is divided into the following sections:

I. General features of hormones
- Relationships between endocrine and neural regulation
- The evolution of hormones

II. The hormones of vertebrates
- Hormones of the pituitary gland
- Hormones of the thyroid gland
- Parathormone of the parathyroid gland
- Hormones of the pancreas
- Hormones of the adrenal glands
- Hormones of the reproductive system
- Hormones of the digestive system
- Endocrine-like glands and secretions

III. The hormones of invertebrates
- Hormones of insects
- Hormones of crustaceans
- Other invertebrate hormones

IV. The hormones of plants
- Growth promotors
- Growth inhibitors

I. General features of hormones

RELATIONSHIPS BETWEEN ENDOCRINE AND NEURAL REGULATION

Hormonal regulation is closely related to that exerted by the nervous system, and the two processes have generally been distinguished by the rate at which each causes effects, the duration of these effects, and their extent; *i.e.*, the effects of endocrine regulation may be slow to develop but prolonged in influence and widely distributed through the body, whereas nervous regulation is typically concerned with quick responses that are of brief duration and localized in their effects. Advances in knowledge, however, have modified these distinctions.

Neurohormones

Nerve cells are secretory, for responses to the nerve impulses that they propagate depend upon the production of chemical transmitter substances, or neurohumors, such as acetylcholine and noradrenaline (norepinephrine), which are liberated at nerve endings in minute amounts and have only a momentary action. It now has been established, however, that certain specialized nerve cells, called neurosecretory cells, can translate neural signals into chemical stimuli by producing secretions called neurohormones. These secretions, which are often polypeptides (compounds similar to proteins but composed of fewer amino acids), pass along nerve-cell extensions, or axons, and are typically released into the bloodstream at special regions called neurohemal organs, where the axon endings are in close contact with blood capillaries (Figure 1A). Once released in this way, neurohormones function in principle similar to hormones that are transmitted in the bloodstream and are synthesized in the endocrine glands.

The distinctions between neural and endocrine regulation, no longer as clear-cut as they once seemed to be, are further weakened by the fact that neurosecretory nerve endings are sometimes so close to their target cells that vascular transmission is not necessary (Figure 1B). There is good evidence that hormonal regulation occurs by diffusion in plants and (although here the evidence is largely indirect) in lower animals (*e.g.*, coelenterates), which lack a vascular system.

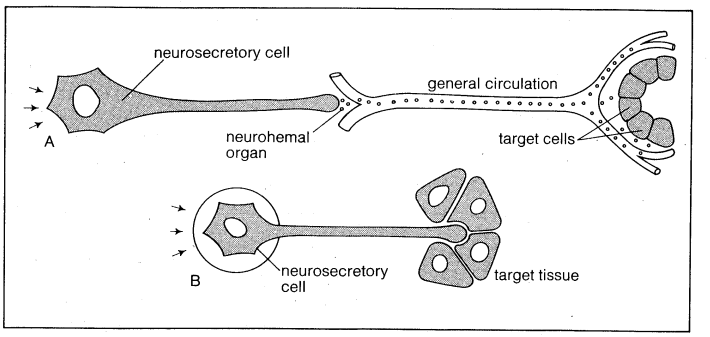


Figure 1: The release of neurohormones from neurosecretory nerve cells.
(A) Nerve signals (arrows) are translated into neurohormones, which enter the circulation at the neurohemal organ, reaching target cells via the bloodstream. (B) Release of neurohormone close to target cells without intervention of bloodstream.

THE EVOLUTION OF HORMONES

Hormones have a long evolutionary history, knowledge of which is important if their properties and functions are to be understood. Many important features of the vertebrate endocrine system, for example, are present in the lampreys and hagfishes, modern representatives of the primitively jawless vertebrates (Agnatha), and these features were presumably present in fossil ancestors that lived more than 500,000,000 years ago. The evolution of the endocrine system in the more advanced vertebrates with jaws (Gnathostomata) has involved both the appearance of new hormones and the further evolution of some of those already present in agnathans; in addition, extensive specialization of target organs has occurred to permit new patterns of response.

Role in chemical regulation

The factors involved in the first appearance of the various hormones is largely a matter for conjecture, although hormones clearly are only one mechanism for chemical regulation, diverse forms of which are found in living things at all stages of development. Other mechanisms for chemical regulation include chemical substances (so-called organizer substances) that regulate early embryonic development and the pheromones that are released by social insects as sex attractants and regulators of the social organization. Perhaps, in some instances, chemical regulators including hormones appeared first as metabolic by-products. A few such substances are known in physiological regulation: carbon dioxide, for example, is involved in the regulation of the respiratory activity of which it is a product, in insects as well as in vertebrates. Substances such as carbon dioxide are called parahormones to distinguish them from true hormones, which are specialized secretions.

II. The hormones of vertebrates

HORMONES OF THE PITUITARY GLAND

The pituitary gland, or hypophysis (Figure 2), which dominates the vertebrate endocrine system, is formed of two distinct components. One is the neurohypophysis, which forms as a downgrowth of the floor of the brain and gives rise to the median eminence and the neural lobe; these structures are neurohemal organs. The other is the adenohypophysis, which develops as an upgrowth from the buccal cavity (mouth region) and usually includes two glandular portions, the pars distalis and the pars intermedia, which secrete a number of hormones. The hormones secreted by the adenohypophysis are protein or polypeptide in nature and vary in complexity; as a result, their chemical constitution has not always been as fully characterized as has that of structurally simpler molecules of some other endocrine secretions. Functional analysis of these hormones also is difficult, for the

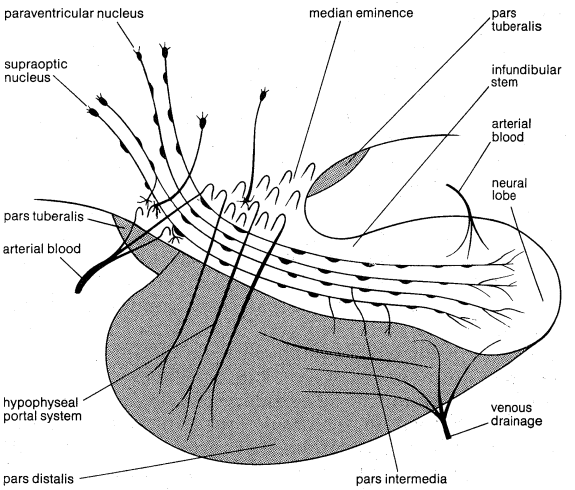


Figure 2: Elements in a generalized mammalian pituitary gland.

From E.J.W. Barrington, *Hormones and Evolution*, p. 29, fig. 4; The English Universities Press Ltd.

targets of certain hormones of the adenohypophysis, called tropic, or trophic, hormones, are other endocrine glands. The action of such tropic hormones can be understood only in the light of the mode of function of the endocrine glands they regulate.

Adenohypophysis. *Growth hormone (somatotropin; STH).* Growth hormone is protein, the primary structure of which has been fully established for the human and bovine forms of the hormone. It is probably universally distributed in gnathostomes (vertebrates with jaws), in which it is essential for the maintenance of growth, but its presence in agnathans (jawless vertebrates) has not yet been established with certainty. The physical and chemical properties of growth hormone (Table 1), which differ from species to species, are associated with marked differences in biological activity. Only part of the molecule, however, is actually responsible for its biological activity, for up to 25 percent of it can be lost without causing any decline in potency. Man responds to growth hormones obtained from other primates, but the rat responds to those from a wide range of species. Even more striking, growth of teleost (bony) fishes, which stops if the pituitary gland is removed, can be restarted by treatment with mammalian growth hormone; on the other hand, preparations of pituitary glands from these fishes have no effect on the growth of mammals. The growth hormones of lungfishes, which are closely related to the terrestrial vertebrates, and of sturgeons, which are primitive members of the evolutiontary line that led to bony fishes, affect mammalian growth, perhaps because these hormones have a more generalized molecular structure.

Table 1: Some Physicochemical Properties of Growth Hormones of Mammals

	human	monkey	bovine	sheep	pig	whale
Molecular weight	21,500	23,000	45,000	47,800	41,600	39,900
Isoelectric point (pH)	4.9	5.5	6.8	6.8	6.3	6.2
Sedimentation coefficient	2.18	1.88	3.19	2.76	3.02	2.84
Diffusion coefficient	8.88	7.20	7.23	5.25	6.54	6.56
Disulfide linkages	2	4	4	5	3	3

Source: G.H. Li, *Perspectives in Biology and Medicine*, 11, 1968.

Effects of growth hormone

Growth is such a complex process that definition of the growth hormone's mode of action is difficult. One of its known effects is an increase in the rate of protein synthesis, which is to be expected, since growth involves the deposition of new protein material. In addition, growth hormone affects the metabolism of certain ions (including sodium, potassium, and calcium), promotes the release of fats from fat stores, and influences carbohydrate metabolism in ways that tend to cause an increase in the level of glucose in the bloodstream. The last action creates a demand for an increased output of insulin (a hormone secreted by the pancreas), which acts to return the blood-glucose level to normal. Prolonged treatment of dogs with growth hormone can overstrain the pancreatic tissue in which insulin is synthesized and bring about a diabetic condition, in which insulin is formed in inadequate quantities. It is unlikely, however, that this is a factor in establishing diabetes mellitus in man. Excess secretion of growth hormone does, however, have damaging effects in man, for it produces overgrowth of the skeleton. If this occurs in young people, before the closure of the epiphyses (ends) of the long bones, it results in gigantism. If it occurs afterward, it causes acromegaly, in which the disturbance of body form and functioning is more serious, with enlargement of the bones and soft tissues, and consequent distortion of the skull.

Prolactin. Prolactin is a protein hormone that in sheep has a molecular weight of about 23,000 (based on a molecular weight for hydrogen of one). It is doubtful that it is present in agnathans, but it is widely distributed in jawed vertebrates. In female mammals, prolactin initiates and maintains the secretion of milk, the mammary glands having been previously prepared for this function by the action of other hormones. In the female rat prolactin also maintains the secretion of the hormone progesterone, which is formed by the corpus luteum, an endocrine gland of the ovary; *i.e.*, prolactin, termed luteotropin in the rat, is a gonadotropin (see below *Hormones of the reproductive system*) in this animal, because its target is an endocrine gland. Evidence is accumulating that the molecular structures of prolactin and growth hormone are similar. This explains why they show some overlap in biological properties; in particular, administration of prolactin promotes some growth in many terrestrial vertebrates. Human growth hormone has prolactin-like luteotropic properties, and it is not yet certain that man actually has a distinct prolactin hormone.

Prolactin itself shows remarkable variety in biological action from one vertebrate group to another. It promotes the production of so-called crop-milk with which pigeons feed their young, and the associated changes in structure and arrangement of the wall of the crop provide a convenient means to assay the hormone. In certain newts (*Triturus* species) prolactin induces the change of behaviour that drives young animals into the water (water-drive action). In bony fishes, prolactin is concerned with the regulation of the level of sodium in blood plasma; it therefore is essential in some teleost species (*e.g.*, *Poecilia lattipinna*) for the maintenance of life in fresh water. Although other teleosts (*e.g.*, eels) can survive in fresh water after hypophysectomy, this means only that prolactin is but one factor in a complex regulatory mechanism involving several factors. Mammalian prolactin can regulate sodium metabolism when given to eels and can maintain the life of hypophysectomized *Poecilia*. Yet, although other convincing evidence suggests that the hormone must be present in the pituitaries of these teleosts, preparations of their glands tested on pigeons do not have a typical crop-stimulating action. This evidence is best accounted for by supposing that the prolactin molecule has undergone evolutionary changes in its molecular structure and biological properties and has also established specific adaptive relationships with target organs such as the crop and mammary glands.

Adrenocorticotropic hormone (ACTH; corticotropin). ACTH is present in all jawed vertebrates but has not yet been decisively demonstrated in agnathans. It regulates the activity of part of the outer region (cortex) of the adrenal glands (considered below under *Hormones of the adrenal glands*). In mammals its action on the adrenal cortex is limited to areas called the zona reticularis and zona fasciculata, in which important steroid hormones (*e.g.*, cortisol and corticosterone, known as glucocorticoids) are formed; ACTH does not affect the synthesis of the mineralocorticoid hormone aldosterone, which takes place chiefly in the outer cortical region (zona glomerulosa). Evidence strongly suggests that the action of ACTH is mediated by a substance known as CAMP (cyclic 3′,5′-adenosine monophosphate), the rate of synthesis of which

increases in adrenal tissue in the presence of ACTH; cAMP in turn promotes synthesis of enzymes necessary for the formation of cortisol and corticosterone. The relationship between ACTH and the adrenal cortex is an example of the negative feedback characteristic of endocrine systems; *i.e.*, a decrease in the level of glucocorticoids circulating in the bloodstream evokes an increase in the secretion of ACTH, which, by stimulating the secretory activity of its target gland (the adrenal cortex), tends to restore to normal the level of glucocorticoids in the bloodstream. The release of ACTH can also be influenced by the level of circulating adrenaline, which is not surprising in view of the close functional relationship between the hormones of the adrenal cortex and medulla.

Negative feedback

The ACTH of mammals is a polypeptide molecule consisting of 39 amino acids, only the first 20 of which are required for full activity. This region, often referred to as the active centre, is constant in composition in all mammals studied thus far; the remainder of the molecule varies slightly in amino-acid composition among different species. Since, however, the mammalian hormone is active in all vertebrates, ACTH structure probably varies little from one class to another. The concept that biological activity is localized in an active centre of a complex molecule is applicable to other polypeptide and protein hormones, including growth hormone, whose structure, as noted previously, can be partly lost without causing loss of activity. The concept of an active centre, however, raises the question of the function of the rest of the molecule. It may serve as the site of antigenic properties or of structural features important in establishing relations with specialized receptors in target cells.

Thyrotropin (thyroid-stimulating hormone; TSH). Thyrotropin regulates the thyroid gland through a feedback relationship similar to that for ACTH; thyrotropin increases the secretion of the hormones from the thyroid gland and, if its action is prolonged, evokes increase in cell number (hyperplasia) and increase in size of the gland. One consequence of an overactive thyroid in man is a bulging of the eyes (exophthalmos). The cause of this is obscure, although it has been thought to result from the action of a distinct exophthalmos-producing substance that, while closely associated with thyrotropin, can be chemically separated from it. Thyrotropin, which is probably absent from agnathans, is a glycoprotein; *i.e.*, a protein combined with carbohydrate. Its molecular weight is estimated to be about 26,000 to 30,000 in mammals. Some variability occurs in the degree of response obtained when a hormonal preparation from one species is tested on other species. This suggests, as with prolactin, that it has undergone molecular evolution.

Follicle-stimulating hormone (FSH). FSH is termed a gonadotropin because it is concerned with the regulation of the activity of the gonads, or sex organs, which are endocrine glands as well as the sources of eggs and sperm. FSH stimulates development of the graafian follicle, a small vesicle containing an egg, in the ovary of the female mammal; in the male, it promotes the development of the tubules of the testes and the differentiation of sperm. FSH, like thyrotropin, is a glycoprotein, with an estimated molecular weight (in man) of 41,000 to 43,000. The effects of FSH are discussed further in *Hormones of the Reproductive System*, below.

Luteinizing hormone (LH; interstitial-cell-stimulating hormone; ICSH). Luteinizing hormone is another gonadotropin, a glycoprotein with a molecular weight of 26,000 in man. In the female mammal it promotes the transformation, following release of the egg (ovulation), of the graafian follicle into the corpus luteum, an endocrine gland; its complex functional interrelationship with FSH is dealt with below in *Hormones of the reproductive system.* In the male, luteinizing hormone promotes the development of the interstitial tissue (Leydig cells) of the testes and hence promotes the secretion of the male sex hormone, testosterone. In males, luteinizing hormone (LH) may be associated with FSH in promoting the secretion of testosterone. The interrelationship of LH and FSH has made it difficult to establish with certainty that two separate hormones exist, particularly since both are glycoproteins. Although the existence of two hormones has been established in mammals, the situation in lower vertebrates is not yet certain. All vertebrates undoubtedly have gonadotropic activity in their pituitary glands; but, although FSH-like and LH-like effects are detectable, it is not yet clear that two distinct hormones always exist.

An unexpected property of mammalian FSH and LH is that both have a thyrotropic action (*i.e.*, stimulate secretion of thyroid hormones) in lower vertebrates. This so-called heterothyrotropic effect has led to the supposition that FSH, LH, and thyrotropin may have evolved by modification of a common ancestral glycoprotein molecule, resulting in an overlap of properties. Similar examples are pointed out in later sections.

Thyrotropic action of FSH and LH

Melanocyte-stimulating hormone (MSH; intermedin). This hormone, secreted by the pars intermedia region of the pituitary gland, regulates colour changes in animals by promoting the concentration of pigment granules in pigment-containing cells (melanocytes, chromatophores) in the skin of lower vertebrates; MSH acts in conjunction with the nervous system in bony fishes and reptiles. No response involving physiological colour change is found in birds and mammals, although the hormone is secreted by them, even in species in which a pars intermedia region is no longer distinguishable in the adenohypophysis. The reason for the presence of MSH in birds and mammals is not clear since the function of the hormone in these animals has not yet been established. MSH is known to influence the behaviour of mammals and the total amount of pigment in their skin, which darkens in man after administration of large doses of the hormone. This type of change, however, which results from a change in the total amount of pigment present, is called a morphological colour change, in contrast to the physiological one that occurs in the skin of lower vertebrates.

MSH is a polypeptide that exists in two forms. α-MSH contains 13 amino acids, which are found in the same sequence in all species studied thus far; β-MSH has 18 amino acids, in sequences that differ in different species. Remarkable are the facts that the 13 amino acids of α-MSH are identical with the first 13 amino acids of ACTH and that both α and β forms of MSH have a heptapeptide (seven-amino-acid) sequence that has some melanocyte-stimulating activity, and that is identical with an amino-acid sequence of ACTH (see Figure 3). This close correspondence in sequence can hardly be coincidental and suggests, as has been postulated above for FSH, LH,

Adapted from R.S. Harris, ed., *Vitamins and Hormones* (1961); Academic Press

hormone	amino acid sequence																								
ACTH:								Ser	Tyr	Ser	Met	Glu	His	Phe	Arg	Try	Gly	Lys	Pro	Val	Gly	Lys	Lys	Arg	Arg	Pro
(pig, sheep, beef)								1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
alpha-MSH:						CH_3CO		Ser	Tyr	Ser	Met	Glu	His	Phe	Arg	Try	Gly	Lys	Pro	Val	NH_2					
(pig, beef, horse)								1	2	3	4	5	6	7	8	9	10	11	12	13						
beta-MSH:						Asp	Glu	Gly	Pro	Tyr	Lys	Met	Glu	His	Phe	Arg	Try	Gly	Ser	Pro	Pro	Lys	Asp			
(pig)						1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18			
beta-MSH:						Asp	Ser	Gly	Pro	Tyr	Lys	Met	Glu	His	Phe	Arg	Try	Gly	Ser	Pro	Pro	Lys	Asp			
(beef)						1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18			
beta-MSH:						Asp	Glu	Gly	Pro	Tyr	Lys	Met	Glu	His	Phe	Arg	Try	Gly	Ser	Pro	Arg	Lys	Asp			
(horse)						1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18			
beta-MSH:	Ala	Glu	Lys	Lys	Asp	Glu	Gly	Pro	Tyr	Arg	Met	Glu	His	Phe	Arg	Try	Gly	Ser	Pro	Pro	Lys	Asp				
(human)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22				

Figure 3: Relationships between the structure of melanocyte-stimulating hormones (MSH) and the first 19 amino acids of adrenocorticotropin (ACTH).

and thyrotropin, that ACTH and α- and β-MSH may have differentiated within the adenohypophysis by evolutionary modification of a common ancestral molecule. A change in biological activity results from modifications in the amino-acid composition; β-MSH preparations from the pig and the horse, for example, are five times more effective than those of the ox in evoking pigment dispersion in frogs. MSH molecules do not show ACTH activity, which is dependent on the presence of amino acids that occur in the region of the molecule not found in MSH. On the other hand, ACTH does have a slight effect on pigment dispersion, presumably because its structure contains the heptapeptide sequence mentioned above.

Evidence shows that each of the adenohypophysial hormones is secreted by a specific cell type. The cell types can be differentiated by staining sections of the pituitary gland, and known changes in the output of an individual hormone, induced experimentally or correlated with phases in the life cycle, can be shown to correspond with changes in the appearance of the corresponding cell type.

Regulation of adenohypophysial activity

The regulation of the activity of the secretory cells of the adenohypophysis depends upon its association with the floor of the brain and results from the existence of a neurosecretory system located mainly, perhaps entirely, in the hypothalamic region there. Much remains to be learned about this system, which involves the passage into the adenohypophysis of neurosecretions from the hypothalamus called hypothalamic releasing factors. Chemical characterization of these factors shows them to be simple polypeptides, in which respect they resemble the hypothalamic polypeptide hormones (discussed in the next section). This neurosecretory system is best understood in mammals, in which good evidence has been found for the existence of a separate releasing factor for each hormone secreted by the pars distalis region of the adenohypophysis; a similar arrangement probably exists in other gnathostomes. The situation in agnathans is obscure, but the anatomical organization of the pituitary glands of these animals implies at least some form of chemical communication between the hypothalamus and the pituitary gland.

Chemical communication is achieved by two routes. One route is by the entry of neurosecretory-cell fibres from the hypothalamus into the adenohypophysis, so that the hypothalamic factors, when released, are either in immediate contact with the secretory cells (Figure 1B) or in blood capillaries very closely related to them. This route is characteristic of the pars intermedia region, in which neurosecretory fibres from the hypothalamus control the functioning of the secretory cells. If the pars intermedia is separated from its direct connection with the floor of the brain, for example, MSH secretion in amphibians increases, and prolonged darkening of the skin results. Secretory activity of the pars intermedia cannot then be regulated again until the nerve fibres have regenerated.

Direct innervation similar to that of the pars intermedia is also found in the pars distalis of bony fishes. Here neurosecretory fibres arise from a localized region of the hypothalamus, called the nucleus lateralis tuberis, and end in contact either with the various types of secretory cells or with blood capillaries related to them. The other route of chemical communication to the pars distalis is found in many fishes and in all terrestrial vertebrates; it is a vascular route that depends upon the median eminence, which lies at the front end of the neurohypophysis (see Figure 2). The median eminence is a neurohemal organ containing a capillary bed into which hypothalamic neurosecretory fibres discharge their releasing factors. These are then transmitted through blood vessels known as the hypophysial portal system, into the capillaries of the pars distalis, where each factor influences its specific target cells (compare Figure 1A).

Both hypothalamic neurosecretory routes have the same physiological significance; *i.e.*, they provide chemical communication between the adenohypophysis and the central nervous system, thus making it possible for the latter to regulate the activity of the gland (and also of the endocrine glands its tropic hormones influence) in response to the demands of both the internal and external environments. The hypothalamic neurosecretory system is also involved in the function of the negative-feedback mechanisms that regulate the secretion of the tropic hormones. As already mentioned for ACTH, the secretions of tropic hormones from the adenohypophysis are controlled by bloodstream levels of the hormones secreted by their target glands; the hormones of the target glands may act directly on specific adenohypophysial cells or indirectly by influencing the output of releasing factors from the hypothalamus.

Neurohypophysis and the polypeptide hormones of the hypothalamus. Another neurosecretory system, which involves the hypothalamic region of the brain and the neurohypophysis of the pituitary gland, originates in groups of neurosecretory cells in the hypothalamus called, in mammals, the nucleus supraopticus and the nucleus paraventricularis and, in lower vertebrates, the nucleus preopticus. Neurohormones from these regions pass along the axons of the neurosecretory cells to the neural lobe (see Figure 2) bound to a protein called neurophysin (molecular weight of 20,000 to 25,000). In the neural lobe, which is the neurohemal organ of this neurosecretory system, the hormones separate from neurophysin and are released into the bloodstream.

Effects of hypothalamic polypeptide hormones

In most mammals, the neurohormones are oxytocin and arginine vasopressin. Both have relatively simple and very similar molecular structures; each is composed of nine amino acids arranged as a ring, which is formed by the linkage of two molecules of the amino acid cysteine (a disulfide linkage $-S-S-$), and a short side chain Table 2). The two hormones differ in structure only at amino acids numbered 3 and 8. In some species of the family Suidae (pig, peccary, hippopotamus) arginine vasopressin is replaced by lysine vasopressin; in others, both may be present. The difference between the two vasopressin hormones is that one has the amino acid lysine (Lys) at position 8; the other has arginine (Arg). Both the vasopressins and oxytocin show some overlap of activity, which is a consequence of the similarities in their molecular structures. Preparations of the three hormones evoke responses from the mammalian kidney, from the epithelial-cell layer of the frog bladder, and from the smooth muscle in blood vessels, uterus, and milk glands. The slight variation in amino-acid composition, however, affects the levels of the responses; *i.e.*, the vasopressins differ slightly from each other in response, and oxytocin differs markedly from both. Each, therefore, is said to have a characteristic pharmacological spectrum, and all have some medical use.

The primary actions of oxytocin are the promotion of uterine contraction (of value in obstetrical medicine) and the release of milk during suckling. The stimulation exerted upon the nipples during suckling leads to the transmission of nerve impulses to the hypothalamus. These bring about the discharge of oxytocin, which causes contraction of the smooth muscle of the small ducts of the mammary glands and the release of milk. Although the vasopressins cause an increase in blood pressure in mammals through vasoconstriction (*i.e.*, contraction of blood vessels), this action requires a high concentration of hormone and is probably not a normal physiological effect. The primary action of the vasopressins is on the kidneys; it brings about a reduction in the output of urine. As a result arginine vasopressin is commonly called the antidiuretic hormone (ADH). A lack of this hormone in man results in a copious flow of urine, a condition called diabetes insipidus, which is readily alleviated by preparations containing arginine vasopressin from bovine sources.

The antidiuretic action of vasopressin is thought to depend upon its binding to the outer surface of the kidney tubule, resulting in an increase in the uptake of sodium from the urine into the tubule cells and, concurrently, an increase in the uptake of water. The amount of water, however, is greater than can be accounted for merely by increased diffusion of sodium into tubule cells, suggesting that ADH increases either the number of or the size of

pores on the surfaces of the cells. One stimulus that increases the release of vasopressin is a rise in the concentration of certain substances—chloride, for example—in blood plasma. These substances act directly upon the neurosecretory cells, although other receptors may also be involved. Another stimulus is a lowering of plasma volume, which probably acts chiefly through receptors in the vascular system, particularly in the heart and in the carotid blood sinuses. Both conditions necessitate increased retention of fluid; as soon as normal conditions are restored in the bloodstream, the secretion of ADH is reduced by negative feedback.

Origin of hypothalamic hormones

Oxytocin and the vasopressins are members of a series of hormones of which seven members have thus far been fully characterized. The existence of others is suspected (see Table 2). All show the same molecular structure but differ with respect to individual amino acids. The hormones comprising the series are believed to have been derived from each other by changes in genetic material (mutations) that resulted in one amino-acid substitution at a time; the starting point in the series is arginine vasotocin, which is the only one of the series found in agnathans. Two types of molecule are found in gnathostomes —a result, presumably, of a genetic duplication that established two lines of evolution. One line (basic vasopressor principles) is constituted mainly of arginine vasotocin, which is present in all gnathostomes except mammals; amino-acid substitution in the molecule gave rise to the vasopressins of mammals. The second line (neutral oxytocin-like principles) is represented by oxytocin, isotocin, glumitocin, and mesotocin. As shown in Table 2, each evolutionary line tends to have characteristic molecules, however, the molecular history in the second line is not clear. Oxytocin is thought to exist in some lower gnathostomes, and it is not yet certain whether it or mesotocin is phylogenetically the older molecule.

The functions of the hypothalamic polypeptide hormones in lower vertebrates are not yet clear, except to some extent in amphibians, in which arginine vasotocin evokes the so-called Brunn (water-balance) response; that is, water accumulates within the body as a result of a combination of increased water uptake through the skin and the wall of the bladder and decreased urinary output. This response, which also involves the uptake of sodium by the skin, is found only in the more terrestrial members of the Amphibia, in which it is an adaptation that enables them to conserve water. It is not yet known whether or not comparable adaptive specializations are associated with the molecules characteristic of the other groups of lower vertebrates. There is some evidence that hypothalamic polypeptides may be involved in the movements of water and ions (charged particles) in fishes. Changes in the functions of the polypeptide hypothalamic hormones during vertebrate evolution have occurred, partly as a result of evolution of their targets; *e.g.*, water balance in amphibians is mediated by a hormonal molecule that was already present in agnathans and was thus a part of the earliest hormonal endowment of vertebrates.

HORMONES OF THE THYROID GLAND

Biosynthesis. The two thyroid hormones, thyroxine (3,5,3′,5′-tetraiodothyronine) and 3,5,3′-triiodothyronine, are formed by the addition of iodine to an amino-acid (tyrosine) component of a glycoprotein called thyroglobulin. Thyroglobulin is stored within the gland in follicles as the main component of a substance called the thyroid colloid. This arrangement, which provides a reserve of thyroid hormones, perhaps reflects the frequent scarcity of environmental iodine, particularly on land and in fresh water. Iodine is most abundant in the sea, where thyroidal biosynthesis probably first evolved. Although the possibility that the thyroid hormones originated as metabolic by-products is suggested by the widespread occurrence in animals of the binding of iodine to tyrosine, the binding commonly results only in the formation of iodotyrosines, not the thyroid hormones. On present evidence, only the vertebrates and the closely related protochordates have a mechanism to synthesize significant amounts of biologically active thyroid hormones.

The synthesis of thyroid hormones in vertebrates begins with the active uptake by thyroid-gland cells of inorganic iodide circulating in the bloodstream; the inorganic iodide is oxidized (combined with oxygen) during a reac-

Table 2: The Structure and Distribution of the Hypothalamic Polypeptide Hormones of Vertebrates*

Basic Vasopressor Principles

1. Arginine vasotocin

Cys	Tyr	Ile	Gln	Asn	Cys	Pro	Arg	Gly(NH₂)
1	2	3	4	5	6	7	8	9

Exists in all fishes; possibly in all vertebrates, at some stage of development

2. Arginine vasopressin

Cys—Phe—Cys—Arg—|

Exists in most mammals except the pig family

3. Lysine vasopressin

Cys—Phe—Cys—Lys—|

Exists in mammals of the pig family

Neutral Oxytocinlike Principles

4. Oxytocin

Cys—Ile—Cys—Leu—|

Exists in mammals, birds, reptiles, amphibia (?), Dipnoi (?), and holocephalians (?)

5. 8 Ile oxytocin (Mesotocin)

Cys—Ile—Cys—Ile—|

Exists in reptiles, amphibia, and Dipnoi

6. 4 Ser, 8 Ile oxytocin (icthyotocin, isotocin)

Cys—Ile · Ser—Cys—Ile—|

Exists in teleost, holostean, and brachiopterygian fishes

7. 4 Ser, 8 Glu(NH₂) oxytocin (glumitocin)

Cys—Ile · Ser—Cys—Gln—|

Exists in elasmobranch fishes, particularly the skate

8. Unknown elasmobranch oxytocic principle (EOPI) (Val, Ser peptide[s]?).
Exists in *Squalus acanthias* and perhaps other sharks

*Amino acids indicated only by lines are the same as those present in arginine vasotocin.

Source: Hoar and Randall, (eds.), *Fish Physiology*, 1969.

tion catalyzed by an enzyme (iodide peroxidase). The product of this reaction (active iodine) combines with tyrosine components of the thyroglobulin molecule to form two compounds (3-monoiodotyrosine and 3,5-diiodotyrosine), which then join to form the active hormones. The synthesis of the thyroid hormones is inhibited by certain chemical agents called goitrogens, which reduce the output of thyroid hormones, thereby causing, through negative feedback, an increased output of thyrotropin and hence an enlargement of the thyroid gland (see below). Some goitrogens (*e.g.*, thiocyanates) reduce or inhibit the uptake of iodide; others (*e.g.*, thiourea, thiouracil) inhibit the peroxidase system and thus prevent the binding of iodine to thyroglobulin.

Release of the thyroid hormones into the bloodstream begins when the thyroid cells take up droplets of the stored thyroid colloid. The thyroglobulin in these droplets is then hydrolyzed (broken down in a reaction involving the elements of water) by an enzyme to form both iodotyrosines and the hormones. Normally, only the latter pass out of the cells in significant quantities. The iodine is removed from the iodotyrosines, which are not hormonally active, by an enzyme (deiodinase), and the iodine thus is conserved and used again. The hormones, usually bound to proteins (globulin and albumin) in the bloodstream, where they constitute the protein-bound iodine of the plasma, must be unbound from the proteins before they can function. The iodine is removed from the hormones largely in the liver and in the kidneys, and most of it returns to the thyroid gland, an economy that again emphasizes the need for conservation; some iodine, however, is lost in the alimentary tract.

Conservation of iodine

Synthesis of the thyroid hormones is regulated by the level of circulating hormones (*i.e.*, a negative-feedback mechanism) operating, as indicated earlier, partly by direct action on the thyrotropin-secreting cells of the pituitary gland and partly by indirect action on the hypothalamus and its thyrotropin-releasing factor. Thyrotropin attaches to the cells of the thyroid gland and may exert its effect by stimulating cAMP synthesis. It causes resorption of thyroid colloid and increases the rates of both glucose metabolism and protein synthesis as secretion of thyroid hormones increases in response to it. After the thyroid gland of the rat has been under thyrotropin stimulation for two or three hours, an increase in the size of the cells of the gland occurs, along with an increase in iodide uptake into them; prolonged thyrotropin action causes a marked enlargement of the gland (goitre), which in man may become externally apparent as a swelling. Goitres, which are of various types, result from a negative-feedback reaction that attempts to maintain output from the thyroid gland.

Effects. One established effect of the thyroid hormones in mammals is an increase in metabolic rate and in oxygen consumption, but the effects of the hormones undoubtedly are more wide-ranging than this. On the one hand, impairment of the thyroid function in mammals results in disturbances in the processes of growth and maturation. Both growth and maturation disturbances occur in the cretinous dwarfism resulting from thyroid deficiency in newborn infants; on the other hand, the metabolic effect is not apparent in lower vertebrates (*e.g.*, fish), even though treatment of these animals with thyroid hormones promotes an increase in the growth rate, provided pituitary growth hormone is also secreted. In addition, evidence suggests that, in lower vertebrates, the thyroid hormones are active during moments of stress in the life cycle (*e.g.*, migration and reproduction) and affect the activity of the central nervous system. Disturbance of thyroid output also affects reproduction in mammals, impairing the functioning of the ovary, for example, and causing irregularities of the ovarian cycle.

Effects of thyroid hormones on frog metamorphosis

The complex effects of thyroid hormones are well documented in the metamorphosis, or change in body form, of the amphibian tadpole into a frog. Metamorphosis, which involves a diversity of integrated morphological and biochemical changes, requires the presence of the thyroid gland and depends upon a delicate balance between the changing output of its hormones and changing sensitivities of the target tissues. Studies involving the tail of the frog tadpole show that the thyroid hormones directly promote the formation of the enzymes needed for reduction of the tail and suggest that the diverse effects produced in vertebrates by the thyroid hormones might depend upon their capacity to regulate protein metabolism, in which case the target cells would have to be adapted to respond by appropriate patterns of enzyme synthesis.

Ultimobranchial tissue and calcitonin. The discovery of calcitonin (thyrocalcitonin) in 1961 demonstrated the importance of comparative studies in endocrinology. It originally had been thought that this hormone, which is present in preparations made from mammalian thyroid glands, was secreted by the parathyroid glands, which in some species are combined with the thyroid gland. Later, the hormone was concluded to be a secretion of the thyroid gland itself. In fact, calcitonin is not a product of either of them. Its actual source is the ultimobranchial tissue, represented in vertebrates from fishes upward by the ultimobranchial gland, which develops from the hinder part of the pharynx. Ultimobranchial tissue is the source of distinctive cells (called light, C, or parafollicular cells), which are found in the thyroid gland of mammals; in birds, however, the ultimobranchial gland is separate, thus making it possible to remove the gland and to show that it is the source of the hormone. The molecular structure of hog calcitonin is that of a polypeptide, containing 32 amino acids and having a molecular weight of about 3,600. The calcitonin of the salmon, which is more potent than that of the pig, has the same number (but some different types) of amino acids, and the molecular weight is 3,427.

Effects of calcitonin

Calcitonin lowers the level of calcium in the blood (hypocalcemic action) when it rises above the normal level. Its secretion probably is regulated by a negative-feedback relationship between the gland and the blood plasma. The hormone affects bone, which is an active tissue. It undergoes not only growth but also remodelling as it adapts to the changing patterns of stress to which it is subjected; its calcium exchanges continuously with that of the plasma. The effect of calcitonin is to decrease the mobilization (resorption) of calcium from the skeleton into the blood plasma. In this respect, as is discussed in the next section, it is opposite in direction to the effect of parathormone of the parathyroid glands. Little is known of the action of calcitonin in the lower vertebrates, but its presence in fish raises interesting functional problems. Elasmobranch fishes (*e.g.*, sharks) lack bone, and many bony fishes have a type of bone that cannot be remodelled; the hormone, therefore, cannot act in these vertebrates as it does in higher ones. It is possible that in these fishes the hormone may control the level of plasma calcium by regulating its movement across cell membranes.

PARATHORMONE OF THE PARATHYROID GLANDS

The parathyroid glands, which are found only in terrestrial vertebrates (amphibians, birds, reptiles, and mammals), develop from certain pharyngeal pouches, which are embryonic remnants of the gill slits of fish. The parathyroid glands secrete a hormone called parathormone (PTH), which is a polypeptide of variable amino-acid composition. PTH, which consists of 83 to 85 amino acids in the human, regulates calcium metabolism in conjunction with calcitonin; its evolution in terrestrial vertebrates may have been an adaptation to the increased demand for continuous skeletal adjustments imposed by the evolution of terrestrial locomotion. Skeletal adjustments must be made without disturbing the delicate calcium balance of the rest of the body, for calcium is involved in maintaining the transport of substances through cell membranes; hence, it has an important role in muscle contractility, excitability of motor end plates in the nervous system, and coagulation of blood.

Removal of the parathyroid glands in mammals causes a fall in the level of calcium in the blood plasma, which, if sufficiently severe, is accompanied by convulsions and other symptoms resulting from increased excitability of the motor nerves. These symptoms can be corrected by

injection of appropriate preparations of parathyroid glands. The activity of the glands, like that of the ultimobranchial tissue, is regulated by negative feedback; *i.e.*, lowering of the plasma calcium level increases the output of parathormone (but decreases the output of calcitonin). The hypercalcemic effect (*i.e.*, increase in level of blood calcium) of the hormone depends largely upon its action on bone, since it promotes the transfer of calcium from this tissue into the plasma, probably by a direct action on the active bone-forming cells (osteocytes). In addition, however, parathormone promotes the formation of new bone tissue, and thus also increases its metabolic activity and the turnover of its structural material. Other effects of parathormone, at least in part, contribute to the elevation of plasma calcium; *i.e.*, PTH increases both the absorption of calcium by the intestine and its resorption by the kidney tubule. Since, however, the hypercalcemia induced by the hormone results in more of it passing into the kidney tubule, the net result may be increased excretion of calcium despite the increased resorption. Other actions of the hormone, less easy to relate to its well-defined influence upon calcium metabolism, include a regulatory influence upon the level of magnesium in blood plasma and upon the rate of removal of phosphate from urine.

In general, therefore, the action of parathormone is opposite in direction to that of calcitonin. Parathormone keeps the level of blood calcium up to its normal value; on the other hand, calcitonin ensures, through its hypocalcemic action, that the level does not rise far above this critical point. The combined actions of the two hormones serve to illustrate the importance of endocrine regulation in homeostasis. Vitamin D is a third factor in calcium regulation; its absence in young children results in skeletal malformations (rickets). Parathormone is unable to regulate the absorption and mobilization of calcium in the absence of vitamin D, which is also associated with the hormone in promoting mobilization of magnesium from bone and perhaps in the movement of phosphate within the kidney tubule.

Role of vitamin D in calcium regulation

HORMONES OF THE PANCREAS

Insulin. The vertebrate pancreas contains, in addition to the zymogen cells that secrete digestive enzymes, groups of endocrine cells called the islets of Langerhans. Certain of these cells (the B, or beta, cells) secrete the hormone insulin, inadequate production of which is responsible for the condition called diabetes mellitus. Insulin and the characteristic B cells are present in gnathostomes and in agnathans; in the latter, however, the islet cells are not associated with zymogen cells to form a typical pancreas. Insulin is a polypeptide molecule composed of two chains of amino acids, an A chain of 21 amino acids containing an intrachain disulfide linkage (— S — S —) and a B chain of 30 amino acids. The two chains are linked by two other disulfide linkages, the destruction of which destroys the activity of the molecule. It is thought that the molecule first appears in the B cell as a single-chain compound called proinsulin, which is disrupted by an enzyme-catalyzed reaction to form the two chains of the active hormone. As with other polypeptide hormones, extensive variation in amino-acid composition of the molecule occurs among different species, with the differences tending to be greater between the more widely separated species—*e.g.*, between fish and mammal. The variations in amino-acid composition have little effect on the biological activity of the molecules, but certainly influence their immunological reactions; this suggests that the two properties depend on the amino-acid sequences at different parts of the molecule.

Injection of insulin lowers blood-sugar (glucose) levels, but this so-called hypoglycemic effect is only one expression of the wide-ranging influence of insulin on storage and mobilization of energy, in which the target tissues of primary importance are muscle, adipose (fat) tissue, and liver. The actions of insulin on these tissues are varied. First, it promotes the use of the sugar glucose as an energy source; at the same time, it encourages the storage of excess carbohydrate as glycogen, the storage carbohydrate of animals. Second, insulin reduces the use of fat as an energy source and promotes its storage. Third, it reduces the use of protein as an energy source and promotes the formation of proteins from amino acids.

Insulin probably acts on carbohydrate metabolism in muscle by increasing the ability of glucose to pass through the muscle-cell membranes. This effect depends on a specific interaction between the cell membrane and the hormone; although the same effect occurs in adipose (fat) tissue, it does not occur either in the liver or in the central nervous system, despite the latter's complete dependence upon glucose for its energy supply. After the entry of glucose into a muscle cell, phosphate is added to the molecule, and two compounds form in succession, first glucose-6-phosphate, then glucose-1-phosphate; after these reactions, the metabolism of glucose is probably aided by two secondary actions of insulin (see also METABOLISM). The hormone stimulates the synthesis of an enzyme (glycogen synthetase), thus promoting the transformation of glucose-1-phosphate into glycogen; it also aids in the breakdown of glucose, thus providing energy to the cell. All of these effects contribute to the hypoglycemic (blood-glucose-lowering) action of the hormone. Insulin has other effects on muscle cells—it slows the breakdown of fat and increases the formation of proteins from amino acids. Insulin affects carbohydrate and protein metabolism in adipose tissue much as it does in muscle and also promotes storage of fat.

Effects of insulin in liver

The action of insulin in liver differs from that in muscle in that it has no direct influence upon the transport of glucose into liver cells; probably, however, insulin promotes the metabolism of glucose within liver cells in much the same way that it does in those of muscle, resulting in increased uptake of glucose from the bloodstream. In addition, insulin decreases gluconeogenesis (the formation of glucose in the liver from amino acids and other noncarbohydrate sources). The combination of these various effects causes a decrease in the level of blood glucose. Other actions of the hormone upon the liver include, as in adipose tissue, increases in fat deposition and in protein synthesis.

The diverse effects of insulin apparently are adaptively linked to regulating the storage and release of energy, but it is difficult to judge whether or not all of the effects result from a single mode of action of the hormone. The interaction of insulin with the muscle-cell membrane suggests that all of its effects might be produced by similar interactions between it and membranes within cells. The mechanism, however, has not yet been established with certainty.

The B cells of the islets of Langerhans respond directly through negative feedback to the level of glucose in the blood that reaches them; *i.e.*, an increase in blood glucose above the normal level (80 to 100 milligrams per 100 millilitres in man) brings about increased synthesis and release of insulin with the result that the level of blood glucose falls. As a consequence, the rate of insulin output then decreases. This, however, is only part of the complex hormonal mechanism that regulates carbohydrate metabolism. Another factor is the hormone glucagon, which is secreted in the islets of Langerhans by a second cell type, the A (alpha, or A_2) cells. (The function of a third type, the D (gamma, or A_1) cells, has not yet been established.)

Glucagon. Glucagon, which is present in gnathostomes but absent from agnathans, is a polypeptide molecule consisting of 29 amino acids. It strongly opposes the action of insulin, primarily through a hyperglycemic (blood-glucose-raising) effect that results from its promotion of the breakdown of glycogen (glycogenolysis) in the liver, a process that results in the formation of glucose. Glucagon exerts its action by increasing the availability of the enzyme required for the reaction by which glucose units are released from the glycogen molecule. It also reduces the rate of synthesis of glycogen, promotes the breakdown of protein, promotes the use of fat as an energy source, and evokes increased glucose uptake by muscle cells. The last effect, however, may be a

consequence of hyperglycemia induced by the increased secretion of insulin.

Another form of glucagon, called gastrointestinal glucagon, is secreted into the blood when glucose is ingested. Its only action appears to be to stimulate insulin secretion, an effect that may provide information to the islet cells of the pancreas about the entry of glucose into the bloodstream. It is also possible that pancreatic glucagon, which is secreted in the islets by the A cells, may directly stimulate the release of insulin from the adjacent B cells without actually entering the bloodstream.

Other hormones that affect insulin release

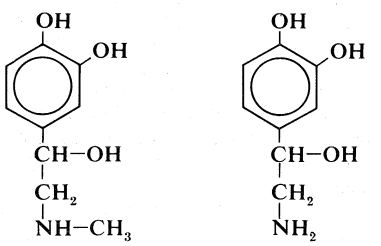
A number of other hormones also influence the release of insulin, mainly through their own actions upon blood-sugar levels. Growth hormone, thyroxine, adrenaline, and cortisol, for example, may increase insulin release because they can promote a rise in blood sugar through effects upon carbohydrate metabolism. In addition, growth hormone and cortisol can probably act directly upon the B cells.

The complexity and delicacy of the control of metabolism by insulin and other hormones in mammals illustrate again the importance of homeostasis, the control of which may not be as well organized in the lower vertebrates. Some of the responses in mammals, however, do occur in lower forms; for example, removal of pancreatic islet tissue from fishes produces hyperglycemia. Thyroxine induces hyperglycemia in amphibians, and corticosteroids promote gluconeogenesis in them. Far more information is needed, however, before the evolution of these remarkable regulating mechanisms can be determined.

HORMONES OF THE ADRENAL GLANDS

Chromaffin tissue of the medulla. The adrenal gland of mammals is composed of an outer region, the cortex, which consists of adrenocortical tissue that secretes steroid hormones (steroids are fat-soluble organic compounds), and an inner region, the medulla, which is composed of chromaffin tissue, so called because its cells contain granules that can be characteristically coloured by certain reagents. Chromaffin tissue secretes two hormones, adrenaline (epinephrine) and noradrenaline (norepinephrine), which are members of a class of compounds called catecholamines. Both chromaffin and adrenocortical tissues are present in gnathostomes and probably in agnathans (although the evidence on the latter point is not yet decisive), but the tissues vary in the degree to which they are associated, being completely separated in elasmobranch fishes.

Noradrenaline and adrenaline are each composed of a benzene ring containing two hydroxyl ($-OH$) groups and an amine (NH_2-containing) side chain (Figure 4).



adrenaline (epinephrine) noradrenaline (norepinephrine)

Figure 4: The structures of noradrenaline and adrenaline.

During the synthesis of these hormones, a sequence of enzyme-catalyzed reactions in the chromaffin granules of the secretory tissue transforms tyrosine into a compound commonly called dopa (dihydroxyphenylalanine), which then forms dopamine; dopamine then is hydroxylated (*i.e.*, an $-OH$ group is added) to form noradrenaline. Adrenaline is formed from noradrenaline by methylation (the addition of a methyl, or $-CH_3$, group), a reaction that occurs outside the granules of the chromaffin cells. Noradrenaline (but not adrenaline) is also formed in certain neurons (nerve cells), where it functions as one of the chemical transmitter substances.

After their release, both hormones are so rapidly metabolized that they probably remain in the bloodstream only for a few seconds. The first step in the breakdown, which usually occurs in the liver and kidneys, is methylation of one of the hydroxyl groups of the benzene ring; the products (metanephrine or normetanephrine), or compounds derived from them, are excreted in the urine. Small quantities (about 2 to 5 percent of the daily secretion of the gland in man) of nonmetabolized hormones are also found in the urine.

Effects of noradrenaline and adrenaline

Adrenaline and noradrenaline evoke diverse and widespread responses but differ from each other in certain of their effects (see Table 3 for their effects on man). Both influence the heart and blood vessels in ways which, although opposed to each other in a few respects, generally result in an increase in blood pressure and in output of blood from the heart. Both hormones also have metabolic actions; adrenaline, for example, like glucagon, stimulates glycogenolysis (breakdown of glycogen to glucose) in the liver, thus raising the level of blood sugar, and also increases oxygen consumption and the output of blood from the heart, probably contributing thereby to the regulation of body temperature in mammals. Adrenaline has effects upon the nervous system, recognizable subjectively in man by feelings of anxiety and of increased mental alertness.

Table 3: Effects of Adrenaline and Noradrenaline in Man

	adrenaline	noradrenaline
Heart rate	increase	decrease
Cardiac output	increase	variable
Total peripheral resistance	decrease	increase
Blood pressure	rise	greater rise
Respiration	stimulates	stimulates
Skin vessels	constriction	less constriction
Muscle vessels	dilation	constriction
Bronchus	dilation	less dilation
Eosinophil count	increase	no effect
Metabolism	increase	slight increase
Oxygen consumption	increase	no effect
Blood sugar	increase	slight increase
Central nervous system	anxiety	no effect
Uterus in late pregnancy	inhibits	stimulates
Kidney	vasoconstriction	vasoconstriction

Source: G.H. Bell, J.N. Davidson, and H. Scarborough, *Textbook of Physiology and Biochemistry*, 7th ed., 1968.

The chromaffin tissue is closely related to the sympathetic nerves of the autonomic nervous system, which innervates the components of circulation and digestion and controls their involuntary functions; in fact, the two may be said to form a sympatheticochromaffin complex. It is generally assumed that this complex acts to increase the capacity of the animal for effective action in emergencies. At such times, cardiac output increases, blood is distributed with maximum effectiveness, respiration is enhanced, and the nervous system is stimulated. The sympathetic nerves initiate these reactions and directly promote the release of adrenaline and noradrenaline because these nerves directly innervate the chromaffin cells. The hormones are thus able to develop and prolong an integrated set of responses; noradrenaline functions both as a neurohumor chemical transmitter of the sympathetic nervous system and also as a hormone of the chromaffin tissue.

The fact that adrenaline and noradrenaline, which have very similar structures (see Figure 4), can exert different actions is probably in part a consequence of the specialization of their target tissues. It has been suggested that the target tissues possess two different kinds of receptors, the alpha type, which responds to noradrenaline, and the beta type, which responds to adrenaline. Evidence for this theory is that adrenaline has a vasodilator effect (*i.e.*, it expands blood vessels), which can be blocked by certain drugs, and noradrenaline has an opposing vasoconstrictor effect, which can be blocked by other drugs.

The actions of both hormones are thought to be mediated by cAMP; alpha responses are associated with reduced synthesis of this mediator and beta ones with increased synthesis.

The interpretation of the function of these hormones in mammals has not yet been established as applicable to lower vertebrates in which the hormones are present, but they are known to influence metabolism and heartbeat in some genera. It is possible that in early stages of vertebrate evolution, the sympatheticochromaffin complex evoked more generalized physiological responses than it now does and that more precise action developed in mammals as part of their high level of homeostatic organization. Laboratory studies show that even in mammals the complex is not essential for life; animals from which it has been removed, however, are much less able to resist environmental stresses than are those whose complex is functional.

Adrenocortical tissue of the cortex. The adrenocortical tissue develops from coelomic epithelium (a cell layer surrounding the body cavity, or coelom). In this respect it resembles the endocrine tissue of the gonads, a resemblance emphasized by the fact that both the adrenocortical hormones (corticoids) and the sex hormones are steroids produced by similar metabolic pathways (the structures of some steroid hormones are found in Figure 5).

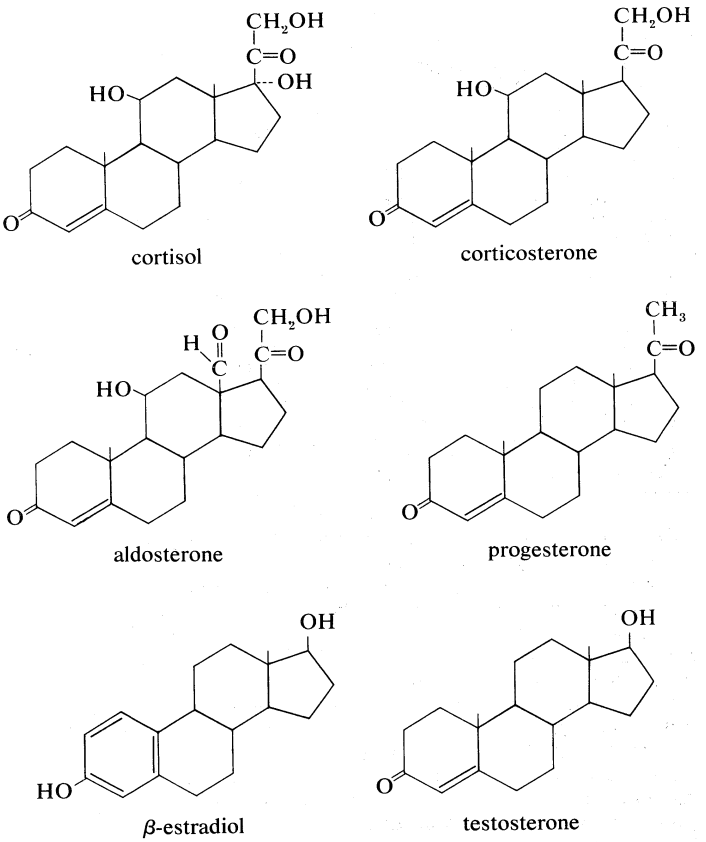


Figure 5: Some steroid hormones of vertebrates.

Active adrenocortical hormones

Many steroids have been isolated from the adrenal cortex, but in most vertebrate groups only three of them are active as hormones; they are cortisol (hydrocortisone; compound F), corticosterone (compound B), and aldosterone. Their biosynthesis is outlined in Figure 6.

The principal sterol of animals is cholesterol, which is formed by a complex series of reactions from a two-carbon compound (acetate). Progesterone, which is derived from cholesterol, can be used to form either corticosterone and aldosterone or cortisol. All three corticoids are bound to proteins during transfer in the bloodstream to their targets; cortisol, for example, is bound to a glycoprotein called transcortin. Some inactivation of the corticoids takes place in the kidney and in the alimentary tract, but most of it occurs in the liver. The metabolic products, which eventually appear in the urine, provide a basis for determination of the output of adrenocortical hormones in man.

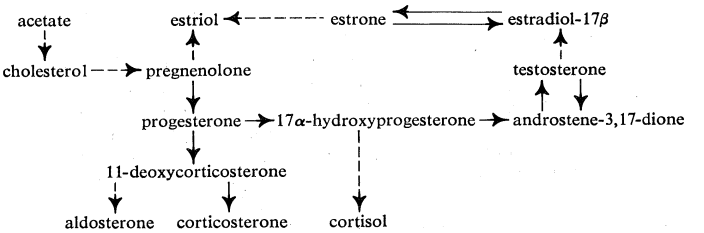


Figure 6: Metabolic pathways involved in the biosynthesis of steroid hormones. Broken lines indicate that more than one step is involved.

Adapted from D.M. Greenberg, ed., *Metabolic Pathways* (1960); Academic Press

The normal secretion of the hormones is best determined by direct measurement of the contents of the venous blood leaving the adrenal gland. In man, the daily secretion rates of the hormones, as determined by this procedure, are cortisol, 20 milligrams (one milligram = 0.001 gram); corticosterone, two to five milligrams; aldosterone, 75 to 150 micrograms (one microgram = 1,000,000th of a gram). Very small amounts of aldosterone are secreted, because the molecule has a high level of activity. Animal tissues maintained in culture fluid together with compounds from which the hormones are formed (*e.g.*, acetate, cholesterol, or progesterone containing radioactive isotopes of carbon or hydrogen) show that cortisol and corticosterone are produced in all vertebrates, including the agnathans, although the proportion of each is species-dependent; elasmobranch fishes are unique, however, in having 1-α-hydroxycorticosterone as the principal hormone. Aldosterone is produced by all terrestrial vertebrates. It has also been found in bony fishes, although its function in them has not yet been established as a hormonal one. The presence of aldosterone has not yet been established in elasmobranchs and agnathans, but whether or not this particular molecule occurs in them, the ability to synthesize corticoids must have evolved very early in vertebrate history.

In contrast to the chromaffin tissue of the adrenal medulla, the adrenocortical tissue is essential for life. Two primary functions of the corticosteroids are distinguishable in mammals. One, which contributes to the regulation of carbohydrate metabolism, is an action of cortisol and corticosterone, which are therefore called glucocorticoids. These hormones promote gluconeogenesis (formation of glucose) in the liver and are thus important in maintaining normal blood-sugar levels, particularly during glucose shortage; lack of them results in low levels of blood sugar and an increase in the sensitivity of the liver to insulin (whose effect there is to decrease gluconeogenesis). In addition, lack of the glucocorticoids is associated with a decrease in the entry of amino acids into muscles and an increase in their uptake by the liver, where enzymes required to convert amino acids to glucose must be synthesized.

In contrast to glucocorticoid action is the so-called mineralocorticoid action of aldosterone, which is manifested in mammals in the regulation of sodium metabolism. In the absence of aldosterone, sodium is lost from the body by excretion in urine; secondary consequences include a decrease in blood volume and in the filtration rate of substances through kidney structures called glomerule. Cortisol and corticosterone also play a minor part in mineral regulation, so that slight overlap in function occurs between the two corticoid types.

The action of aldosterone is exerted mainly upon the distal segment of the nephron (kidney tubule), where it promotes an increase in the permeability of the tubule membrane to the passage of sodium, and also an increase in the quantity of sodium removed into the blood from the fluid passing through the kidney tubule. At the same time, potassium and hydrogen pass into the fluid from the blood. Aldosterone also exerts other effects. It pro-

motes sodium retention in salivary glands, in sweat glands, and in the colon of the large intestine; it also promotes the excretion of magnesium in the urine. The effects of aldosterone result in an increase in the rate of synthesis of enzymes required to transport these substances through membranes.

Other actions of the corticoids are apparent in patients suffering from Addison's disease, which is caused by a general deficiency in corticoid production. A deficiency of corticoids causes disturbances in urinary output and fat metabolism, diminished resistance to stress, muscular weakness, and nervous disturbances manifested by depression and a general lack of mental alertness. The adrenocortical hormones, then, like the hormones of the chromaffin tissue of the medulla, are involved in resistance to stress. It has been postulated that the response to alarm stimuli initially involves both the sympathetico-chromaffin complex and the adrenocortical secretion; then a stage of full resistance occurs that may be followed by mental exhaustion if the alarm stimuli are prolonged. Although a close functional relationship is known to exist between the adrenocortical and chromaffin tissues in mammals, the function of the corticoids in the lower vertebrates has not yet been established. Indications are, however, that the general pattern of action may be similar; for example, the cortisol type of corticoid promotes gluconeogenesis in fish and removal of adrenocortical tissue impairs the metabolism of water and ions in the eel. Any interpretation of corticoid action in teleost, or bony, fishes has to incorporate prolactin, for, as has been noted previously, this hormone also influences the movement of ions.

Resistance to stress

In contrast to the chromaffin cells, the adrenocortical cells are not innervated. Both cortisol and corticosterone production are regulated by the action of ACTH from the pituitary gland on the zona reticularis and the zona fasciculata. The regulation of aldosterone secretion in the zona glomerulosa, however, is associated with the so-called renin–angiotensin system, which is best characterized in mammals. Renin, an enzyme with a molecular weight of about 40,000, is formed in the kidney and is released into the bloodstream, where it catalyzes the formation of angiotensin, a polypeptide molecule. Angiotensin acts upon smooth muscle and raises blood pressure. In man it reduces sodium excretion, probably by a direct action on kidney filtration, and may, in fact, be a true hormone, acting to aid sodium retention. In addition, however, angiotensin contributes to sodium retention by increasing aldosterone secretion. The exact physiological significance of the renin–angiotensin system is not yet known. In one form or another the system is probably widely distributed in vertebrates.

HORMONES OF THE REPRODUCTIVE SYSTEM

The hormones of the reproductive system of vertebrates (sex hormones) are steroids that are secreted, like those of the adrenal cortex, by tissues derived from the coelomic epithelium. Both types of secretory tissues also share biosynthetic pathways (see above *Adrenocortical tissue of the cortex*).

Female hormones. The sex hormones, together with the hypothalamic region of the forebrain and the pituitary gland, form a regulatory system, which is most complex in the female mammal. It is common for sexual activity of vertebrates to be cyclical and for the cycles to be coordinated with the seasons of the year; this ensures that the young are born at the most favourable time. In mammals, however, reproduction is complicated by the need to provide for the intrauterine life of the developing fetus and to ensure that interference by another generation of embryos cannot occur.

Estrogens. Two types of gonadal hormone, estrogens and progestins (Figure 6) are secreted in the female mammal. Estrogens are substances that evoke the cyclical onset of heat, or estrus, during which the animal is sexually active and receptive to the male. Estrus in this sense is not found in human females, but estrogens contribute to the events of the menstrual cycle, bringing about cyclical changes in the reproductive system that are comparable with those accompanying estrus in other mammals.

Hormones are secreted from the mammalian ovary by the ovarian follicle, or vesicle, including the granulosa cells immediately surrounding the ovum, or egg, and the cells of the theca, which forms a supporting outer wall for the follicle. The main estrogen secreted is called β-estradiol. The close relationship between the female and the male sex hormones is revealed by the fact that testosterone (the main male hormone) is an intermediate compound in the pathway that leads to the synthesis of estradiol, although another route, which avoids the formation of testosterone, is possible. Other estrogens are also known; the most familiar ones in man and other mammals, estrone and estriol, are much less active than estradiol, estriol being the weakest. Estrone can be converted to estradiol and vice versa in the ovary and in other tissues; *e.g.*, estradiol is converted, particularly in the liver, to estriol, which is an excretory product. The metabolism of these compounds is complex; they may be combined in part with other substances, or they may pass through the bile into the intestine for reabsorption and circulation through the body before excretion in the urine occurs. Their urinary concentrations provide an important clinical index of reproductive function.

Types of estrogens

Estrogens are concerned not only with reproductive behaviour but also with the general maintenance of the sexual organization of the female. When estradiol is administered to a mammal, the hormone becomes bound to uterine tissue, where it increases the rates of protein synthesis, of uptake of water and glucose, and, eventually, of growth of the lining epithelium and underlying muscular tissue (endometrium) of the uterus. Estradiol also evokes changes in the vagina, including hardening of the epithelium, a phenomenon that, in the laboratory rat, is used to determine its sexual condition. Estradiol and other estrogens have also been found in fishes and in other lower vertebrates. As with the corticosteroids, the sex hormones evolved very early in vertebrate history. Indeed, they have even been identified in invertebrates—in the eggs of the lobster, for example, and in the ovaries of starfishes, where, however, they may be no more than by-products of the metabolism of other steroids. Estrogenic activity is not necessarily restricted to steroids; for example, the estrogen mirestrol from a Thailand plant is not a steroid, nor is the very potent synthetic estrogen, stilbestrol, which is widely used in medicine.

Progestins. Progestins, of which the most important is progesterone, are concerned with the maintenance of pregnancy. Progesterone, therefore, evolved in viviparous mammals; *i.e.*, those that produce living young. Its chemical origin is demonstrable, since it is also an important intermediate compound in the biosynthetic pathways leading to corticoid and estrogen production. Mammals thus converted to hormonal use a substance that was synthesized by vertebrates long before the evolution of terrestrial vertebrates.

Some progesterone is probably formed in the ovarian follicle, but the main site of production is the corpus luteum, which is formed by a transformation of the follicle after ovulation; the secretory cells are formed from granulosa cells. The functions of the two important follicular phases, preceding and following ovulation, therefore, are continuous. The hormone is metabolized in several ways, but one important product is pregnanediol; formed mainly in the liver, it appears in part in the urine, where it can be measured to determine the degree of ovarian function.

The transformation of the follicle into the corpus luteum is an important turning point in the diphasic menstrual cycle of women and in the ovarian cycles of other mammals, from which the human cycle evolved. Progesterone prepares the uterus for the implantation of fertilized eggs, and it is also needed for the maintenance of pregnancy once implantation has taken place. It evokes a reduction in the ability of the uterine walls to contract, a proliferation of the glands of the endometrium, and the formation of glycogen. In addition, through its

Role of progesterone

feedback action upon pituitary secretion, progesterone inhibits further ovulation (see below), thus ensuring undisturbed fetal development. Ovulation in women occurs at about the middle of the monthly cycle, and the follicular phase is succeeded by the luteal phase. The vaginal bleeding at the end of the cycle is an indication that ovulation has not been followed by implantation of a fertilized egg and is immediately followed by the inception of a new cycle. If implantation does take place, the uterus provides metabolic support for the fetus until birth (see also MENSTRUATION).

The vertebrate reproductive cycle depends upon delicate interrelationships between the sex hormones and the pituitary gonadotropic hormones (FSH and LH). As mentioned, it is uncertain whether or not there are two distinct gonadotropins in lower forms, but their separate action is well defined in mammals. Broadly speaking, FSH (follicle-stimulating hormone), with some support from LH (luteinizing hormone), promotes growth and secretory activity of the follicle. The increasing output of estrogen from the ovary eventually tends, by feedback to the pituitary gland, to reduce FSH output and to stimulate the secretion of LH; it is a sudden peak release of the latter hormone that evokes ovulation in many mammals. In others, such as the cat and rabbit, however, ovulation occurs as a response to the stimulus of copulation. Although some progesterone may be secreted by the granulosa cells of the follicle, the development of the corpus luteum greatly increases its secretion. Luteotropic activity in one form or another (the action of prolactin, for example, in the rat) is important in the early stage of this phase. Progesterone, in conjunction with the estrogen that is also being secreted (in part, probably, by the corpus luteum), suppresses further ovulation. This interaction of the two hormones is the basis of the design of contraceptive pills.

The corpus luteum continues to function during pregnancy, supplemented (in eutherian, or placental, mammals but not in marsupials) by endocrine secretions of the placenta (the organ through which contact between mother and fetus is maintained). The hormonal activity of the placenta varies with the species; in man, for example, the placenta secretes two gonadotropins called human chorionic gonadotropin (HCG) and human placental lactogen (HPL). HCG, like the pituitary gonadotropins, is a glycoprotein, with a molecular weight of 25,000 to 30,000. HPL is a protein, with a molecular weight variously estimated at about 19,000 or 30,000. One or perhaps both of these hormones, which become detectable during the early weeks of human pregnancy, probably stimulate luteal secretion. After two months the human placenta begins to manufacture estrogen and progestin; as a consequence, the corpus luteum is no longer needed for the maintenance of pregnancy. Much of the estrogen, although synthesized in the placenta, is derived from a compound (dehydroepiandrosterone) formed in the adrenal glands of the fetus. The placenta and the fetus thus form an integrated endocrine complex, a striking index of the high level of specialization found in the regulation of mammalian reproduction. (See also PREGNANCY.)

The placenta probably secretes a luteotropin in all mammalian species, thereby contributing to prolongation of the life of the corpus luteum. In the mare and the monkey the placenta also secretes estrogen and progesterone, as in man, but in the mouse and rabbit it secretes only estrogen, and in the hamster and rat it secretes neither. In these last four species and in others like them, in which the placenta cannot substitute completely for the corpus luteum, ovariectomy (removal of the ovaries) of a pregnant female leads to the termination of pregnancy unless progesterone is administered to the female.

The interrelationships between the ovarian and the pituitary hormones are based upon negative feedback involving both the cells of the pituitary and those of the hypothalamus, which contains centres that are excited or inhibited by gonadotropin released from the pituitary gland. It is the hypothalamic involvement that enables vertebrate reproductive cycles to be adjusted by the central nervous system relative to external stimuli, particularly the seasonal fluctuations of daylight and other environmental factors that determine the onset of reproduction in many vertebrate species.

Male hormones. The sex hormones of the male follow a much simpler pattern than do those of the female, although the same principle of interaction exists between the pituitary gland and the gonads. The latter organs, the testes, secrete steroids called androgens, which are responsible for the maintenance of male characteristics and behaviour. FSH (follicle-stimulating hormone) from the pituitary gland stimulates the growth of the seminiferous tubules that constitute much of the structure of the testes and promotes within them the cell divisions that result in the production of mature sperm. LH (luteinizing hormone) from the pituitary gland promotes the development within the testes of endocrine tissue, which is composed of groups of cells (interstitial tissue) between the seminiferous tubules. The interstitial tissue of certain bony fishes, however, is represented by cells, called lobule boundary cells, situated within the tubule tissue.

Androgens

Under the influence of LH (often called ICSH, or interstitial-cell-stimulating hormone, in males), the interstitial tissue secretes the steroid hormone testosterone, which is the most important vertebrate androgen. The fact that it is an intermediate compound in the metabolic pathway of estrogen synthesis accounts for the origin of some forms of abnormal sexual organization in man; for example, the testes may secrete predominantly estrogen instead of androgen, resulting in markedly female appearance and behaviour in a male. Although testosterone may be secreted by the adrenal cortex, occasionally producing sexual disturbances, the amount of secretion is not normally significant. Testosterone, which is bound to a protein as it circulates in human blood, can be converted to the compound (androstenedione) from which it is formed, especially in the liver and in muscle; both compounds are metabolized, mainly in the liver, to substances that are excreted in urine. Very small quantities of testosterone can also be excreted in urine, and the quantities of testosterone and compounds derived from it frequently are measured to provide an index of testicular condition.

In addition to promoting male characteristics, male behaviour, and the maintenance of the spermatic tubules, testosterone, in the presence of normal amounts of growth hormone, also promotes growth of the bony skeleton. The reason for rapid growth at puberty is that the secretion of androgen markedly increases. The hormone brings about the closure of the epiphyses (ends) of the long bones, which completes the process of growth (estrogens have a similar action in the female). Thus, as often occurs among animals, growth ceases before full reproductive activity is attained, and competition between two processes, both of which make heavy demands upon the resources of the body, is avoided.

HORMONES OF THE DIGESTIVE SYSTEM

In vertebrates, the muscular and secretory activities of the alimentary canal and its associated glands are regulated by nervous and hormonal mechanisms. The hormones comprise a self-contained complex that functions at a relatively primitive level of organization and is distinguished by peculiar features; for example, specialized glandular tissues that secrete the hormones cannot be identified, although certain cells that can be seen in the wall of the alimentary canal are thought to be involved in their production. In addition, the digestive hormones regulate the system that produces them and are largely independent of the rest of the endocrine system, although certain relationships may not yet have been discovered.

Functions of digestive hormones

The functions of digestive hormones are best understood in mammals, in whom at least three are well characterized; the existence of others has been postulated. The three hormones—gastrin, secretin, and cholecystokinin/pancreozymin (CCK-PZ)—are polypeptide molecules whose amino-acid sequences are known. When food enters the stomach, the wall of its pyloric end (the area at which the stomach joins the small intestine) releases a hormone called gastrin, which promotes the flow of acid

from the gastric glands in the stomach. These glands also release pepsinogen, which is the inactive form of the protein-digesting enzyme pepsin, but this process is primarily under nervous control. The entry of the acidified stomach contents into the first part of the small intestine (duodenum) releases secretin and cholecystokinin/pancreozymin. Secretin evokes the discharge of fluid and bicarbonate ions from the pancreas (hydrelatic action) and promotes the secretion of bile from the liver (chloretic action). Cholecystokinin/pancreozymin, so-called because its two main actions were formerly attributed to two separate hormones, evokes the release of enzymes from the pancreas (ecbolic action) and causes contraction of the gallbladder (cystokinetic action), thereby promoting the entry of bile into the duodenum.

Little is known regarding hormonal control of alimentary activities in lower vertebrates; however, hydrelatic, ecbolic, and cystokinetic activities are present in preparations of the alimentary tracts of both agnathans and gnathostomes, indicating that substances able to regulate digestive activity appeared very early in the evolution of the vertebrate alimentary tract. Evidence suggests that the appearance of these hormones may have resulted in molecular diversification similar to examples previously discussed. The structure of the glucagon molecule from the pancreas, for example, is similar to that of secretin in that each molecule includes the same 15 amino acids located in the same positions. It has therefore been suggested that the two hormones may have evolved from a common ancestral molecule.

ENDOCRINE-LIKE GLANDS AND SECRETIONS

In addition to the well-defined hormones, other substances, which are found in blood and in tissues and are of uncertain function, may be concerned in various ways with physiological regulation in vertebrates, although their hormonal status has not yet been established.

Blood contains substances called kinins, which are polypeptides that originate in the blood and perhaps elsewhere; bradykinin, for example, causes contraction of most smooth muscles and has a very potent action in dilating certain blood vessels. Its function, which is not yet established, may be to regulate the rate of blood flow or to participate in the inflammatory response of an animal to injury.

Some endocrine-like glands are associated with organs. One example in mammals is the carotid bodies, which are found on the carotid arteries that supply blood to the head. The carotid glands are stimulated by a decrease in the oxygen content of the blood and are considered to be the source of a substance, the nature of which has not yet been established with certainty, that promotes the process of red-blood-cell formation (erythropoiesis).

The pineal organ is an endocrine-like body found in the brain of all vertebrates. In lower vertebrates, it contains sensory and supporting cells and functions as a light-sensitive organ; in higher vertebrates, beginning with amphibians, the pineal gland has secretory functions, and in mammals, it is exclusively a secretory organ, producing from an amino acid (tryptophan) the compound serotonin (5-hydroxytryptamine, or 5HT) and a derivative of serotonin called melatonin. Preparations of melatonin, when given to amphibians, stimulate the concentration of pigment granules in chromatophores, an effect comparable with that of intermedin (MSH) but much more powerful. The normal physiological function of melatonin in higher vertebrates has not yet been established, although involvement in the regulation of reproduction is suspected. Serotonin is widely distributed in animals, especially in the brain and alimentary tract of vertebrates; it may function as a neurohumor in the invertebrate mollusks, but its significance in other animals is not yet certain.

The thymus is essential for the normal development in mammals of the system responsible for immunological responses. Its removal in newborn mice results in a deficiency of one type of white blood cells (lymphocytes) and a consequent likelihood of early death from infection. Preparations of thymus glands from various species contain a protein component, called thymosin, that promotes the development of lymphocytes. Although thymosin is sometimes regarded as a possible thymus hormone, the evidence is not yet complete.

Urohypophysis and corpuscles of Stannius

The urohypophysis, an organ found only in elasmobranch and bony fishes, probably developed independently in each group. The neurosecretory cells comprising the urohypophysis are concentrated at the hind end of the spinal cord, where they are associated with a vascular plexus to form a neurohemal organ. The urohypophysis resembles the neurosecretory system of the hypothalamus and the neural lobe (see above *Neurohypophysis and the polypeptide hormones of the hypothalamus*), but its functions have not yet been established.

The corpuscles of Stannius, found only in bony fishes, are sac-like bodies in the kidney. Although they were once thought to be a form of adrenocortical tissue, they differ from it in embryological origin as well as in cytological characteristics; moreover, although the corpuscles of Stannius are capable of limited steroid biosynthesis, they cannot convert cholesterol into corticoids, a process that occurs in adrenocortical tissue. Evidence suggests that these corpuscles secrete some substance, as yet uncharacterized, which plays a part in maintaining ionic homeostasis, perhaps in conjunction with the corticoid hormones.

III. The hormones of invertebrates

Some form of endocrine regulation probably occurs in all invertebrates; in arthropods (as exemplified in insects and crustaceans) it attains a level of complexity similar to that of vertebrates.

HORMONES OF INSECTS

Insects secrete hormones from neurosecretory cells and also from endocrine glands. Important neurosecretory centres occur in the pars intercerebralis region of the brain. The several cell types found in these centres indicate that more than one hormone is produced there.

Neurohormones. One of the brain hormones is thoracotropic hormone. This is released from nerve endings located in a neurohemal organ called the corpus cardiacum; the relationship between the corpus cardiacum and the brain closely parallels that between the neural lobe of the pituitary gland and the hypothalamic region of the brain of vertebrates. The thoracotropic hormone is transferred in the blood to the thoracic glands in the body region called the thorax. It stimulates the production and release from the glands of ecdysone, a hormone that initiates molting, which is the periodic shedding of the outer skeleton that typically occurs in insects and other arthropods. The thoracotropic hormone is probably a polypeptide molecule.

A neurosecretion of the insect brain distinct from the thoracotropic hormone and called bursicon acts directly on the adult cuticle (skin) of arthropods to stimulate darkening and hardening processes. Bursicon is almost certainly a polypeptide, with a molecular weight of about 40,000. The brain of insects also produces a third neurohormone, which has a hyperglycemic (increase in level of blood glucose) effect in a tissue mass called the fat body, and a fourth neurohormone, which acts on the malpighian tubules (excretory organs) and rectum to facilitate the removal of excess fluid taken in with food; these two hormones, which may actually be one hormone with two effects, are also probably polypeptides. Another more active hyperglycemic hormone is formed within the corpora cardiaca, perhaps under the control of cerebral neurosecretion.

Molting hormones. Ecdysone is a steroid compound derived from cholesterol. Two forms are found in insects—α-ecdysone and β-ecdysone; ecdysones of unknown biological significance are also present in plants. Unlike vertebrates, insects cannot synthesize cholesterol, and they thus must obtain it from their food. Evidence concerning the mode of action of ecdysone indicates that it has a direct action upon the synthesis of the ribonucleic acid (RNA) that controls protein synthesis in the cell.

The distinction in insects between molts that occur

Juvenile hormone

within the larval stage of development and those that result in the transformation of larvae to other stages (pupae, adults) in the life cycle is controlled by another hormone, called juvenile hormone, which is secreted in epithelial glands, called the corpora allata, near the brain. The hormone controls the appearance of juvenile characters in larval stages, presumably by suppressing the activity of genes concerned with the expression of adult characters; reduction in the amount of or absence of the hormone at later molts results in the appearance of mature characters. The hormone nevertheless may continue to function in adults and often is necessary for normal egg production in females. Juvenile hormone is a lipoidal (fatlike) compound of similar structure from all sources. Many synthetic compounds mimic its effect, as do certain natural products—*e.g.*, substances in the balsam fir tree (*Abies*). Substances that mimic the action of juvenile hormone sometimes are used as insecticides, for if they are present in abnormal amounts in the later stages of the life cycle, they kill the insects.

Pheromones. Pheromones are important as insect sex attractants and as regulators of the social organization of social insects; *e.g.*, bees. The sex attractant of the female silk moth (*Bombyx mori*) is called bombykol. A related compound, gyptol, is the sex attractant of the female gypsy moth (*Porthetria dispar*), and gyplure is a synthetic compound that acts as an even more powerful attractant. The use of these compounds in the chemical control of insect pests is probably more promising than is the use of juvenile hormone (see also PEST CONTROL).

The odour of the sex attractant of the honeybee (*Apis mellifera*), 9-oxodecenoic acid, stimulates the olfactory receptors of the drones (males). Secreted by the queen bee in the hive, the pheromone inhibits the development of the ovaries of the worker bees (sterile females) but is entirely effective only when it acts in conjunction with another inhibitory pheromone, 9-hydroxydecenoic acid. Removal of the queen from the hive results in the building of new queen cells by the workers and the development of functional ovaries in the drones. The mechanism by which these inhibitory substances function is not yet understood; some effect upon the nervous system presumably is involved.

HORMONES OF CRUSTACEANS

The endocrine systems of crustaceans resemble those of insects; important differences occur, however, implying extensive independent evolution in the two groups. The main sources of neurohormones are groups of cells (the X-organs) located in the optic ganglia of the eyestalks; the most important neurohemal organ is the sinus gland beside the eyestalks. Less important neurosecretory centres and neurohemal organs also occur. The pericardial organ of decapods, for example, is a group of nerve fibres and nerve endings in the walls of the pericardium, which encloses the heart; the pericardial organ secretes a substance, perhaps a polypeptide neurohormone, which accelerates the heartbeat.

In crustaceans, neurosecretory control is exerted over many functions, including the movement of pigment in the chromatophores, which determine body colour, and in the retina of the compound eye. Neurosecretions also regulate molting and the metabolic functions associated with it by actions exerted upon the so-called Y-organ in the head; this organ so closely resembles the thoracic gland of insects that the two may share a common ancestry. In crustaceans, however, the neurosecretion inhibits secretions from the Y-organ, and the molt is initiated by the withdrawal of the inhibitory hormone (in insects, the thoracotropic hormone from the corpus cardiacum stimulates the secretion of the molting hormone, ecdysone, from the thoracic gland). Neurosecretory hormones of crustaceans have diverse chemical and biological characteristics but apparently are polypeptides, as are the neurosecretory hormones of vertebrates.

Androgenic gland

Unlike insects, crustaceans have an androgenic gland, which typically is located on the genital duct (vas deferens) of the male. The androgenic gland secretes a hormone, possibly steroid in nature, that controls both the differentiation of the gonad of the male into a testis and the male characteristics of its limbs. The absence of the androgenic gland in the female results in the formation of an ovary, which subsequently synthesizes one or more hormones that, in female amphipods, promote the development of brood chambers (in which the young are hatched) and other structures associated with reproduction.

OTHER INVERTEBRATE HORMONES

The characterization of the hormones of other invertebrates awaits further study. Evidence indicates that the brain of polychaete worms produces neurosecretions that regulate growth and reproduction; in *Nereis* and *Nephtys* the neurosecretory fibres apparently have a close and presumably functional relationship with an epithelial gland (infracerebral organ), which is formed from coelomic epithelium and is situated on the wall of the brain.

Neurosecretory cells probably are present in mollusks such as gastropods and lamellibranchs. Experimental studies indicate an endocrine relationship in gastropods between the gonad (ovotestis) and possible neurosecretory cells in the tentacles and the brain; one ganglion of the gastropod *Lymnaea* may secrete a neurohormone with a diuretic (urine producing) action. Epithelial glands in mollusks are important; in the cephalopods, which are the most advanced invertebrates in some respects, optic glands on the optic stalks (eyestalks) secrete a hormone that promotes development and maturation of the gonads. In immature cephalopods the activity of the glands is inhibited by the central nervous system, apparently by a chemical mediator that diffuses from nerve fibres.

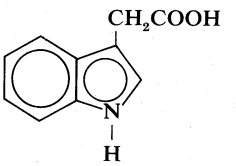
The nerve net, which constitutes the very primitive nervous system of the coelenterates, probably the most primitive multicellular animals, apparently contains neurosecretory cells; indirect but convincing evidence suggests that the cells release a secretion that promotes growth and inhibits sexual reproduction.

IV. The hormones of plants

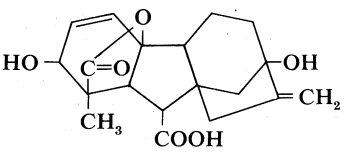
Growth in plants is regulated by four categories of phytohormones—auxins, gibberellins, cytokinins, and inhibitors.

GROWTH PROMOTORS

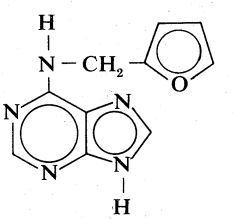
Auxins. The distribution of auxins, which promote the lengthwise growth of plants, is correlated with the distribution of the growth regions of the plant. The most important auxin, whose structure is represented in Figure 7, is β-indolylacetic acid (IAA), which is formed either from the amino acid tryptophan or from the breakdown of carbohydrates known as glycosides. The hormone affects plants by its action on chemical bonds of



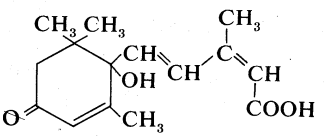
β-indolylacetic acid



gibberellic acid



kinetin (6-furfurylaminopurine)



abscisic acid

Figure 7: The structures of plant hormones.

carbohydrates comprising plant cell walls. The process permits the cells to be irreversibly deformed and is accompanied by the entry of water and the synthesis of new cell-wall material. Many animal hormones may exert their effects by influencing protein synthesis, and evidence suggests that auxins may act in a similar way.

Many other naturally occurring and synthetic compounds called auxins also have growth-promoting properties, but they are not always as active as IAA. Some of these compounds, however, resist the enzymatic destruction that is the normal fate of IAA within the plant; this feature is of great value in research and in horticulture, because auxin action can be prolonged. Other auxin-like compounds are used as selective weed killers (*e.g.*, to disturb the leaf growth of dicotyledonous plants either in fields containing monocotyledonous cereal crops or on lawns) and as agents that remove leaves from dicotyledonous plants (defoliating agents).

Effects of indolylacetic acid

The hormonal characteristics of IAA are readily demonstrated in grass seedlings, in which the hormone is synthesized at the tip of the coleoptile (the protective sheath of the emerging plumule, or embryonic bud) and passes downward to its point of action in the growing region, where it evokes elongation of the coleoptile cells; growth stops if the tip is removed. The movement of the hormone downward from the tip of the coleoptile depends upon an interaction between the hormone and the cells through which the movement normally takes place.

In addition to promoting normal growth in plant length, auxins influence the growth of stems toward the light (phototropism) and against the force of gravity (geotropism). The phototropic response occurs because greater quantities of auxin are distributed to the side away from the light than to the side toward it; the geotropic response occurs because more auxin accumulates along the lower side of the coleoptile than along the upper side. The downward growth of roots is also associated with a greater quantity of auxin in their lower halves. This effect, which is the opposite to that found in coleoptiles, is attributed to an inhibitory action of auxins on root growth, but this aspect of auxin action is not yet fully understood. Auxins have actions other than those associated with promoting growth; *e.g.*, they play a role in cell division, in cell differentiation, in fruit development, in the formation of roots from cuttings, and in leaf fall (abscission). In experimental conditions, auxins tend to inhibit the progress of plant aging, perhaps because of their stimulating effect upon protein synthesis.

Gibberellins. Gibberellins are named after the fungus *Gibberella fujikuroi*, which produces excessive growth and poor yield in rice plants. One gibberellin (Figure 5) is gibberellic acid (GA_3), which is present in higher plants as well as in fungi; many related compounds have structural variations that correlate with marked differences in effectiveness.

Gibberellins, abundant in seeds, are also formed in young leaves and in roots; they move upward from the roots in the xylem (woody tissue) and thus do not show the movement characteristic of auxins. Evidence suggests that gibberellins promote the growth of main stems, especially when applied to the whole plant. Unlike the auxins, gibberellins have little effect upon pieces of coleoptile in tissue culture. Gibberellins promote the growth of dwarf peas and are involved in the bolting (elongation) of rosette plants such as the carrot. Elongation of rosette plants occurs after exposure to certain environmental stimulation (*e.g.*, cold, or long periods of daylight), which is accompanied by an increase in the gibberellin content of the affected plant. In experimental conditions gibberellins tend, like auxins, to retard senescence.

Cytokinins. Cytokinins are compounds derived from a nitrogen-containing compound (adenine). One cytokinin is 6-furfurylaminopurine (kinetin, see Figure 7); other compounds derived from adenine with effects similar to those of kinetin, and certain compounds derived from another nitrogen-containing compound, urea, are conveniently referred to as cytokinins, although not all are natural products. Cytokinins are synthesized in roots, from which, like the gibberellins, they move upward in the xylem and pass into the leaves and the fruit. Required for normal growth and differentiation, cytokinins act, in conjunction with auxins, to promote cell division and to retard senescence, which, at least in its early stages, is an organized phase of metabolism and not just a breakdown of tissue. An example of senescence is the yellowing of isolated leaves, which occurs as proteins are broken down and chlorophyll is destroyed. Cytokinins, which prevent yellowing by stabilizing the content of protein and chlorophyll in the leaf and the structure of chloroplasts, are used commercially in the storage of green vegetables.

GROWTH INHIBITORS

Growth inhibitors of various types have been identified in plants. The best characterized one is abscisic acid (Figure 7), which is chemically related to the cytokinins. It is probably universally distributed in higher plants and has a variety of actions; for example, it promotes abscission (leaf fall), the development of dormancy in buds, and the formation of potato tubers. The mode of action of abscisic acid has not yet been clarified but is thought to involve the direct inhibition of the synthesis of RNA and protein.

Another growth inhibitor is ethylene, which is a natural product of plants, formed possibly from linolenic acid (a fatty acid) or from methionine (an amino acid). Ethylene promotes abscission in senescent leaves, perhaps by facilitating the destruction of auxin. Its effects extend beyond that of inhibiting growth; in fruit, for example, ethylene is regarded as a ripening hormone. Involved in its action in fruit is another factor, perhaps auxin or another growth-regulating hormone, which influences the ethylene sensitivity of the tissues.

Hormonal interaction

The hormonal interaction conspicuous in animals is found also in plants; one example is the control of abscission, which requires the synthesis of enzymes at an abscission zone, at the base of the structure concerned, to catalyze reactions involving breakdown of cell walls. Auxin reaching the abscission zone from the tip of the structure promotes abscission; if auxin reaches the structure from the opposite direction, however, it tends to inhibit the process, probably by its influence on metabolism. Other hormones are also involved in abscission; ethylene stimulates the synthesis of the enzymes, and abscisic acid accelerates the associated senescence. Gibberellin tends to inhibit abscission by promoting growth.

Another example of hormonal interaction occurs during the germination of cereal seeds. The embryo (germ) is first activated by uptake of water, which enables it to produce gibberellin. Gibberellin acts on the living cells (aleurone layer) surrounding the food reserves (endosperm). This action induces the aleurone cells to produce enzymes that break down starch to sugars and release tryptophan from the protein of the endosperm. The tryptophan migrates to the coleoptile tip and is transformed into indolylacetic acid, which in turn moves to the growth zone and weakens the cell walls, thus permitting water uptake.

The target tissues probably play a role in such sequential actions, and it is likely that changes in their responsiveness to hormonal action, perhaps correlated with environmental stimuli, contribute to adaptive integration. The similarities in the hormonal mechanisms of plants and animals, two groups that are so profoundly different in their structure and mode of life, effectively illustrate the fundamental uniformity of biological organization.

BIBLIOGRAPHY. E.J.W. BARRINGTON, *An Introduction to General and Comparative Endocrinology* (1963), a textbook of university level, treating general principles, mainly vertebrate, but with two chapters on invertebrates; *Hormones and Evolution* (1964), for students and general readers, concerned with the molecular structure and mode of action of hormones in relation to evolutionary theory; G.K. BENSON and J.G. PHILLIPS (eds.), *Hormones and the Environment* (1970), a wide-ranging review and research symposium, primarily for specialists but also of general interest; M. BLACK and J. EDELMAN, *Plant Growth* (1970), an elementary treatment, with at-

tention to plant hormones; W.R. BUTT, *Hormone Chemistry* (1967), a correlation of research on the major mammalian hormones, requiring some knowledge of chemistry; R.E. COUPLAND, *The Natural History of the Chromaffin Cell* (1965), an integrated treatment of research on structure and function, considered at all levels of analysis, from the gross anatomical to the molecular; J. EBLING and K.C. HIGHNAM, *Chemical Communication* (1969), a concise elementary introduction for high school and first-year university students; B.E. FRYE, *Hormonal Control in Vertebrates* (1967), an elementary introduction, with emphasis on general principles and physiological adaptation; M. GABE, *Neurosecretion* (1966), an authoritative and comprehensive treatment for advanced students; A. GORBMAN and H.A. BERN, *A Textbook of Comparative Endocrinology* (1962) a textbook of university level; G.W. HARRIS, *Neural Control of the Pituitary Gland* (1955), an authoritative monograph for students and research workers; K.C. HIGHNAM and L. HILL, *The Comparative Endocrinology of the Invertebrates* (1969), a comprehensive review of endocrine principles and experiments for students; I. CHESTER JONES, *The Adrenal Cortex* (1957), an authoritative review of the literature up to 1955 and a valuable foundation treatment of the subject; M. PICKFORD, *The Central Role of Hormones* (1969), an elementary treatment for students and general readers with some knowledge of vertebrate biology; C.T. SAWIN, *The Hormones: Endocrine Physiology* (1969), a clear treatment of hormone action for university students; E. and B. SCHARRER, *Neuroendocrinology* (1963), an authoritative account, that discusses, by reference to selected examples, comparative biological aspects rather than chemical or clinical ones.

(E.J.W.B.)

Horse

Zoologically, the horse is a mammal of the family Equidae. It comprises a single species, *Equus caballus*, whose numerous varieties are called breeds. The earliest traces of the ancestors of the horse have been found in rocks from the Eocene epoch (50,000,000 years ago); evidence suggests a small fox-like animal that walked on four toes but possessed a fifth in the form of a useless vestige. Fossil skeletons of this animal, called *Eohippus*, were found in the Mississippi Valley area of the U.S., from which the animal disappeared in an unexplained manner after first giving rise to the line that produced *Orohippus*, *Protohippus*, *Hipparion*, and, finally, *Equus caballus*. In this series the animal developed in size and shape and came to stand on one toe enveloped in a protective hoof.

GENERAL FEATURES

Historic background. *Equus caballus* first appeared in Central Asia as a sturdy, sand-brown animal called Przewalski's horse (*E. caballus przewalskii*). It spread eastward, giving rise to Chinese and Mongolian breeds, and farther westward there arose the mouse-coloured European Tarpan. Wild horses migrated to the southwest, to Asia Minor, and from there to Egypt and the Mediterranean countries, establishing local breeds in those parts of the world. In 1519 the Spanish explorer Hernando Cortéz transported the horse to the New World.

Utility of the wild horse

In prehistoric times the wild horse was probably first hunted for food. When its domestication took place is unknown, but it certainly was long after the domestication of the dog or of cattle. It is supposed that the horse was first used by a tribe of Indo-European origin that lived in the steppes north of the chain of mountains adjacent to the Black and Caspian seas. Influenced by climate, food, and man, the horse rapidly acquired its present form.

Significance. The relationship of the horse to man has been unique. The horse was man's partner and friend, carrying him above his fellowmen on foot and giving him power and speed. It plowed his fields and brought in his harvest, hauled goods and conveyed passengers, followed game and, later, tracked cattle, and carried combatants into battle and adventurers to unknown lands. It has provided recreation in the form of jousts, tournaments, carrousels, and in the sport of riding. The widespread influence of the horse is expressed in the English language in such terms as chivalry, cavalier, and cavalry, which generally connote honour, respect, good manners, and straightforwardness. (For an account of the training and use of the horse in riding, see RIDING AND HORSEMANSHIP; for the racing of horses see HORSE RACING.)

The horse is the "proudest conquest of Man," according to the French zoologist Le Comte de Buffon. Its place was at its master's side in the graves of the Scythian kings or in the tombs of the Pharaohs. Many of man's early cultures were centred on possession of the horse. Superstition read meaning into the colours of the horse, and a horse's head suspended near a grave or sanctuary or on the gables of a house conferred supernatural powers on the place. Greek mythology created the centaur, the most obvious symbol of the oneness of horse and rider. White stallions were the supreme sacrifice to the gods, and the Greek general Xenophon recorded that "Gods and heroes are depicted on well trained horses." A beautiful and well trained horse was, therefore, a status symbol in ancient Greece. Kings, generals, and statesmen of necessity had to be horsemen. The names of famous horses are inseparably linked to those of their famous riders: Bucephalus, the charger of Alexander the Great; Incitatus, foolishly made a senator by the Roman emperor Caligula; El Morzillo, Cortéz' favourite horse, to whom the Indians erected a statue; Roan Barbery, the stallion of Richard II, mentioned by Shakespeare; Copenhagen, the Duke of Wellington's horse, which was buried with military honours.

The horse in art

The horse has occupied a special place in the realm of art. From the Stone Age drawings to the marvel of the Parthenon frieze, from Chinese T'ang dynasty tomb sculptures to Leonardo da Vinci's sketches and Verrocchio's Colleoni, from the Qur'ān to modern literature, the horse has inspired artists of all ages and in all parts of the world.

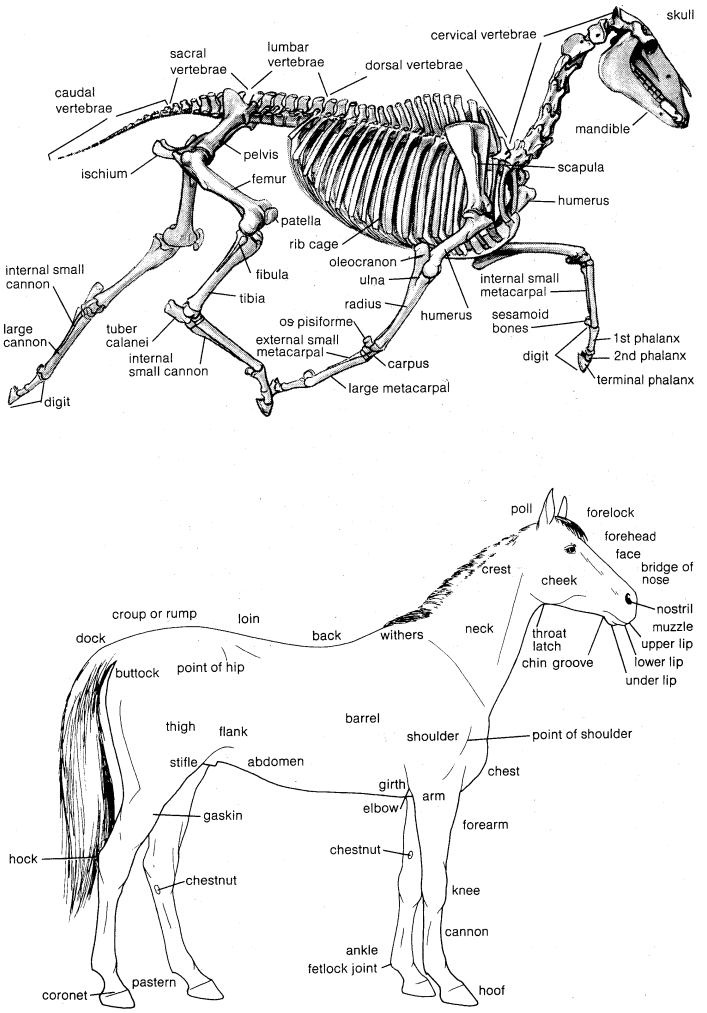
The horse in life serves man in his travels, wars, and labours and in death provides him with many commodities. Long before their domestication horses were hunted by primitive tribes for their flesh, and horsemeat is still consumed by man in parts of Europe and in Iceland and is the basis of many pet foods. Horse bones and cartilage are used to make glue. Tetanus antitoxin is obtained from the blood serum of horses previously inoculated with tetanus toxoid. From horsehide a number of articles are manufactured, including fine shoes and belts. The cordovan leather fabricated by the Moors in Córdoba, Spain, was originally made from horsehide. Stylish fur coats are made of the sleek coats of foals. Horsehair has wide use in upholstery, mattresses, stiff lining for coats and suits; high-quality horsehair, usually white, is employed for violin bows. Horse manure, which today provides the basis for cultivation of mushrooms, was used by the Scythians for fuel. Mare's milk was drunk by the Scythians, the Mongols, and the Arabs.

FORM AND FUNCTION

A mature male horse is called a stallion, the female a mare. A stallion used for breeding is known as a stud. A castrated stallion is commonly called a gelding. In former times stallions were employed as riding horses, while mares were kept for breeding purposes only. Geldings were used for work and as ladies' riding horses. In recent years, however, geldings generally have replaced stallions as riding horses. Young horses are known as foals; male foals are called colts and females fillies.

Anatomical adaptations. The primitive horse probably stood 12 hands tall (about 120 centimetres, or 48 inches) at the withers, the high point on the back at the base of the neck, and was dun coloured. Domestic horses gone wild, such as the mustangs of western North America, tend to revert to those primitive features under random mating; they generally are somewhat taller (about 15 hands), usually gray, dun, or brownish in colour, and move in herds led by a stallion.

The horse's general form is characteristic of an animal of speed: the long leg bones pivot on pulley-like joints that restrict movement to the fore and aft, the limbs are levered to muscle masses in such a way as to provide the most efficient use of energy, and the compact body is supported permanently on the tips of the toes, allowing fuller extension of the limbs in running (see illustration).



(Top) Skeleton of the horse; (bottom) points of the horse.

The rounded skull houses a large and complex brain, well developed in those areas that direct muscle coordination. While the horse is intelligent among subhuman animals, it is safe to say that the horse is more concerned with the functioning of its acute sensory reception and its musculature than with mental processes. Though much has been written about "educated" horses that appear to exhibit an ability to spell and count, it is generally agreed that in such cases a very perceptive animal is responding to cues from its master. But this ability is remarkable enough in its own right—for the cues are often given unconsciously by the human trainer, and detection of such subtle signals requires extremely sharp perception.

The horse, like other herbivores, has typical adaptations for plant eating: a set of strong, high-crowned teeth, suited to grinding grasses and other harsh vegetation, and a relatively long digestive tract, most of which is intestine concerned with digesting cellulose matter from vegetation. Young horses have milk (or baby) teeth, which they begin to shed at about age two and a half. The permanent teeth, numbering 36 to 40, are completely developed by age four to five years. In the stallion these teeth are arranged as follows on the upper and lower jaws: 12 incisors that cut and pull at grasses; 4 canines, remnants without function in the modern horse and usually not found in mares; 12 premolars and 12 molars, high prisms that keep moving out of the jaw to replace the surfaces worn off in grinding food.

Diversification into three types

Under domestication the horse has diversified into three major types, classified on the basis of size and build: draft horses, heavy limbed and up to 20 hands high (200 centimetres or 80 inches); ponies, by convention horses under 14.2 hands high (about 144 centimetres or 57 inches); and light horses—the saddle or riding horses that fall in the intermediate size range. Domestic horses tend to be nearsighted, less hardy than their ancestors, and often high strung, especially among thoroughbreds, in which intensive breeding has been focussed upon speed to the exclusion of other qualities. The stomach is relatively small, and, since much vegetation must be ingested to maintain vital processes, foraging is almost constant under natural conditions. Domestic animals are fed several (at least three) times a day in quantities governed by the exertion of the horse.

Senses. The extremely large eyes placed far back on the elongated head admirably suit the horse for its chief mode of defense: flight. And its long neck and high-set eyes, which register a much wider range than man's, enable it to discern a possible threat even while eating low grasses. The horse's vision is binocular, like man's, but only in the narrow area directly forward, and evidence suggests that it does not register colour. While visual acuity is high, the eyes do not have variable focus, and objects at different distances register only on different areas of the retina, which requires tilting movements of the head. The senses of smell and hearing seem to be keener than man's. As the biologist George Gaylord Simpson has put it in *Horses* (1961): "Legs for running and eyes for warning have enabled horses to survive through the ages, although subject to constant attack by flesh eaters that liked nothing better than horse for supper."

Colour and pattern. From the dun of the primitive horse have sprung a variety of colours and patterns, some highly variable and difficult to distinguish. Among the most important colours are black, bay, chestnut (and sorrel), palomino, cream, and white.

The black colour is a true black, although a white face marking (blaze) and white ankles (stockings) may occur. The brown horse is almost black but has lighter areas around the muzzle, eyes, and legs. Bay refers to several shades of brown, from red brown and tan to sandy. Bay horses have a black mane, tail, and (usually) stockings. Chestnut is similar to bay but with none of the bay's black overtones. Lighter shades of chestnut are called sorrel. The palomino horse runs from cream to bronze, with a flaxen or silvery mane and tail. The cream is a diluted sorrel, or very pale yellow, nearly white. White in horses is variable, ranging from aging grays (see below) to albinos with blue eyes and pink skin and to pseudoalbinos with a buff mane or with brown eyes. The chief patterns of the white horse are gray, roan, pinto, and appaloosa. Gray horses are born dark brown or black and develop white hairs as they age, becoming almost all white in advanced years. Roan refers to white mixed with other colours at birth: blue roan is white mixed with black; red roan is mixed white and bay; and strawberry roan is white and chestnut. The pinto is almost any spotted pattern of white and another colour; other names, such as paint, calico, piebald, skewbald, overo, and tobiano, refer to subtle distinctions in type of colour or pattern. Appaloosa is another extremely variable pattern, but the term generally refers to a large white patch over the hips and loin, with scattered irregular dark spots.

Nutrition. The horse's food, which in general consists of hay and grain, is usually not given immediately before or after work. Fresh water is important, especially when the horse is shedding its winter coat, but the animal is never watered when it is overheated after working. Oats provide the greatest nutritional value and are given especially to foals. Older horses, whose teeth are worn down, or those with digestive troubles, are provided with crushed oats. Chaff (minced straw) is added to the oat ration of animals that eat greedily or do not chew the grain properly. Crushed barley is sometimes substituted in part for oats. Hay provides the bulk of the horse's ration and may be of varying composition according to locale. Mash is bran mixed with water and with various invigorating additions or medications. It is given to horses with digestive troubles or deficient eating habits. Maize, or corn, is used as a fattening cereal, but it makes the horse sweat easily. Salt is needed by the horse at all times and especially when shedding. Bread, carrots, and sugar are tidbits used to reward an animal by the rider or

trainer. In times of poverty horses have adapted to all sorts of food—potatoes, beans, green leaves, and in Iceland even fish—but such foods are not generally taken if other fare is available. A number of industrial products are available to modern breeders and owners that contain balanced additions of minerals, vitamins, and other chemical elements, together with the customary natural foods.

Behaviour. The horse's nervous system is highly developed and gives proof to varying degrees of the essential faculties that are at the basis of intelligence: instinct, memory, judgment. Foals, which stand on their feet only a short while after birth, and are able to follow their mothers within a few hours, even at this early stage in life exhibit the traits generally ascribed to horses. They have a tendency to flee to evade danger. They express fear sometimes by showing panic, sometimes by immobility. Horses attack rarely and do so either when flight is impossible or when driven to assault a person who has treated them brutally.

Habit and instinct

Habit governs a large number of their reactions. Instinct, together with a fine sense of smell and hearing, enables them to sense water, fire, even distant danger. An extremely well-developed sense of direction permits the horse to find its way back to its stables even at night or after a prolonged absence. The visual memory of the horse prompts it to shy repeatedly from an object or place where it had earlier experienced fear. The animal's auditive memory, which reminded ancient army horses or hunters to follow the sounds of the bugles, is used in training. When teaching, the instructor always uses the same words and the same tone of voice for a given desired reaction. Intelligent horses soon attach certain movements desired by their trainers to particular sounds and even try to anticipate their rider's wishes.

While instinct is an unconscious reaction more or less present in all individuals of the same species, the degree of its expression varies according to the individual and its development. Most horses can sense a rider's uncertainness, nervousness, or fear, and are thereby encouraged to disregard the rider or even deliberately disobey him. Highbred animals, which give evidence of greater intelligence than those of low breeding, are capable not only of acts of vengeance and jealousy against their riders but also of expressions of confidence, obedience, affection, and fidelity. They are less willing than a lowbred horse to suffer rough handling or unjust treatment.

Cunning animals have been known to employ their intelligence and physical skill to a determined end, such as opening the latch of their stalls or the lid of a chest of oats.

Reproduction and development. The onset of adult sex characteristics generally begins at the age of 16 to 18 months; and the horse is considered mature, according to the breed, at approximately three years and adult at five. Fecundity varies according to the breed and may last to beyond age 20 with thoroughbreds and to 12 or 15 with other horses. The gestation period is 11 months, 280 days being the minimum in which the foal can be born with expectation to live. As a rule a mare produces one foal per mating, twins occasionally, and triplets rarely. The foal is weaned at six months.

The useful life of a horse varies according to the amount of work it is required to do and the maintenance furnished by its owner. A horse that is trained carefully and slowly and is given time necessary for development may be expected to serve to an older age than a horse that is rushed in its training. Racehorses that enter into races at the age of two rarely remain on the turf beyond eight. Well-kept riding horses, on the contrary, may be used up to 20 years and beyond.

Life expectancy

The life-span of a horse is calculated at six to seven times the time necessary for his physical and mental development; that is, at 30 to 35 years at the utmost, the rule being rather about 20 years. Ponies generally live longer than horses. There are a number of examples of horses that have passed the usual limit of age. The veterinary university of Vienna conserves the skeleton of a thoroughbred mare of 44 years of age. Unauthenticated reports have been made of horses living to the age of 40 and 50.

Diseases and parasites. Horses are subjected to a number of contagious diseases, such as influenza, strangles, glanders, ringworm, and swamp fever. Their skin is affected by parasites, including certain mites, ticks, and lice. Those with sensitive skin are especially subject to eczemas and abscesses, which may result from neglect or contamination. Sores caused by injuries to the skin from ill-fitting or unclean saddles and bridles are common ailments. The horse's digestive tract is particularly sensitive to spoiled feed, which causes acute or chronic indigestion, especially in hot weather. Worms can develop in the intestine and include the larvae of the botfly, the pinworms, tapeworms, and roundworms (ascarids). Overwork and neglect may predispose the horse to pneumonia and rheumatism. The ailment known as roaring is an infection of the larynx that makes the horse inhale noisily; a milder form causes the horse to whistle. Chronic asthma, or "broken wind," is an ailment that is considered to be all but incurable. A horse's legs and feet are sensitive to blows, sprains, and overwork, especially if the horse is young or is worked on hard surfaces. Lameness may be caused by bony growths, such as splints, spavins, and ringbones, by soft-tissue enlargements, known as windgalls, thoroughpins, and shoe boils, and by injury to the hooves, including sand crack, split hoof, tread thrush, and acute or chronic laminitis.

BREEDS OF HORSES

The first intensively domesticated horses developed in Central Asia. They were small, lightweight, and stocky. In time, two general groups of horses emerged: the southerly Arab-Barb types (from the Barbary coast) and the northerly, so-called cold-blooded types. When, where, and how these horses appeared is disputed. Nevertheless, all modern breeds—the light, fast, spirited breeds typified by the modern Arabian; the heavier, slower, and calmer working breeds typified by the Belgian; and the intermediate breeds typified by the thoroughbreds—may be classified according to where they originated (Percheron, Clydesdale, and Arabian), by the principal use of the horse (riding, draft, coach horse), and by their outward appearance and size (light, heavy, pony).

Light horses. *Arabian.* Known for its stamina, intelligence, and character, the Arabian breed is first mentioned in 400 BC and is said to have stemmed from five mares given to Muḥammad by his followers, but the Arabs doubtless had horses earlier. The Arab is a compact horse with a small head, protruding eyes, wide nostrils, marked withers, and a short back. It usually has only 23 vertebrae, while 24 is the usual number for other breeds. Its legs are strong with fine hooves. The coat, tail, and mane are of fine silky hair. While many colours are possible in the breed, gray prevails. The most famous stud farm is in the region of Najd, Saudi Arabia, but many fine Arabian horses are now bred in the United States.

Thoroughbred lineage

Thoroughbreds. The history of the English Thoroughbred is a long one. Records indicate that a stock of Arab and Barb horses was introduced into England as early as the 3rd century. Conditions of climate, soil, and water favoured development, and selective breeding was long encouraged by those interested in racing. Under the reigns of James I and Charles I, 43 mares (the Royal Mares) were imported into England, and a record, the General Stud Book, was begun in which only those horses are inscribed that may be traced back to the Royal Mares in direct line, or to only three other horses imported to England: the Byerly Turk (imported in 1689), the Darley Arabian (after 1700), and the Godolphin Barb (about 1730). The English Thoroughbred has since been introduced to most countries, where it is bred for racing or used to improve local breeds. The Thoroughbred has a small fine head, a deep chest, and a straight back. Its legs have short bones that allow a long easy stride, and its coat is generally bay or chestnut, rarely black or gray.

Asian. Asian breeds were strongly influenced by Arabian or Persian breeds, which together with the horses of

the steppes produced small, plain looking horses of great intelligence and endurance. Among them are the Tartar, Kirghis, Mongol, and Cossack horses. A Persian stallion and a Dutch mare produced the Orlov trotter in 1778, named after Count Orlov, the owner of the stud farm where the mating took place.

Anglo-Arab. The Anglo-Arab breed originated in France with a crossing of English Thoroughbreds to pure Arabians. The matings produced a horse larger than the Arab and smaller than the Thoroughbred, of easy maintenance, and capable of carrying considerable weight in the saddle. Its coat is generally chestnut or bay.

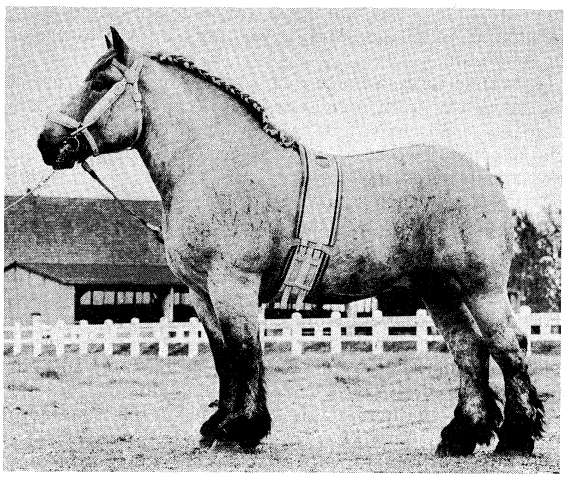
Standardbred and Saddle horses. From the English Thoroughbred stem the American Thoroughbred; the American Standardbreds, especially good at the pace and trot gaits; the American Saddle horses with three and five gaits. These are very fine animals with small heads and spectacular high-stepping movements. The three-gaited horses perform the walk, trot, and canter; the five-gaited horses in addition perform a rack and a slow gait. Since they are used mainly for shows, their hooves are kept rather long and the muscles of their tails are often clipped so that the base of the tail is carried high. Chestnut and bay are the usual coat colours.

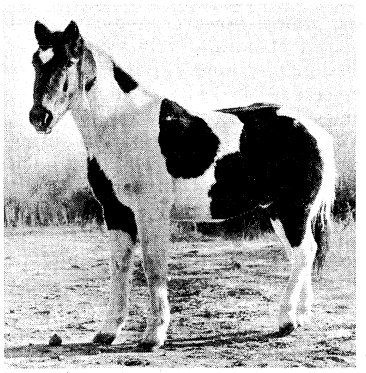
Also influenced by the Thoroughbred is the Tennessee Walking horse, used as a comfortable riding horse to cover great distances at a considerable speed. Its specialty is the walking pace, a very long and swift stride, which allows the horse to move at a rate of 8–10 miles (16 kilometres) per hour. Bay is the most common colour.

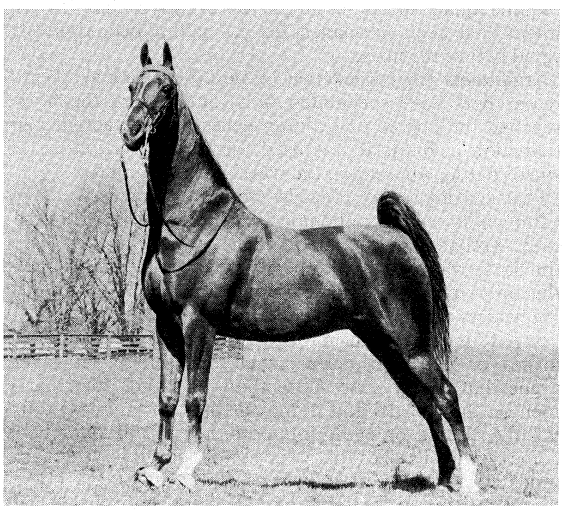
The Quarter Horse was bred for races of a quarter of a mile and is said to descend from Janus, a small Thoroughbred stallion imported into Virginia toward the end of the 18th century. It is 14.2 to 15 hands high, with sturdily muscled hind quarters, essential for the fast departure required in short races. It serves as a polo pony equally well as for ranch work.

The Morgan horse originated from a stallion given to one Justin Morgan around 1795. This breed has become a most versatile horse for riding, pulling carriages, farm labour, and cattle cutting. It was the ideal army charger. It is 15 hands high, robust, good-natured, willing, and intelligent. Its coat is dark brown or liver chestnut.

Sally Anne Thompson







Representative horse breeds.
(Top) The Belgian, a heavy breed; (centre) an Indian pony; (bottom) the American saddlebred, a light breed.

The Appaloosa

Appaloosa is a colour breed (see above) said to have descended in the Nez Percé Indian territory of North America from wild mustangs, which in turn descended from Spanish horses brought to the new world by explorers. The Appaloosa is 12 to 14 hands high, of sturdy build and of most diverse use; it is especially good in farm work. The English Hackney is a light carriage horse, influenced by the Thoroughbred, capable of covering distances of 12 to 15 miles (19 to 24 kilometres) per hour at the trot and canter. It measures 15.2 to 15.3 hands high and is appreciated for its high knee action. The English Hackney Bellfounder, imported to the United States in 1822 together with the Thoroughbred Messenger, contributed to the American Standard Breed.

The Cleveland Bay carriage horse, up to 17 hands high and generally bay in colour, is similar to the Yorkshire Coach horse. Both breeds are now used for the sport of driving.

Other versatile breeds include the German Holstein, Hanoverian, and East Prussian (Trakehner), which serve equally well for riding, light labour, and carriage. These horses, 16 to 18 hands high and of all colours, are now mostly bred for sport.

The Andalusian, a high-stepping, spirited horse, and the small but enduring Barb produced the Lipizzaner, named after the stud farm founded near Trieste, Italy, in 1580. Originally of all colours, the Lipizzaner is gray or now, exceptionally, bay. It is small, rarely over 15 hands high, of powerful build but slender legs, and long silky mane and tail. Intelligence and sweetness of disposition as well as gracefulness destined it for academic horsemanship, notably as practiced at the Spanish Riding School of Vienna.

Heavy breeds. The horses used for heavy loads and farm labour descended from the ancient war horses of the Middle Ages. The English Shire (the world's largest horse) and Clydesdale, the French Percheron, the Belgian horse, the German Noriker, and the Austrian Pinzgauer are now little used for their original purpose, being almost entirely replaced by the tractor. They measure well over 17 hands high, some 19 hands and more. They are of all colours, sometimes spotted, and generally have a very calm temperament.

Originating in the South Tyrol, the Haflinger is a mountain horse, enduring, robust, and versatile, used for all farm labour, carriage, sledge, and for pack hauling. It is rarely over 14.2 hands high and is chestnut with a flaxen mane and tail.

Ponies. Ponies are any horses other than Arabians that are shorter than 14.2 hands. They are generally very stur-

dy, intelligent, very energetic, and sometimes stubborn. The coat is of all colours, mainly dark, and the mane and tail are full. Ponies are used for pulling carriages and pack loads and as children's riding horses or pets. There are numerous varieties, including the Welsh, Dartmoor, Exmoor, New Forest, Highland, Dale, Fell, Shetland (under seven hands high), Iceland, and Norwegian. Ponies of the warmer countries include the Indian, Java, Manila, and Argentine.

HORSE SOCIETIES AND SHOWS

International standards

With the development of riding as a sport, organizations devoted to horses formed, and international standards were set up for various competitions. The chief governing body, the International Equestrian Federation, heads the various national federations that are subdivided into associations and clubs. It is the object of these organizations to promote and maintain interest in riding, to further the art of riding, to encourage instruction and progress of horsemanship, and to safeguard the welfare of the horse. The national federations organize horse shows and competitions, appoint judges, and delegate riders from their countries to international events. The selection of individual riders or a team for the Olympic, Pan-American, or regional games is done by the national federations.

Horse shows range from small local shows to international competitions and are held for many kinds of activity with a horse. The object of a horse show is to set a standard of presentation of horse and rider or trainer. The international competitions are open for dressage (deportment and obedience), jumping, and combined training. The International Federation acknowledges national events, held under the regulations of the national federation of the host country; friendly events, in which competition is open to riders from an invited nation; and official international events (Concours Hippique International Officiel) or Championships, which are authorized by both the national and the International Federations, are limited to once a year in Europe and twice a year in the rest of the world. At the 23rd Olympiad, 688 BC, a four-horse chariot race was on the program for the first time, and at the 33rd games, some 40 years later, mounted-horse races were included. The Olympic Games now include three equestrian events: dressage, stadium jumping, and combined training. Besides uncountable local horse shows, often in connection with agricultural fairs, a large number of international horse shows are held each year in Europe, the U.S., and Canada. Apart from the associations affiliated with the national federations, numerous organizations and clubs exist for racing, breeding, polo, hunting, or schooling.

BIBLIOGRAPHY. G.G. SIMPSON, *Horses* (1951; paperback edition, 1961), a very readable and popular account of the horse family today and through 60,000,000 years of development, with a good bibliography; M.C. SELF, *The Horseman's Encyclopedia*, rev. ed. (1963), an invaluable collection of information on domestic horses. Breed associations issue pamphlets on selected breeds of horses.

(A.W.P.)

Horse Racing

Although horse racing is one of the oldest of all sports, its basic concept has undergone virtually no change over the centuries. It has developed from a primitive contest of speed or stamina between two horses into a spectacle involving large fields of runners, colourful costumes, rigid protocol for contestants and spectators alike, sophisticated electronic equipment, and immense sums of money; but its essential feature has always been the same: the horse that finishes first is the winner.

This article covers the history of horse racing, including its beginnings and the development of the sport and of the Thoroughbred racing horse; racecourses; racing rules and procedures; and strategy. There are also short sections on steeplechasing, Quarter Horse racing, and harness racing and a section on wagering, including parimutuel and off-track betting (see also RIDING AND HORSEMANSHIP).

Horse racing remains one of the most popular sports and continues to grow in attendance and financial statistics, but, relatively speaking, it has had to surrender an increasing share of public attention to newer sports and other leisure activities.

It is difficult, now, to appreciate the pre-eminence once enjoyed by racing. Throughout the bulk of recorded history of many countries the horse was the primary means of transportation, a vital adjunct to agriculture, and the mainstay of armies. Even those who did not actually ride were thoroughly familiar with horses as an integral part of everyday living. The attention that today is lavished on automobiles, football teams, and other agents of sport formerly was concentrated on horses.

It was only natural that pride in one's own animal would engender the desire to match it against another and that community pride would be manifested by matching the champion local horse against that of a neighbouring village. And, it was only human to wager on the result.

In its rude form, a horse race required no special arena or time away from regular work to train teams of participants. And, apart from wagering, it entailed no special expense. Two horses and two riders provided all the ingredients necessary.

Today, horse racing survives in its rude form, and coexists as an enormous, highly technical industry.

History

ORIGINS AND EARLY HISTORY

The origin of horse racing cannot be determined with precision, and the subject does not lend itself to categorical treatment. Conclusions based on archaeological findings are subject to revision from time to time in the light of additional findings and developments in the techniques for interpreting them.

Evidences of racing. For many years the chariot race described in Homer's *Iliad* was regarded as the first recorded instance of the sport. Subsequent discovery of inscribed tablets in Asia Minor, however, indicates that Assyrian kings maintained elaborate stables under the supervision of "professional" trainers 1,500 or more years BC; *i.e.*, centuries before the Trojan War.

Evidence predating the *Iliad*

At the Olympic Games in Greece, there was racing among chariots drawn by teams of four horses during the 7th century BC, prior to which the two-horse hitch had prevailed. It was not until about 40 years after the appearance of the four-horse hitch that racing among mounted horses (ridden bareback) was introduced, and for years this was regarded as having marked the approximate origin of horseback riding. Other evidence, however, indicates that man was riding horseback in Egypt about 1345 BC and in Babylonia during the 3rd millennium BC, although in the latter case the claim is submitted only to the extent that the animal ridden was an "equine quadruped," with the observation that it well might have been an onager, the ass having been domesticated before the horse.

Arguments for precedence of mounted racing. Some researchers have speculated that long before the horse reached the degree of training necessary for actual participation in battle it was used for transportation to the site of battle, whereupon the warrior dismounted.

That chariot racing preceded mounted horse racing by as much as several millennia is disputed by some horsemen. Archaeologists contend that for thousands of years the horse was too small to bear man's weight on his back; the only truly wild horse known in the 20th century, Przewalski's horse (*Equus caballus przewalskii*), is a small pony, found in western Mongolia, about 12 hands (48 inches or 120 centimetres) high. Horsemen, however, argue that donkeys no more than 12 hands high carry men on their backs and point out that, in any case, during the course of evolution both horse and man have come in various sizes that conceivably could have accommodated a variety of matchups. The gist of the horsemen's contention is simply that it is not natural that riding in wheeled vehicles should have preceded riding astride by so long a period, if at all. At whatever time the

horse was domesticated (after the ox and the ass), it would have been virtually instinctive on the part of man to mount and ride it. The chariot itself, with its relatively complicated harness attachments, would not have been necessary. A braided vine or rawhide thong (evidence of which is not likely to be found in excavations nor easily interpreted if found) would have been all the equipment needed. Archaeological evidence by and large has been found on the sites of what formerly were relatively stable, civilized areas, horsemen point out, whereas nomadic tribes, which were the least likely to have left traces of their activity, are the most likely to have originated mounted horsemanship.

Spread of racing stock. While many of the horses that performed at the Olympic Games probably were bred in Greece, many others came from Asia Minor or North Africa or were close descendants of horses imported from those areas. Arabians, Barbs, and Turks, since time immemorial, were the most greatly prized equine stock for purposes of speed. As the nucleus of culture and power shifted from Greece to Rome, the horses went along, and from Rome their influence was disseminated throughout Europe. Much later, from Spain, some of this same blood crossed the Atlantic to the Americas.

Horses from Asia Minor and North Africa

BEGINNINGS OF ORGANIZED RACING

Origins. As is true of impromptu racing, there is no specific date to mark the beginning of organized racing. It developed at different times in different countries. Presumably, areas such as China, Persia, Arabia, other countries of the Near East, and North Africa, where horsemanship was highly developed, were among the first to have organized racing.

Greece and Rome. Modern racing, in the sense that the general public became part of the proceedings, was foreshadowed in ancient Greece and Rome. During the heyday of the Roman Empire, there were chariot races, mounted races, and so-called Roman races in which riders stood with one foot on each of two horses. Public registries were kept for recording bloodlines of horses as well as racing records; and exceptional horses were accorded the honour of burial with a stele citing their records—one such having indicated 1,300 victories, 88 second- and 37 third-place finishes. There was ample opportunity to race: from 12 races a day under Augustus (27 BC–AD 14), the programs increased to 100 under the Flavians (AD 69–96), morning until sundown, with the distances of the races having been reduced to accommodate the increased numbers. There were professional racing officials, starting chutes, disputes at law, accusations of doping, widespread gambling (off-track as well as on, judging from the carrier pigeons that were released after certain races), and riots among the spectators.

Great Britain. Racing of the variety that came to be known as "the sport of kings" originated in England, although the fact that, long after the Roman invasion, Englishmen continued to import horses from the Continent, the Near East, and North Africa suggests that a similar form of sport was being carried on in other countries. An Arabian filly described as a racehorse reportedly was sold for the equivalent of £67,200 in the year 1290.

For centuries the avowed purpose of improving the breed of horses was to upgrade the cavalry; but in England, as the war horse gravitated toward the massive charger capable of carrying a man in heavy armour, a distinct type of horse, lighter and faster and selected for the specific purpose of sport, emerged. A description of London, about 1174, mentions racing at Smithfield every Friday; horses for sale were ridden in competition by professional riders to display their speed to buyers.

Like many other participants in the Crusades, Richard the Lion-Heart had acquired a great admiration for the Arabian horse, and it was during his reign (1189–99) that the first known racing purse was offered, 40 pounds in ready gold, run for over a three-mile (4.8-kilometre) course, with knights as riders. Richard's successor, John, was the first English king reported to have "running horses" in his stables. The Lanark (Scotland) Silver Bell has been run for since the late 12th century; and at some time prior to 1500 an annual "meeting" (one race each year) was instituted at Chester, as the report of the 1512 race stated that the event had been contested "time out of mind."

The patronage of English kings

Henry VIII in the 16th century imported "coursers" from Italy and Spain (which suggests that the horses were Barbs) and established royal studs in several locations. James I, who previously had patronized racing in Scotland, sponsored meetings in England during his reign (1603–25). Charles I, who reigned 1625–49, had a stud of 139 horses at the time of his death; and Charles II, who reigned 1660–85 and is known as "the father of the British turf," actively participated in races as a rider and offered King's Plates (silver cups or other prizes) to be run for. His patronage established Newmarket as the headquarters of racing.

North America. An emissary of Charles II, Richard Nicolls, earned the cognomen "father of the American turf." The first governor of the colony of New York, in 1665 he offered a silver cup, to be run for each spring and fall over a course laid out on Hempstead Plain, Long Island. This was the first known race trophy in North America and the beginning of so-called course racing. There had been previous racing in other American colonies, notably Virginia, but it was of the informal frontier variety.

France. There is documentation of racing, of a type associated with rural festivals, in France during the 14th century; a formal race, resulting from a bet between Prince d'Harcourt and another nobleman, was reported in 1651. During the long reign of Louis XIV (1643–1715), racing on which gambling was based was prevalent. Louis XVI (reigned 1774–93) organized a jockey club; and decrees issued by the royal printing house included rules of racing, among which were specifications concerning certificates of origin and extra weight to be carried by "foreign" horses—an early example of discrimination in favour of native stock. Napoleon established a series of regional competitions for money prizes, with a runoff for a "Grand Prix" at Paris.

Match racing. As noted, racing by fields of horses for prizes dates back to ancient times, but as the sport began to re-evolve into its present format, during the 16th century, matches were the most common form of competition. In a match, the owners of the competing horses themselves provided the prize (purse).

A typical early match was simply an agreement between two owners to run their horses against one another for, say, "200 guineas each, half forfeit," which represented a wager, pure and simple, with a penalty provided should either owner withdraw his horse. "Official" records of such agreements (some of which involved sums considerably larger than 200 guineas) were entered in books by unbiased third parties, and specialists at this function emerged in various racing centres as "keepers of the match book." In 1727, John Cheny, keeper of the match book at Newmarket, published *Cheny's Horse Matches*, which was a consolidation of the match books of various racing centres. The work was carried on as an annual volume by a succession of publishers under different titles—some volumes carried results of cock fighting and cricket matches as well as of horse racing—until it was taken over in 1773 by James Weatherby. As the *Racing Calendar*, devoted to horse racing only, it has been published by the Weatherbys ever since.

Origins of racing calendars

Although three or more horses could participate in early matches, the usual number was two; and in modern times the term match race applies only to contests between two horses.

A defection from a two-horse race, of course, resulted in cancellation of the contest. To discourage casual forfeiture, such terms as one-fourth forfeit were replaced in some races by the phrase play or pay, which meant that an owner whose horse failed to fulfill the engagement forfeited his entire stake. A similar change took place in the rules of betting, which originally provided that a wager on a horse that did not actually start should be refunded but later stipulated "play or pay," which meant that the bet was firm whether or not the horse ran.

Open field racing. Because of their essentially private nature, match races progressively lost favour to more open events, involving larger fields of runners. The races for the King's Plates, which dated back to the time of Charles II, were especially popular. Townships and individual sportsmen also offered prizes to be run for by any horses that met certain qualifications; and general sweepstakes, to which a number of owners could nominate horses at a fee, came into vogue.

There developed a wide spectrum of eligibility conditions for races, based on age, sex, birthplace, qualification of riders, previous usage, and previous accomplishment or (as was more common) lack thereof. Some of the conditions, *e.g.*, those based simply on age and sex, were designed to attract larger fields of runners. Other conditions were designed to broaden the base of the sport by encouraging participation of those who might have eschewed open competition or to provide for more interesting contests among horses of essentially similar class, or both. Examples of such conditions were "owners to ride," or "for Horses, etc., bred in the county of Glamorgan . . .," or "for Horses, etc.; that never won Plate or Match of 50 pounds. . . ."

Eligibility conditions and racing rules

The articles for the King's Plates in England were among the earliest racing rules of national scope. George II's Act of 1740 was relatively simple in that it provided for only one type of race but remarkably comprehensive in the contingencies it covered. Horses were required to be the bona fide property of persons who entered them (a precaution against "ringers" [*i.e.*, horses entered into competition with false identification] of superior ability); certificates as to ownership and age were required; penalties were specified for rough riding; and riders were not allowed to dismount before arriving at the place of weighing-in.

First records of jockeys. Although jockeys were identified in contemporary accounts of certain important races, they were ignored in the early official records. No doubt this was in part a reflection of the succinct format of early records, but it also reflected prevailing opinion that the jockey was not an especially important factor in determining the outcome of a race. Certainly this was true of four-mile (6.4-kilometre) heat racing, in which the strength and stamina of the horse were more important than quick thinking and judgment by the rider. If a horse became boxed in during a four-mile heat, there was plenty of time in which to extricate him; and even should the heat be lost, there were other heats in which to make amends. In dash racing, by contrast, depending on the distance, the loss of a few yards could mean the race. Jockeys more or less crept into the official records. First, the names of the winning trainer and winning rider only were added to race reports. By the late 1850s it had become customary for the *Racing Calendar* to record all riders.

One of the first jockeys to appear in official English records was for many years identified simply as "Nat," until finally, in 1841, the *Racing Calendar* recorded his full name: Elnathan Flatman. A famous jockey in early American records was a slave listed only as "Abe."

Development of the Thoroughbred. The widely circulated version that the Thoroughbred breed was founded by three "Oriental" stallions and 43 "royal" English mares is misleading. The word founded is dubious to the extent that it suggests establishment according to some preconceived plan. The Thoroughbred breed simply evolved. In the beginning, it was the horses that made the pedigrees rather than vice versa; and it was not until many generations later that certain horses were determined—in retrospect—to have been founders. The breed emerged from considerably more than 43 mares, not all of which were "royal" (a designation given to a group of mares imported by Charles II and applied also to their descendants), and considerably more than three stallions, not all of which were "Oriental."

Emergence of the breed

The Thoroughbred did begin as a mixture of Arabs, Turks, and Barbs with native English stock; although it might have had remote Arabian antecedents, the Barb was generally acknowledged as a distinct breed indigenous to Algeria and Morocco rather than to the Orient. Barbs were especially popular in England and more easily obtainable, through Spain and Italy, than the other breeds. Since a stallion has so many more offspring than a mare, this foreign blood was more widely disseminated by sires; but there also were elements of it, resulting from previous importations, among the "native" English broodmare population.

Bloodlines and studbooks. As the English sporting horse emerged as a distinct type, the term bred horse came into being, and it became customary to include pedigrees of horses in papers of sale. There had been private studbooks, one dating back to 1605, but the records were not invariably reliable.

In 1791, Weatherby, publisher of the *Racing Calendar* mentioned above, offered a small collection of pedigrees he had extracted from *Racing Calendars* and sale papers as *An Introduction to a General Stud Book*. It was revised several times over a period of many years, increasing in scope with the revisions, and it became Volume I of the *General Stud Book*. This volume purported to be nothing more than a collection of pedigrees of horses (with a view toward correcting "false and inaccurate" pedigrees but with no claim that it was free of error). Horses were included "from the earliest records," which in some cases dated back a century, suggesting that Weatherby relied at least to some extent on private studbooks for his information.

The term bred horse, or blooded horse, also was used in other countries, and although it was not formally defined, horsemen evidently had an understanding as to its meaning. Conditions for a Pennsylvania race in 1764 specified more entrance money for "full-blooded" horses than for three-quarters blooded, and half-breds received a further discount. In colonial North America, where the progress of racing and breeding followed that of the mother country so closely as to be virtually contemporaneous, the degree of breeding almost certainly referred to the English pattern; Bulle Rock, the first documented "bred" horse in America, was imported from England in 1730.

Studbooks were established in other countries, including in some cases native stock but relying heavily in all cases on horses of "English" breeding. Volume I of the *Stud Book Français*, begun in 1838 but, like the British book, revised over a period of years before it became firm, had two classifications of "purebred" horses: "Orientale," which included pure Arabs, pure Turks, and pure Barbs; and "Anglais," which included mixtures according to the English pattern. The latter category, by far the larger, was determined according to type of breeding rather than place of birth; and most of the horses were native to France.

The mixed breed was destined to predominate, and the modern *Stud Book Français* is presented simply as a register of "chevaux de pur sang Anglais" ("horses of pure English blood").

Because the top (tail-male, through male issue) and bottom (tail-female, through female issue) lines are the longest lines in a pedigree, horsemen pay particular attention to those lines, especially the top one since, as already noted, the influence of an individual stallion is far more widespread than that of a mare.

Male and female pedigree lines

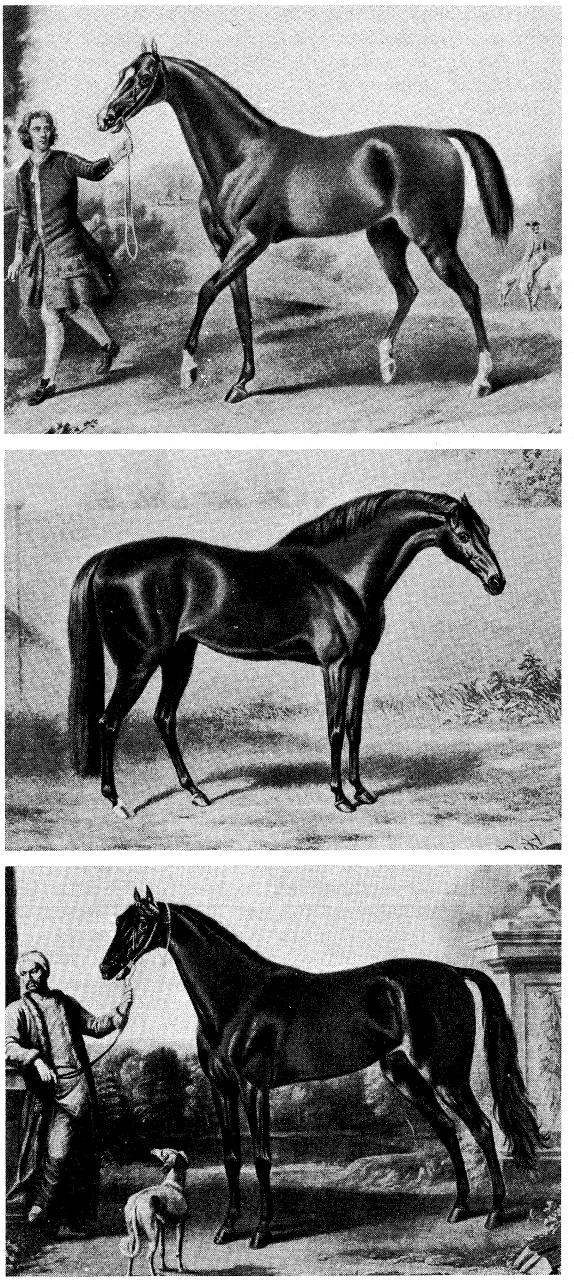
By sheer coincidence it developed that every horse registered in the studbooks of all countries traced in direct male line to one of three stallions, viz:

The Byerly Turk, used as a charger by the English Captain Byerly (also spelled Byerley) at the Battle of the Boyne in 1690.

The Darley Arabian, imported as a young horse, about 1704, from Aleppo to England by Richard Darley of Yorkshire.

The Godolphin Barb (also called Arabian) brought to England from France about 1730 by Edward Coke, who, according to a popular account, discovered the horse pulling a cart through the streets of Paris. After Coke's death, the horse passed into the ownerhip of the Earl of Godolphin.

By another coincidence, the male lines of each of these three sires survives through a single descendant. All tail-male descendants of the Godolphin Barb trace to him



Three foremost stallions.
(Top) The Darley Arabian. (Centre) The Godolphin Barb. (Bottom) The Byerley Turk. Every horse registered in the stud books of all countries can be traced to one of these stallions in direct male line.
Fores Ltd.

through his grandson, Matchem (1748–81); those of the Byerly Turk through his great-great-grandson, (King) Herod (1758–80); and those of the Darley Arabian through his great-great-grandson, Eclipse (1764–89). Actually, the Matchem line could be identified by other names since there probably is not a modern horse tracing to him that does not do so through West Australian, six generations removed from Matchem. The male-line descendants of Eclipse comprise the vast majority of the Thoroughbred breed, and his most prolific branch is that of an oddly named son, Potooooooooo. The name supposedly resulted from the inability of a stable lad to spell potatoes, but this sire is listed in modern pedigrees as Pot-8-0's. Eclipse was bred by the Duke of Cumberland, who at one time owned Herod.

In tracing pedigrees only through tail-male or tail-female lines, connections between stallions and their daughters or mares and their sons are disregarded. While all Thoroughbreds today descend from one of the three so-called foundation sires, in tail-male line, they also descend from other sires in other lines of their pedigrees. For example, Herod, Matchem, and Eclipse, in other than their top lines, trace to D'Arcy's Yellow Turk and to the Leedes Arabian, whose tail-male lines have died out.

When the Australian Bruce Lowe and the German Hermann Goos (independently of each other) traced every mare in tail-female line back to her so-called taproot as registered in volume I of the *General Stud Book*, they found only 43 female "families" that still survived in unbroken chains of dam-to-daughter at the time of their research, in the latter part of the 19th century. No account was taken of taproot mares whose bottom lines had died out but who nevertheless had descendants through lines that at some point(s) traced through a son. Moreover, Lowe's family numbering system ignored female families in other countries, whose taproots were not recorded in the first volume of the *General Stud Book*.

For many years there had been reciprocity among the studbooks of various countries, but in 1913 the English Jockey Club passed a measure, sponsored by the Earl of Jersey, that declared that only horses that traced in all their lines to animals previously registered in the *General Stud Book* would be accepted for future registrations therein. The effect of the Jersey Act, as it was called, was to disqualify as Thoroughbreds many horses bred outside of England or Ireland, including the majority of North American horses. The ostensible purpose of the Act was to protect the British Thoroughbred from infusions of American "sprinting" blood, but the background suggests more mundane motives. Political repression had caused a shutdown in 1911 and 1912 of racing in New York, which was the major American racing centre and bloodstock market. During the uneasy years that preceded the shutdown, American owners and jockeys had gone to England in increasing numbers; after the shutdown an invasion of American bloodstock into England was a threat. As just one example, James Ben Ali Haggin (whose private catalog for 1903 had included approximately 2,000 horses) had been in the habit of shipping trainloads of horses to the New York sales, and he already had sold horses in England, including Grand National winner Rubio for $75. By himself, Haggin could have flooded the English market, and there were other American breeders with horses to sell but no buyers at home.

Exclusion of American horses

Apart from its economic significance, for decades the Jersey Act rankled national pride and rather defied logic. Americus, for example, had been sent to England and duly registered in the *General Stud Book* before the Act. He was therefore a Thoroughbred, but both of his parents and numerous half-siblings, which had remained in America, as of 1913 suddenly were not. Americus (through a daughter) became the ancestor of such distinguished English Thoroughbreds as Mahmoud and Nasrullah.

The ticklish question finally was resolved in a proper setting, on the racetrack. Following a rash of victories in prestigious English races by French horses with "tainted" American ancestry, in 1949 the Jersey Act was rescinded. Animals with "eight or nine crosses of pure blood" became eligible for admission into the *General Stud Book*. This did not open the book automatically to all American Thoroughbreds, but it eliminated the ill feeling.

In 1954, the Derby and St. Leger were won by Never Say Die, a son of Nasrullah, mentioned above, out of a granddaughter of America's most famous horse, Man o'War, who, throughout his long life (1917–47) was not regarded as a Thoroughbred by English standards. In 1962 Never Say Die was champion English sire.

Jockey clubs and racing commissions. Provincial jockey clubs or similar organizations that regulated local turf affairs developed, along with racing, at various times in various places. The Jockey Club formed at Newmarket about 1750 wrote its own rules of racing; in contrast to the articles for King's Plates, which dealt with that form of racing only, these rules took into account different kinds of contests involving horses of different ages and

The Jockey Club rules and stewards

were correspondingly more detailed. These rules originally applied to Newmarket only but were printed in the *Racing Calendar* and served as models for rules everywhere. They gradually were adopted throughout the United Kingdom, and as the Jockey Club acquired ownership of the *Racing Calendar* and the *General Stud Book* (although both continued to be published by the Weatherbys) its authority became supreme. For a time it was policy to delegate authority to a Jockey Club steward of exceptional stature; those who held this position—Sir Charles Bunbury; Lord George Bentinck; and Admiral Henry John Rous, who died in 1877 and had no successor—were known as dictators of the English turf.

The Jockey Club in England continues to exercise virtually complete control of racing and breeding, and it also owns the Newmarket racecourse and the training grounds that surround it.

La Société d'Encouragement pour l'Amélioration des Races de Chevaux en France was founded before its English counterpart, in 1833. It administers racing, owns major tracks and training centres, including that at Chantilly, which is probably the finest in the world; but the *Stud Book Français* is published by the Ministry of Agriculture, and it was only in 1969, after the demise of *La Chronique du Turf*, that the society took over publication of that country's racing calendar.

In other countries the scope and authority of national jockey clubs vary; in North America there is a proliferation of organizations that bear that title. Some are simply commercial corporations that administer business operations of one track only. Control of actual racing in the United States is widely diffused, as rules are promulgated by individual commissions in each of the states that have racing. The commissions are organized into a National Association of State Racing Commissioners (which includes member bodies from Canada, Mexico, and elsewhere in North America), but each state retains its own authority.

The (North American) Jockey Club, headquartered in New York, at one time exercised wide (but not complete) control over American racing. A court ruling that took away its licensing power reduced its authority; but it still has considerable influence as keeper of the *American Stud Book* (which includes foals from Canada, Puerto Rico, and parts of Mexico), and it also publishes rules of racing that are used as a model by many states and have been formally adopted in at least one state.

There are numerous other racing organizations in many countries, as well as associations among owners, breeders, operators of race tracks, jockeys, trainers, veterinarians, and others closely identified with the sport.

TYPES OF RACES

Organized races. *Heat racing.* The original King's Plates were standardized races—all were for six-year-old horses, carrying 168 pounds (76 kilograms), at four-mile (6.4-kilometre) heats, a horse having to win two heats to be adjudged the winner. Beginning in 1751, five-year-olds (set to carry 140 pounds) and four-year-olds (126 pounds) were admitted to King's Plates, with a reduction in distance of the heats to two miles (3.2 kilometres). The desire to race younger horses had long since manifested itself in other kinds of races. Racing for four-year-olds was well established, and a race for three-year-olds, carrying 112 pounds (51 kilograms) in one three-mile (4.8-kilometre) heat, was run in 1731.

The Industrial Revolution was a significant factor in the demise of heat racing. In an agrarian economy a horse could be raised until it reached the age of four or five at no great additional expense, and if the horse turned out to be a failure on the racecourse it could be diverted to some other useful function. In an industrial economy such a horse represented virtually a total loss.

Heat racing persisted as the most popular form of the sport much longer in America, where Virginia was the horse capital, than it did in England. Heat racing, in which four-year-olds competed at the "full" distance of four miles, continued as a popular form until the U.S. Civil War (1861–65).

Dash racing. By that time heat racing had long since been overshadowed in Europe by dash racing, a "dash" being any race decided by only one heat, regardless of its distance. Some early dashes were at eight miles (12.8 kilometres), although the horses generally cantered until within a few furlongs of the finish.

The English classics

The modern age of racing is generally considered to have been marked by the inauguration of the English classic races: the St. Leger in 1776 (although it was not run under that name until two years later), the Oaks in 1779, and the Derby the next year. All were dashes for three-year-olds only. Sir Charles Bunbury, who was a strong advocate of dash racing for young horses, won the inaugural of the Derby with Diomed. Ironically, the horse later was exported to America where he founded a dynasty that dominated the golden age of four-mile-heat racing in that country.

Handicaps. *Weight handicaps.* During the 19th century, races of the English classic pattern—dashes for three-year-olds, at level weights except for the sex allowance (see below)—spread all over the world.

In the early match races the weight to be carried by each horse was a matter of agreement between the owners. In racing for fields of horses the weights generally were determined by age. Older horses had to concede weight to their juniors up to age seven, for example, at which point they were considered "aged" and the weights levelled off. (There were exceptional cases, such as the renowned American campaigner, Walk-in-the-Water, who raced on through age 18; at 15 and beyond, he received a concession from his juniors.)

There also were races in which weight was based upon height, but this proved to be an unsatisfactory criterion.

Scales of weights for age

As racing of younger horses became more prevalent, finer distinctions as to weight were necessary. A preliminary scale of weights for age, published by Admiral Rous in 1850, embodied sharply ascending weights for two-year-olds as the racing season progressed, April through November, while the weights for older horses were relatively stable throughout the year. Among the earliest acknowledgments that two-year-olds were regarded as racehorses in the U.S. was an 1843 rule that prescribed for them "a feather."

Gradually, scales of weights for age evolved into the elaborate modern tables that specify weights, month by month, according to age of the horses and distance of the races. Each country (and in the United States, each state) has its own scale, but the following examples are representative. A three-year old of the Northern Hemisphere, racing against a four-year-old at 6 furlongs (one furlong = ⅛ mile) receives a 15-pound (seven-kilogram) concession in January, but only a two-pound (one-kilogram) concession in November, when it has almost caught up in development. At a given time of year, say, November, the concessions increase from two pounds at 6 furlongs to four pounds (two kilograms) at 1½ miles or more. In general, the scales reckon a horse as being fully aged at five years, or earlier, so far as sprint racing is concerned.

The sex allowance provides that fillies receive a three- or five-pound (one- or two-kilogram) concession from males, and in many countries geldings also receive a weight concession, to encourage elimination of mediocre horses from stud duty.

Ages are reckoned from January 1 in the Northern Hemisphere and either July 1 or August 1 in the Southern. (May 1 formerly was the universal birthday in the Northern Hemisphere.)

The weight scale is designed to determine the intrinsically best horses, and all classic races are run under its provisions. Various devices were developed to neutralize the weight scale, however, in order to provide for more closely contested races or to attract larger fields (thereby stimulating wagering), or both.

Allowance and claiming races. "Condition" or "allowance" races provide for weight penalties or allowances based on previous performance. The "selling plate," or "claiming race," is essentially a method of requiring owners to classify their own horses by setting a price on

them. In early versions, the winning horse was sold at auction. In the modern version, any owner who has a horse in the race (or has started a horse at the meeting) may claim any other horse for the stated price (drawing lots if more than one owner claims the same animal). If any owner runs a horse below its true class, he stands to win the race but lose the horse.

Handicap races. The handicap race represents an outright repudiation of the classic concept that the best horse should win; imposts are assigned on the basis of past performance with the specific objective of giving all horses an equal chance of winning—although this objective is by no means invariably achieved. Racing form fluctuates, and assignment of weights is at best a difficult art, and at worst it can be a farce. In some countries, weights for all handicaps are assigned by a central authority, such as a representative of the jockey club, but in others each track employs its own handicapper and owners of star attractions thus are able to shop around for the most lenient weight assignments.

Assignment of handicap weights

In some countries, the purses of handicaps generally are not large enough to attract the classic winners or other top horses. In countries that offer handicap purses that are sufficient to attract the best horses, such races provide more emphatic tests of intrinsic merit than do classic events. Obviously, it requires a better horse to win against all comers while conceding weight than it does to win against members of only one foal crop at level weights. Unlike many classics, handicaps as a rule are not restricted as to age or sex.

The Melbourne Cup, a handicap inaugurated in 1861, is generally regarded as the most important race of the Southern Hemisphere. It has been won by such epic horses as Carbine, under 145 pounds, and Phar Lap, under 138. In America, the Metropolitan, Brooklyn, and Suburban handicaps—all of which date back to the 19th century—were at one time the most valuable events in the country, and remain reasonably comparable in value to classic events. Only three horses—Whisk Broom II, Tom Fool, and Kelso—have managed to win this "handicap triple," compared to nine winners of the American Triple Crown for three-year-olds. As it happens, not one of the three handicap triple winners ran in any one of the three-year-old classics.

The pioneer among existing races of $100,000 or more in value is the Santa Anita Handicap, which for its first running, in 1935, attracted horses from thousands of miles away. The field included four classic winners of former years (and still another had been withdrawn shortly before the race), but the winner was an Irish-bred former steeplechaser, Azucar.

As is to be expected, handicaps are exceptionally popular as betting media.

Phar Lap, who was bred in New Zealand, won several classic races in Australia, and Kelso's failure to participate in the American classics for three-year-olds was the result of an interruption in his training. As geldings, however, had they raced in Europe (or in any of several countries elsewhere), neither Phar Lap, Kelso, nor Azucar would have been allowed to run in classic races. The restriction against geldings in certain important races in New York was lifted as recently as 1957; and Kelso, the leading money earner in history ($1,977,896), derived a considerable measure of his fame and income from five successive victories in the Jockey Club Gold Cup, a classically conceived weight-for-age event for which he would not have been eligible a decade earlier.

Races for older horses. That three years is the classic age of lucrative purse opportunities is based more on expediency than on sporting considerations. Statistics tend to confirm the implicit message of weights, that a racehorse achieves peak ability at age five. There has been a modern trend toward giving horses older than three more opportunities to race for larger purses. The Prix de l'Arc de Triomphe, in France, is generally acknowledged as the world championship event not only because of its enormous value (more than $200,000 to the winner) but also because it admits older horses. It is not a handicap, as weights are based on age and sex only, but they are adjusted from time to time to maintain balance between frequency of victory and frequency of participation by horses of various age groups. The Arc is not open to geldings, however.

A partial listing of other famous races that admit horses older than three would include, in addition to events already mentioned, the Gran Premio Internacional Carlos Pellegrini in Argentina, Caulfield and Sydney cups in Australia, Grande Premio São Paulo Internacional in Brazil, King George VI and Queen Elizabeth Stakes in England, Gran Premio del Jockey Club and Gran Premio di Milano in Italy, Emperor's Cup and Arima Memorial Stakes in Japan, Wellington Cup in New Zealand, Rothman's July Handicap in South Africa, and Gran Premio Clasico Simón Bolívar in Venezuela. Of the approximately 60 races in the United States with purse values of $100,000 or more, about half are open to horses older than three years.

Purse money and stakes fees. The King George VI and Queen Elizabeth Stakes and the Rothman's July Handicap are examples of sponsored races for which much of the purse money is put up by commercial firms. In the United States, this practice has not caught on, and most of the purse money for the richest events is provided by the owners themselves, in the form of stakes fees. The $352,320 purse of the 1970 Garden State Stakes (which is an event for two-year-olds) included $227,320 in owners' stakes fees.

Distribution of purses

Methods of distributing purses vary widely. As winner-take-all matches were replaced by racing among fields of horses, the provision "second to save his stake" was inserted into conditions for some races. Then, a second prize of stated value, greater than the owner's stake, was offered (in early races, hogsheads of claret and, in America, casks of tobacco were frequent ingredients of race prizes). Gradually, in further efforts to encourage larger fields, third and fourth prizes became common. Fifth money is not unusual today, and in some races at least a token portion of the purse is awarded to every horse that starts.

As an overall average, North American purses, for example, in 1970 were allocated 61 percent to the winner, 20 percent to second, 11 percent to third, 6 percent to fourth, 2 percent to fifth, and an inconsequential percentage to horses out of the first five. Slightly more than half (54.5 percent) of the horses that started in North America in 1970 won at least one race, and as a group these horses also gathered in the lion's share of minor purse awards—when they didn't win, they usually finished close up—with the result that they earned 94.1 percent of the total purse money.

THE HORSES AND THE HORSEMEN

Breeding theory and practice. Over the centuries the guiding principle for breeding Thoroughbreds has been, as expressed by an old cliché: breed the best to the best and hope for the best. Performance of progeny is the most reliable guide to what is best for breeding purposes, of course, but in the case of horses untried at stud, their own racing ability, pedigree, and physical conformation are the only available yardsticks. Emphasis is on racing ability, especially in evaluating potential stallions. Although for many years the stallion was regarded as by far the dominant partner in an individual mating, the modern school of thought credits the mare for approximately half the inherited quality of the offspring.

Breeding systems

When Lowe numbered the female families, he did so in order of the frequency of victories in the English Derby, St. Leger, and Oaks, members of the number one family having scored the most victories, and so forth. Families that had produced no winners of the designated races were assigned the higher numbers, according to Lowe's estimate of their worth. He also designated certain tail-female lines as "running families" or "sire families." As time went on, the rankings in number of classic victories were subject to change, and new families were introduced; but the original family numbers were retained. Matching bloodlines according to Lowe numbers was at one time practiced as a breeding system, but today the

numbers are generally regarded simply as convenient identifications of female lines.

Lieutenant Colonel J. Vuillier, a French cavalry officer, propounded in the 1920s what he called a "dosage system," based on the theory that each ancestor provided a fractional portion of the makeup of a horse. Having noted that 15 stallions and one mare (Pocahontas) appeared most frequently in the pedigrees of successful racehorses, he advocated matings that provided for optimum dosages of the blood of each. The system continues to attract some followers; and the late Aga Khan III, who enjoyed enormous success in racing and breeding, was said to have employed a modification of the Vuillier system.

The most successful breeders appear to be those who apply the best-to-the-best formula, with common sense and judgment added—instead of mating just any top class stallion to any top class mare, they take into account the individual characteristics of each animal.

To the phenomenally successful Italian breeder, Federico Tesio, whose products included undefeated champions Nearco and Ribot (he named many of his horses after artists), breeding horses appears to have been as much an art as a science. Nor did he discount romance, having attributed the merit of Derby and Oaks winner Signorinetta to the strong attraction her parents had manifested for each other.

Outstanding horses. Eclipse, from whom the majority of Thoroughbreds descend in tail-male, antedated the classics and did not begin racing until he was five years old. He was undefeated in 18 races, including a number of walkovers in which no rivals appeared to compete with him.

St. Simon (1881) did not run in any of the three-year-old classics, and he, too, was undefeated—in nine races, including the Ascot Gold Cup. At stud he was extraordinarily prolific and successful. He covered mares through age 27 and got an average of 25 foals per crop. He was champion sire nine times, including a sequence of seven consecutive seasons.

Pocahontas and her descendants

St. Simon's maternal grandsire, King Tom, was a son of Pocahontas (1837), the only mare included in the Vuillier dosage system. Among her other sons was Stockwell (1849), a seven-time champion sire who was known as the "Emperor of Stallions." Although she was a non-winner on the racecourse, Pocahontas produced 15 foals of which 10 achieved fame, and it would be difficult if not impossible to find a Thoroughbred anywhere in the world today that does not trace to her in some line of its pedigree. (This rather aptly illustrates the fallacy of some breeding systems based on bloodlines; every classic winner descends from Pocahontas, but so does every non-winner.)

Pocahontas' sire, Glencoe (1831), was exported to North America, where he was champion sire eight times. He begat 50 known foals when he was 23 years old and was a genetic oddity: early records did not in every case indicate sex; but of Glencoe's 481 American foals whose sex was recorded, 317, or about two-thirds, were fillies.

The longest tenure as leading sire was that of Lexington (1850), who led the North American rankings for 16 seasons, 14 of them in succession. The modern record—the most remarkable of all in view of the vastly larger arena in which he was competing—is held by Bold Ruler, who was leading North American sire seven successive seasons (1963–69). He succeeded his own sire, Nasrullah, who held the unique distinction of having been leading sire in England (once) and in North America (five times).

The most remarkable racing record was compiled by the Hungarian mare Kincsem (1874), undefeated winner of 54 races in six countries: Austria, Czechoslovakia, England, France, Germany, and Hungary. She finished in a dead heat with Prince Giles in the 1878 Grand Prix at Baden-Baden, but the rules required a runoff, which she won by five lengths. Other famous undefeated runners include Barcaldine, Ormonde, Hurry On, The Tetrarch (who ran only at two), and Bahram in England; Ajax, Pharis, Le Pacha, and Caracalla in France; and Colin in the United States. The American horse Kingston (1884) won 89 races in 138 starts, and the mare Pan Zareta (1910) won 76 races in 151 starts.

World record holders and winners of selected famous English, French, and American races are listed in SPORTING RECORD in the *Ready Reference and Index*.

Triple crowns and classics

The highly-prized trilogy known as the English Triple Crown, consisting of the 2,000 Guineas, Derby, and St. Leger, has been won 15 times since 1853. The American Triple Crown, consisting of the Kentucky Derby, Preakness Stakes, and Belmont Stakes, has been won nine times since 1919. In France, where (with notable success) stamina is encouraged, the emphasis is on the classic double of the Prix du Jockey Club (French Derby) and the $1\frac{15}{16}$-mile Grand Prix de Paris, which has been won 19 times since 1875.

Sceptre (1899) won four English classics: the 1,000 Guineas and Oaks restricted to fillies, and the 2,000 Guineas and St. Leger from males. Seven other fillies swept the classics for their sex and won the St. Leger as well, among them Pretty Polly, winner of 22 races in 24 starts and subsequently a landmark broodmare.

The Ranger, winner in 1863 of the Grand Prix de Paris, was an English horse, but two years later France returned the favour with interest as Gladiateur won the Grand Prix de Paris in his own country and crossed the channel to sweep the English Triple Crown. (He also won two other coveted English races, the Goodwood Cup and Ascot Gold Cup.)

Over the years, competition between England and France became quite keen; and with the development of jet transportation, competition in classics has become worldwide.

Outstanding riders. English jockey Fred Archer, riding from 1870 to 1886, achieved 2,748 victories, and the 20th century was well along before his record was surpassed. Among those victories were 21 classics, and Archer compiled his total from 8,004 mounts for a winning average of 34 percent.

Personal records of the American Negro jockey Isaac Murphy, who had relatively few mounts (partially due to a weight problem that later led to his retirement) showed 628 victories in 1,412 races for a phenomenal 44 percent winning average. His official career was from 1876 to 1896, but when 14 Murphy had ridden a few races in 1875 under his original name of Isaac Burns.

Richards, Longden, and Shoemaker

Gordon Richards, who surpassed Archer's record for total victories in 1943, continued to ride through age 50, retiring with 4,870 victories. Sir Gordon, who was knighted by Queen Elizabeth II, was seasonal champion 26 times in 34 years, including a streak of 22 championships in 23 years. Richards' idol, Steve Donoghue, rode Triple Crown winners Pommern and Gay Crusader and won the Derby four other times in five years with Humorist, Captain Cuttle, Papyrus, and Manna (1921–25). Another of Donoghue's records was that in 33 years of riding he never was summoned before the stewards for disciplinary purposes.

John Longden, who surpassed Richards' record in 1956, retired in 1966 at age 59 with 6,032 victories. During his 40-year career Longden accepted 32,406 mounts in North America in addition to several in other countries (six of his victories were scored outside North America).

In 1970, William Shoemaker, at age 39, overtook Longden and finished the year with a career total of 6,072 victories. Shoemaker had amassed his total in only 22 seasons, during which he had 24,661 mounts and a winning average of 25 percent. His 485 victories in 1953 are a record for a single season, and, all told, his mounts through the end of 1970 had earned about $43,800,000, to average almost $2,000,000 per season. During 1971, Shoemaker continued to add to his records and by the end of 1972 he had set a new record of 577 victories in stakes races, the only significant gross record he did not already hold, breaking the previous record which had been held by the famed rider Eddie Arcaro, who retired in 1961. Shoemaker's success was a contradiction to the old belief that "live" weight on a horse is better than "dead" weight. A natural 100-pounder, Shoemaker had a special saddle pad with extra pockets for lead weights.

Among other jockeys are nine-time English champion Lester Piggott, who in 1972 rode his 20th English classic winner to come within striking distance of Archer's record; and young Yves Saint-Martin, who had nine French championships to his credit at age 30.

Riding styles

Among jockeys who influenced style, Tod Sloan, who flourished during the 1890s, did much to popularize short stirrups and the high crouch; but the style had been used previously, notably by black exercise boys and by North American Indians. John Randolph of Virginia (1773–1833) was said to have ridden "like a monkey on a stick," which description later was applied to Sloan. Similarly, although Arcaro did not originate the style, many other riders adopted his method, called "ace-deuce," of riding with the right stirrup shorter than the left.

History is crowded with references to women riders, but women as licensed professional jockeys in official horse races did not make their debut until 1969. A year later there were about a dozen of them in North America, none of whom had proved to be more than a reasonably competent journeyman rider.

Trainers, owners, and breeders. At the time of his death in 1970, the American trainer Hirsch Jacobs had saddled winners of 3,596 races. The former handler of racing pigeons had started out in horse racing with mostly cheap claimers; but, in partnership with Isidor Bieber, he developed a racing–breeding empire worth millions of dollars.

Official records of breeders were not maintained in the United States until 1918; but John E. Madden, who topped the standings for the first ten years such records were kept, bred winners of 3,477 races in just those years, and he had bred numerous winners previously.

Through the end of 1970, owner Marion H. Van Berg (who died in 1971) had won a total of 4,691 races. His stable, which he operated in divisions, included as many as 245 horses in the course of a season, not all of them full-time, as he was active in the claiming ranks. The 393 races won by the Van Berg stable in 1969 were a numerical record for a single season, but the financial record is held by Sigmund Sommer, whose stable earned a total of $1,605,936 in 1972.

Horses owned by Calumet Farm, most of them homebred, through 1970 had earned almost $20,000,000. The Calumet runners have included former world's leading money earners Armed, Whirlaway, and Citation, the latter two also winners of the Triple Crown. Horses bred by Calumet Farm through 1970 had earned almost $30,000,000.

Earnings. Offsetting the enormous gross purse distribution in North America is the fact that training expenses are generally higher. Moreover, the average annual earnings per runner of approximately $4,000 reflects $100,000 purses to which the truly average horse cannot aspire; and median earnings are less than half the average, about $1,600.

Nevertheless, as of July 1973, 12 horses had earned more than $1,000,000 each, and another was approaching that sum. The hope of acquiring such a horse has had a pronounced influence on prices for bloodstock.

Nijinsky, Canadian-bred winner of the English Triple Crown, was syndicated into 32 shares at $170,000 each, for a total valuation $5,440,000 at age three, before his third classic victory. American Triple Crown winner Secretariat was syndicated for a new record total of $6,080,000 before his first classic victory. The record price at public auction is $725,000, bid for the seven-year-old Typecast, who was classified as a race mare at the time of sale but was retired to stud without having raced for her purchasers. Top price for a brood mare as such was $450,000 paid for What A Treat, in foal to Vaguely Noble (and the foal she was carrying was sold subsequently for $326,400).

Even untried yearlings command enormous sums. A colt by Bold Ruler (sire of Secretariat) was sold in 1973 for $600,000. A daughter of European champion Sea-Bird, out of Hyperion's daughter, Libra, brought the record price for a yearling filly of $405,000 in 1968.

The sport of kings

The sport of kings—or king of sports, as racing came to be known as it developed from a diversion of the leisure class into a huge public entertainment business—is conducted on a sufficient scale to warrant a separate report in *The Bloodstock Breeders' Annual Review* in: Argentina, Australia, Austria, Barbados, Belgium, Brazil, Canada, Czechoslovakia, Chile, Colombia, Denmark, France, Germany, Greece, Holland, Hong Kong, India, Ireland, Italy, Indonesia, Jamaica, Japan, Kenya, Malaysia, Malta, Mexico, The Netherlands, New Zealand, Norway, Panama, Peru, Poland, Puerto Rico, Rhodesia, South Africa, Spain, Sweden, Switzerland, Trinidad, Turkey, Uruguay, the United States, the U.S.S.R., Venezuela, and Yugoslavia.

Nations in which racing is a major attraction

The occasions of certain races are regarded as public holidays. Estimates of attendance on Derby day at Epsom, where the public is admitted onto portions of the grounds at no fee, have ranged as high as 500,000. Official (paid) attendance at the 1969 Japan Derby was 168,000, and more than $13,000,000 was wagered on that race alone. Japan observed the 100th anniversary of organized racing in 1961. Among its unusual features is an Equestrian Park, at which there is a school for jockeys; the two-year course includes academic as well as equestrian subjects.

RACECOURSES

Proprietorship. Proprietorship of the sport varies from complete state operation and control, as in the Soviet Union—where the national government owns the tracks and the horses and employs the trainers, jockeys, grooms, and other necessary personnel—to complete private enterprise, as in most of the United States—where racetracks are operated for profit as a business, jockeys and trainers are independent contractors, and horses are owned by individual citizens. There are in-between versions of these arrangements. The government (provincial or national) might own the tracks and citizens might operate the racing stables (in some cases, leasing horses from the government). Some tracks are owned by a township (Doncaster, for example), others by a jockey club or equivalent organization (Newmarket, Longchamp), and still others are owned by private corporations but operated on a nonprofit basis (Australia, the New York Racing Association).

Stands and paddocks. The stands at racecourses range from elegant palaces with every imaginable facility (*e.g.*, Longchamp, Ascot, various tracks in the Americas, Japan, Australia), to relatively modest but functional structures (Cologne, The Curragh), to simple sheds for protection from the elements.

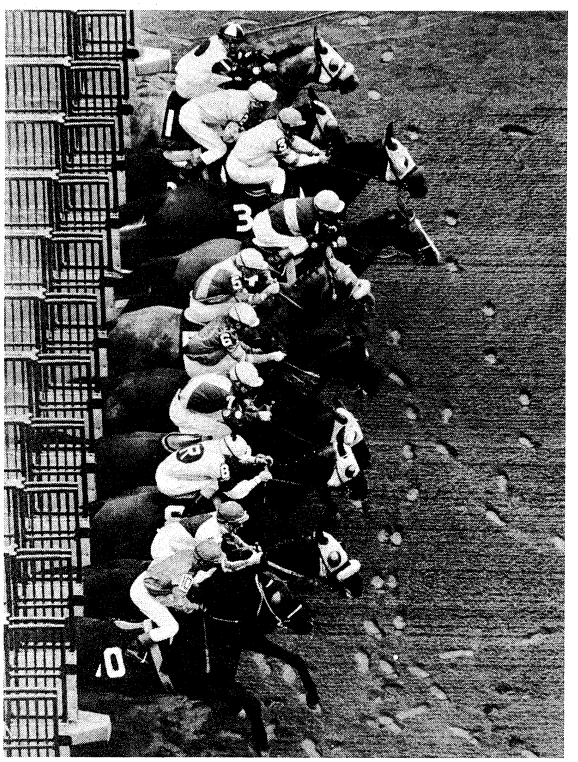
Paddocks at some tracks are ignored areas behind the stands where horses are saddled, while at other tracks they are garden-like enclosures surrounded by stands for spectators.

Track surfaces and layouts. The running surfaces of European tracks (and of tracks in some former dominions of European countries) are generally grass; in the Americas, dirt is the most common running surface although grass courses are quite popular, and in other countries there is racing on grass, dirt, and sand (the latter surface is used for training only in many countries). There are a few tracks in the United States with a synthetic surface designed to be impervious to weather, but some horsemen complain that it "grabs" the horses' hooves and insist that it be sprinkled with sand.

Some older tracks (mostly in Europe) conform to the natural terrain, having been laid out on previously existing race grounds, with facilities for spectators added later. Newer tracks are mostly elliptical, essentially flat, and a mile is the most common circumference.

The English Triple Crown offers a variety of tests—over a straightaway with a dip at Newmarket, the undulating, turning course at Epsom, and the flat, pear-shaped course at Doncaster—at distances from a mile (1,600 metres) to more than 1¾ miles (2,800 metres).

The course at Newmarket offers probably the truest run races, as it can accommodate a run of 2¼ miles



A field of horses breaking from the electrically operated starting gate at Aqueduct, Long Island, New York. As at most other U.S. tracks, the Aqueduct races are run on a dirt track.
Photo Trends

(3,600 metres) with a gentle change of direction of less than 90 degrees. Much of a race on this course is out of view of the spectators, however. The American Triple Crown events are all run on flat, elliptical tracks at distances that vary by only 2½ furlongs (5/16 miles or 500 metres). On the other hand, the English classics are spaced out in the calendar from April or May to September, while the American races frequently are bunched between the first Saturday in May and first Saturday in June, which requires that a horse maintain peak form with no letup for five weeks.

The Triple Crown courses

Grass courses require meticulous care and cannot take the constant pounding to which dirt tracks are subjected. Race meetings at a given European track generally are brief, some lasting only several days, and the track is reserved for racing only. Training is accomplished on gallops, or training tracks, set aside for that purpose; horses are vanned to the track to race and returned to their home yards at the training centres afterward. In the U.S., where it is not uncommon for a track to be used for 100 or more days of racing a year (some double as harness tracks), stabling is provided at the track, which is also used for training. (There are training tracks in some areas, but the main tracks are used, too.)

RACING RULES AND PROCEDURES

The rules of racing, representing as they do centuries of experience, are quite complex in their entirety. Briefly, eligibility of horses having previously been checked by entry clerks, the race procedure begins when jockeys weigh out and report to the paddock for instructions from trainers. An official in the paddock verifies the identity of the horses. After the jockeys mount, the horses enter the track and parade past the stewards for inspection. At some tracks this is a colourful spectacle as horses gallop onto the track one by one; but at the typical North American track they simply amble past the stands, more often than not accompanied by a lead pony, and do not break off into their warm-up gallop until they reach the far side of the track.

The use of electric starting gates is practically universal, although some starts still are effected from a barrier that springs upward when actuated by the starter, or in the "walk-up" fashion, whereby the starter gives a verbal order when the horses are reasonably well aligned. During the course of the race, patrol judges and stewards, supplemented by a motion-picture film patrol at most tracks, are alert to detect rule violations. The finish is photographed by a special camera, and if it is close, development of the film is awaited before listing the order of finish. The result does not become official until the jockeys have weighed in and riders of horses that finished in the money are certified by the clerk of scales to have carried proper weight.

Photo finishes and patrol films

At the time of weighing in a jockey may claim foul against a horse that interfered with his mount or against another jockey (or the owner or trainer may do so). If a horse that finished among the first three is involved, the stewards study films of the race and reach a decision prior to declaring the result official. If the infraction does not affect the order of finish for wagering purposes, official action usually is postponed so as not to delay the program. Patrol films are reviewed the next morning and used as educational tools for jockeys and officials, and also as a basis for disciplinary action.

For wagering purposes, the result of a race is official when so declared by the stewards; however, urine or saliva samples, or both, are extracted from winning horses (and others that might be designated by the stewards); and if a horse subsequently is determined to have won while under the influence of a prohibited substance, he is disqualified from the purse distribution.

Most time records are clustered in North America, opportunity being an obvious factor, as in no other area is there half as much racing as in the combined United States, Canada, northern Mexico grouping.

Although the difference is not enough to account for the generally faster North American running times, the method of timing Thoroughbred races in North America is less precise than that of most other countries, where horses are timed to the nearest 1/100 of a second, from a standing start. In North America they are timed to the nearest 1/5 of a second, from a running start—the length of the run between the starting gate and the official starting point varying from track to track (and, occasionally, varying according to the distance of the race at any one track).

The timing of races

RACING STRATEGY

The typical Thoroughbred is capable of running only a quarter of a mile or so at very top speed; and much of the strategy of racing lies in determination of the best moment at which to unleash this burst. Operating factors are distance of the race, size of the field, layout of the track, and nature of a particular horse and of the opposition. The burst cannot be delayed too long in a sprint, nor should it be turned on when there is heavy traffic ahead or while going wide around a turn and thus losing ground. On the other hand, if a fast horse is allowed to run in front unchallenged, he might exert himself so little as to be fresh enough to repel any challenges at the end. Sprinters have been known to win distance races in this manner, and for important races slow beginners often are accompanied by pacesetters, whose role is to challenge the fast horses in the early going and keep up the pace. Some horses are inclined to pull themselves up upon reaching the front; timing the move on such a horse so as to coincide with the finish pole is an art.

Training, simply expressed, is maintaining a horse in the best condition to run. Exercise and feeding programs and knowledge of the individual horse are factors involved. A good trainer selects a jockey who suits the horse and, perhaps more important, enters the horse in suitable races. A trainer of a horse for a classic race not only must develop the horse into peak condition but must time the development so that the horse reaches its peak on a certain day, which is the most difficult art of all.

Other types of racing

STEEPLECHASING

Steeplechasing, racing over jumps or obstacles, derives its name from impromptu contests over natural country

in which church steeples served as landmarks. This form of racing is quite ancient; Xenophon (431–*c.* 355 BC), much of whose writings on horsemanship remains pertinent today, was a strong advocate of riding over the countryside, jumping obstacles. Steeplechasing was long a favourite sport of cavalry officers, which tradition is reflected by the fact that steeplechasers, as a rule, carry much higher weights than flat racers.

The most famous steeplechase is the Grand National at Aintree, Lancashire, over a distance of four miles and 856 yards (7,180 metres), with 30 obstacles. Huge fields (there were 66 starters in 1929) and the hazards of the course have caused numerous injuries and some deaths, and there has been occasional agitation to outlaw this race. It has been won several times by horses that carried 175 pounds (79 kilograms).

Importance of stamina

Because of the stamina required, mature horses are preferred for steeplechasing, and racing through age ten and beyond is not uncommon. Speed is desirable but not if it is accompanied by impetuosity or temperament; and many steeplechasers are "half-bred," the term being loosely applied to any horse that is not a pure Thoroughbred although it may be three-quarters or seven-eighths so. Large horses also are generally preferred, although there are exceptions. Battleship, winner of both the English (at age 11) and American Grand Nationals was quite small; he was further unusual in that he was a stallion, whereas most steeplechasers are geldings.

Fred Winter, who rode two Aintree winners, also scored an epic victory in the Grand Steeplechase de Paris aboard Mandarin after the horse had lost its bridle.

While steeplechasing is a "poor relation" to flat racing in many countries, it is quite popular in England, France, and Ireland, especially the latter country, where Arkle, winner of 27 races in 35 starts and earner of the phenomenal sum (for a steeplechaser) of £75,207½, became a national hero second to none.

QUARTER HORSE RACING

Quarter racing, the racing of horses for short, traditionally one-fourth-mile (400-metre) distances, is indigenous to North America and was the earliest form of sport in the Colonies, having begun shortly after the settlement at Jamestown, Virginia, was established in 1607. The original quarter racers ran brief distances over whatever pathways through the forest were available. Generally, such pathways could accommodate only two horses (although racing through streets of a settlement later became common).

The Quarter Horse breed

Although they were for years recognized as a distinct type, registration of Quarter Horses as a breed did not begin until 1941. It is the largest equine registry in the world, with approximately 90,000 foals registered annually. Of these, about 12 percent are racing Quarter Horses, with the remainder divided between performing horses (ranch work) and pleasure horses.

Modern Quarter Horse racing is conducted at about 100 tracks throughout North America. Basically, the rules and procedures are the same as those that apply to Thoroughbred racing, but the races themselves are quite different. All are on a straightaway, varying in distance from 220 to 660 yards (200 to 600 metres), and vying for position along the rail or establishing pace plays no part. The horses run straight ahead, all out all the way. The combination of large fields and short distances produces close contests in which photo finishes are more the rule than the exception.

The All-American Futurity at Ruidoso Downs, New Mexico, a 400-yard (370-metre) event to which entry is gained through qualifying trials, in 1971 had a gross purse of $753,910 if breeders' awards are included. It is divided into three divisions, according to best qualifying times, and all qualifiers receive a share of the purse. Value to the winner of the championship division in 1971 was $200,841, from a total of $502,782 allocated to that division.

Unlike North American Thoroughbred races, Quarter Horse races are timed to the nearest 1/100 of a second, from a standing start. (W.H.P.R.)

HARNESS RACING

History. Harness racing—racing, trotting, and pacing horses in harness, pulling the drivers in light, two-wheeled vehicles called sulkies—an organized sport in 46 countries of the world, has an ancient past and an active and flourishing present. Between the two is a chasm of nearly 2,000 years of obscurity.

Archaeological excavations in the early 1930s in Boğazköy, Turkish Asia Minor, unearthed baked-clay tablets containing, intact, some 900 lines of detailed description of the training of racehorses by one Kikkulis, head trainer for Suppiluliumas, King of Mitanni, a country later known as Cappadocia, around 1350 BC.

These tablets, when translated and presented before the Académie des Inscriptions et Belles-Lettres in 1934, revealed a highly organized and sophisticated program for the general preparation of horses for training at speed at both the trot and the gallop. Specific and minute in detail, they included a 144-day program, starting in spring, which led either to the racecourse, hunting field, or battlefield. It was obvious, from the tablets, that the trotting horse occupied the royal place of honour in the racing affairs of that part of the Assyro-Babylonian Empire, the leading horse-producing area of its day.

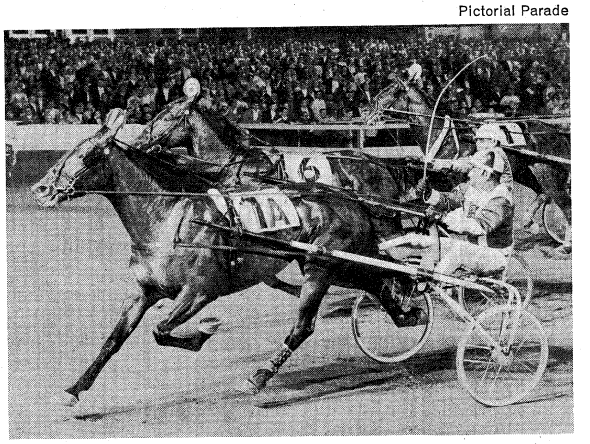
With the disappearance of the ancient racing trotters of Mitanni–Cappadocia, however, the trail of the trotting racehorse is lost until it reappears informally in Holland in the 16th century, on the road in England in the 18th, and in full glory as a racing phenomenon in the United States in the early 19th century.

Reappearance of the trotting horse

Dutch trotters competed as early as 1554 at Valkenburg, when 3,000 horses were brought together at a fair and trotting matches were held among the fastest. Holland's "Golden Whip," most famous of Dutch trotting events, dates back to 1777 at Soestdijk.

About the same time as this trotting activity among Dutch Fresian horses in the late 18th century, Count Grigory Orlov was developing a powerful trotting strain at his stud in Russia, and from his stallion Barss came the Orlov trotter that remains today as the foundation of Russian trotting stock.

Pictorial Parade



Harness Racing.
Trotters streaking to the finish. In the lead is Flamboyant, winning the Realization Trot at Roosevelt Raceway, New York, June 1968.

England's Norfolk Trotter, which emerged as a breed around 1750, was purely a road horse, bred for use on the highways of the day. Its speed led to its being used for road racing as a diversion for its owners, and most of its matches were trotting a given distance within a specified amount of time.

American trotting, too, had road racing as its earliest heritage; but by the end of the first decade of the 19th century, trotting racetracks were in use in the United States. A mile in 2:59 was recorded by a horse called Yankee over the track at Harlem, New York, in 1806, and was lowered to 2:48½ on August 25, 1810, by an unnamed trotting gelding "from Boston" at the Hunting Park track in Philadelphia.

The American Standardbred. By 1840, trotting was an organized sport in New York and New England, and a new era was underway. This era culminated, in 1879, with establishment of the American Standardbred horse in the United States, based on a standard of time performance—2 minutes 30 seconds—for one mile. It marked creation of the first pure breed and variety of trotting racehorse, as opposed to the road horse, since the Cappadocian trotters of 1350 BC.

Eight years before the Standardbred was established came the formation, in 1871, of the Grand Circuit of American harness racing. Originally known as the Quadrilateral Trotting Combination, it grew from four to 23 tracks, celebrated its centennial in 1971, and is the oldest organized sports group in the United States.

Messenger and Hambletonian

The creation and evolution of the American Standardbred and that horse's impact on world trotting rested on the prepotency of the English Thoroughbred stallion Messenger, imported to Philadelphia in 1788. Messenger combined the blood of the three founding fathers of the Thoroughbred horse—the Darley Arabian, Godolphin Arabian, and Byerly Turk—and he became a major contributor to the American Thoroughbred through his undefeated grandson American Eclipse, the descendants of which included a number of great runners such as Whirlaway, Equipoise, Gallant Fox, Exterminator, and Man o'War.

It was as a sire of Thoroughbred runners that became trotting stallions, however, that Messenger attained immortality. Ten of his sons became leading trotting sires of the early 19th century, and his great-grandson, Hambletonian 10, foaled in 1849, sired 1,331 sons and daughters between 1851 and 1875 and obliterated all other strains of the trotting horse in America. He founded a line so dominant that all American Standardbreds today and many trotters in the rest of the world can be traced to him.

Development of the sport. Trotting, meanwhile, was spreading rapidly. The sport was introduced in France, at the mountaintop course at Cherbourg, in 1836; Italy's Padovanella dates back to the same era; Austria and Germany became active in the racing of trotters late in the century; and Sweden began harness racing in the 1880s.

By this time American trotting had moved from the agricultural fairs where it thrived in the mid-1800s to regular harness racetracks, and the rise of colourful and charismatic champions like Lady Suffolk, nicknamed the "old gray mare"; Flora Temple, the "bob-tailed mare" of the well-known song "Camptown Races"; the legendary Goldsmith Maid and Dexter; and particularly the popular Maud S. led to an extremely active interest in American breeding stock in Europe and around the world.

One rival of Lady Suffolk—the gelding Americus—had been sent on a highly successful invasion against English road trotters in the 1850s; and in the same decade the trotting mares Lady Pierce and Miss Bell had been exported to France. Both raced well there, and Lady Pierce produced a line that led to Fuschia, greatest early trotting progenitor of the French breed.

In the last decade of the 19th century, American blood was introduced into virtually all trotting nations of the world—2,000 American Standardbreds had been exported by 1898 and 3,000 by 1903—and the American influence has remained a factor ever since.

The French and Russians, who had established trotting breeds of their own in the strong demisangs (half-bloods) of Normandy and the Orlovs, resisted the American influx; and the French closed their studbook to foreign horses in 1937. Elsewhere, however, American blood predominated; and in Italy, Germany, Scandinavia, the Low Countries, Argentina, and across the world in Australia and New Zealand, most present-day horses of major rank can be traced to American heritage.

Trotting and pacing gaits

Trotters and pacers. To talk of trotters exclusively covers the subject in Europe, where pacers are not raced; but in the United States and Australasia today the pacing horse is predominant, at least in numbers. The pacing horse, sometimes known as a "sidewheeler," moves both legs on one side of its body at the same time, in contrast to the diagonal gait of the trotter, which strides with its left front and right rear leg moving forward simultaneously, then right front and left rear together.

The infusion of American blood into Australia and New Zealand began in the 1880s; and, with little harness-racing background on which to draw and pacers far easier to train, that branch of the Standardbred family attained immediate popularity with trainers and owners. It still prevails.

The American pacer descended a different path from that of his trotting cousin. His heritage fuses the blood of the Narragansett pacer—a saddle horse that disappeared from the scene by 1850—and the Canuck of French Canada. While the American trotter had his beginnings in the East, the great growth of the pacer came in the Midwest and South—primarily Ohio, Indiana, Kentucky, and Tennessee—and he was a despised horse until he attained popularity in the late 1800s.

The arrival of the "Big Four"—Mattie Hunter, Sleepy Tom, Rowdy Boy, and Lucy—on the Grand Circuit of the late 1870s; the coming of the first two-minute harness horse, the pacer Star Pointer, in 1897; and the overwhelming national popularity of the first sports hero of the 20th century, Dan Patch, did much to foster the popularity of pacers.

Harness racing's first zenith, in the late 1800s, was followed by a sharp decline with the advent of the automobile and the passing of the road horse.

Revival with pari-mutuel betting under lights. But the European trotting derbies and traditional American Grand Circuit and county-fair races persisted and survived. With the introduction of regularly scheduled pari-mutuel racing under lights at Roosevelt Raceway in New York City in 1940 (there had been night racing as early as the 1890s and a meeting under lights in Toledo, Ohio, in 1927) and the introduction of the mobile starting gate (a pair of retractable metal wings mounted on the rear of an automobile or light truck that moves off slowly, getting the horses off to an even start, and then accelerates away and off the track), also at Roosevelt, in 1946, a new era was ushered in on the sports scene.

Introduction of the mobile starting gate

Tracks proliferated; breeding boomed; international trade and competition accelerated; attendance, pari-mutuel turnover, and purses soared. Attendance at American harness tracks rose from 5,700,000 in 1948 to 27,200,000 in 1971; state revenue rose from $8,500,000 to $162,000,000; purses went from $9,800,000 to $97,600,000; horses starting from 9,300 to 37,600; and members of the United States Trotting Association from 7,300 to 34,600.

The two leading money winners of all time are the trotting mares Une de Mai of France and Fresh Yankee, Canadian-owned but bred in the United States. Une de Mai has won nearly $1,900,000 and Fresh Yankee ended her career with $1,294,252. The two leading money-winning pacers are Albatross, retired in 1972 with earnings that totalled $1,201,470, and Rum Customer, retired with total earnings of $1,001,548.

American racing, meanwhile, produced its own roster of superhorses in the last three decades, starting with the trotting champion Greyhound in the late 1930s and including the pacers Adios, most successful money-siring stallion of any breed of all time, Good Time, Bye Bye Byrd, and Adios' sons Adios Harry, Adios Butler, and world champion Bret Hanover; and the outstanding trotters Su Mac Lad, Noble Victory, Nevele Pride, and Fresh Yankee.

Jamin of France, Tornese of Italy, and Hairos II of The Netherlands also proved true world-class competitors in international competition. The advent of the Roosevelt International Trot in 1959, the International Pace series at Yonkers Raceway in the 1960s, and the introduction of the World Driving Championship by Harness Tracks of America in 1970 all fostered international competition; and the United States Trotting Association, which administers all central record keeping of the American Standardbred, joined with counterpart Euro-

pean and Australian and New Zealand trotting groups in 1970 and 1971 to bring closer worldwide cooperation and uniformity to the sport.

Today, major tracks abound around the world. More than 60 pari-mutuel and 380 fair meetings are conducted annually in the United States; and virtually all principal cities of the North have racing plants, with New York City's Roosevelt and Yonkers raceways, Hollywood Park in Los Angeles, Sportsman's, Washington, and Maywood parks and Hawthorne Race Course in Chicago, and Liberty Bell Park in Philadelphia the largest. Paris has two great trotting centres at Vincennes and Enghein, Rome its Tor di Valle and Milan its San Siro, Stockholm its Solvalla, Vienna its Krieau, Moscow its Hippodrome; and there are fine tracks in Munich and the German Ruhr. In Australia, Harold Park in Sydney, the Royal Showgrounds in Melbourne, and Gloucester Park in Perth host major harness events; and New Zealand has excellent plants at Auckland and at Addington Raceway in Christchurch.

Herve Filion of Canada, William Haughton, and Stanley Dancer have dominated American driving championships in recent years. Filion's 605 victories and $2,473,265 in purses won by horses he drove in 1972 were world records, and he has won more than 4,000 races. Filion's 4,065 was a North American record entering 1973 and Haughton's $17,878,566 in career purses a world mark. In Europe the veteran Hans Fromming, with more than 5,000 victories, was the all-time leader. For world trotting and pacing records, see SPORTING RECORD in the *Ready Reference and Index*. (S.F.B.)

Wagering

EARLY DEVELOPMENTS

From the beginning, wagering has been an integral part of horse racing. The early rules of racing took cognizance of its gambling aspect (and one rule prohibited "spying" on horses during their training trials, in sharp contrast to the modern attitude that the form of horses is information within the public domain).

The built-in wagers represented by owners' stakes in match races necessarily were at natural odds (even money for a two-horse race, 2-to-1 for a three-horse race, etc.); but betting among spectators quickly developed, in which odds reflected prevailing opinion as to the horses' relative chances of victory. Man-to-man betting was the original form, but this required that a person who wished to wager find another person of opposite opinion but similar purse; and it was a natural step to the appearance of professionals who would accept bets in any reasonable amount from all comers.

The role of the bookmaker. As racing among fields of horses became popular, laying of odds became more complicated, and the bookmaker appeared on the scene. In theory, bookmaking is quite simple: the bookmaker sets his odds so that a percentage is working in his favour, an oversimplified example being a race among seven evenly matched horses in which he would offer 5-to-1 (a point shorter than natural odds) on each. Assuming equal amounts wagered on each horse, regardless of which one won the race, the bookmaker's profit would be one-seventh of the total wagered. In practice, however, such a "percentage book," in which the bookmaker has no financial interest in which horse wins or loses, is very difficult to achieve; and despite constant adjustment of the odds as bets are received, the bookmaker generally stands to win or lose depending on the result of the race. Odds quoted at the time of a wager are firm; and although the bookmaker can reduce the odds on the favourite for subsequent betting, he must pay off each wager according to the original terms. (In an extreme case, he might take the favourite off his slate and refuse further wagers on that horse or lay off all or part of the risk by betting on the favourite with other bookmakers; in either case he curtails his business.) The financial stakes of bookmakers and bettors in a race's outcome create a temptation to tamper with horses.

Theory of bookmaking

Entrepreneurs who preferred a guaranteed percentage to the risky business of laying odds conducted auction pools on races, in which the "odds" were established by the wagerers, so to speak. (Such pools continue to be conducted on other sports.) Each horse was "sold" at auction to the highest bidder; and after deduction of the pool seller's percentage, receipts were turned over to the buyer of the winning horse. The shortcoming of the system was that it could accommodate only one bettor (the highest bidder) per horse in each pool.

Pari-mutuel betting. In 1872, Pierre Oller, a Parisian shopkeeper (and, according to some reports, a bookmaker who had experienced frequent losses), devised a modification of the auction pool whereby betting tickets could be purchased in any quantity the wagerer desired; and after deduction of Oller's commission, the wagering receipts were distributed among holders of tickets on the winning horse in proportion to the number of tickets held by each. Oller's system, which he called "pari mutuel" (bet among ourselves) proved to be extremely popular; and, although the French government at first frowned upon it, the pari-mutuel system subsequently was declared the only legal means of wagering in France. The "takeout" specified in an 1891 regulation was 7 percent, of which 4 percent was allocated to the racing societies, 2 percent to certain charities, and 1 percent to the Minister of Agriculture for development of horse breeding. Although a tax on wagering has been imposed and the takeout has more than doubled since then, the tradition in France of providing reasonable return to the sport that produces substantial tax revenue has been preserved; and France is the envy of other countries in the economic health of its racing and breeding industry.

TOTALIZATORS

The Ekberg totalizator, a machine that mechanically records bets, was first used in New Zealand in 1880; a similar device, known as the "Australian tote," became popular in America. Some form of totalizator, now often integrated with electronic computers, is used in all pari-mutuel operations.

Bookmaking dies hard. Apart from a preference for doing business with a human being rather than with a mechanical contrivance, bettors were distrustful of the early totalizators (sometimes, with good reason). The payoff odds are not known until all betting is completed; and it was dismaying to bet on a horse at 4-to-1, only to see the odds drop and the eventual payoff dwindle to 2-to-1. Dark rumours abounded of unscrupulous track operators who would insert large wagers into the totalizator system after the race was well underway and they had assured themselves that their choice had got off to a good start. Refinements of the totalizators and the increasing supervision of governmental agencies as the tax on pari-mutuel wagering became more and more significant as a revenue source have ended such complaints.

Persistence of bookmaking

The pari-mutuel system received numerous tryouts in the U.S. during the 19th century but was not permanently established until 1908 (in Kentucky); and not until 1940 did New York become the last of the states to abolish bookmaking and make pari-mutuel wagering the only legal form. The impetus, in virtually every case, came from state governments, since pari-mutuel systems offer an exceptionally convenient method of extracting tax revenues.

In countries such as England, where both bookmaking and pari-mutuel wagering are offered, the former is by far the most popular form. Big bettors complain that they are forced to bet against themselves in the totalizator, in that a large bet depresses the odds. Small bettors, with an off-course bookmaker, can wager trivial sums that the tote does not accept; and to bettors large and small the off-course bookmaker offers the convenience of betting by telephone—and on credit. Even in some countries where bookmakers are outlawed, they continue to flourish illegally.

Betting pools. Following a custom that had applied to auction pools, in some of the early pari-mutuel systems a percentage of the money wagered to win was set aside for bettors who had backed the second (or third)

horse. This was in the nature of a consolation award rather than a successful bet, and the amount returned often was less than the amount wagered.

The modern totalizator displays approximate odds to win on each horse and the total amount of wagering on each horse in each of various betting pools. Customary pools are "win," "place," and "show." In some countries a "place" bet is paid off if the horse finishes among the first three; but in America such bets apply only to the first two finishers, and "show" betting applies to the first three finishers.

Win, place, and show entry and field bets

The pools are entirely distinct from each other and there is no relationship between the various payoffs, although the favourite in the win pool usually is also the favourite in the other two. The typical volume of wagering is in the order of win, place, and show; and payoffs are in corresponding order. There are instances, however, of the payoff for a place bet exceeding that for a win bet, and show payoffs higher than place payoffs are not at all rare. The latter occurs most frequently when more than one member of an "entry" or the mutuel "field" finishes among the first three. An entry is the coupling of two or more horses of the same ownership, or trained by the same trainer, into a single unit for betting purposes; a mutuel field is a similar consolidation of lesser regarded runners in a field that is too large for the totalizator to accommodate as individuals. For example, if there are 15 runners at a track whose totalizator is set up to handle 12 betting units, four of the horses would be lumped together as a mutuel field; and a bet on one member of the field would include a bet on the other three also.

The takeout, which varies from 12 percent to as high as 25 percent depending on locale, is divided between government and racing associations. The allocation to the racing association is used to pay purses, defray other expenses of operation, and (as the case might be) generate profit for the track operators.

In a win pool, after deduction of the takeout, the remaining money is distributed among backers of the winner only, in proportion to the amount wagered by each.

In some countries, a "place" pool, after deduction of the takeout, was simply divided into three equal parts, which were distributed among holders of place tickets on each of the first three finishers in the same manner that applied to a win pool. In such cases, the backer of a heavy favourite in a place pool could receive less than the amount of his wager. This method has lost favour to the method that applies to an American "show" pool, in which the money is distributed somewhat differently. After deduction of the takeout, the money wagered on each of the first three finishers also is deducted and set aside. What remains represents the winnings; and one-third of this sum is added to the money wagered on each horse to constitute the payoff pool for that particular horse. Pari-mutuel laws require that a successful bettor receive a minimum profit (5 percent or 10 percent) on his wager. Occasionally, such as when an exceptionally heavy favourite finishes among the first three—or two well-backed runners do so—after deduction of the takeout and setting aside of the money wagered on the first three finishers there is not enough money remaining to provide this minimum profit. In such a case, the track must make up the difference, which is known as a "minus pool." The term is rather a misnomer in that practically never does the track actually lose money on such a race—it merely has to forego some of its customary percentage and realizes less money than was expected. Minus show pools are the most common, and a minus win pool is a rarity. Tracks can protect themselves against minus pools by eliminating show or place wagering, as appropriate.

The "minus pool"

Breakage. Beyond the percentage takeout, in most pari-mutuel systems a wager is subjected to "breakage," which has been aptly described as a form of legalized larceny. On the pretext that it would be too inconvenient to include odd pennies, payoffs are rounded off to the nearest, lower, even sum. In the United States, for example, "breaks" are usually to ten cents, and the calculation of payoffs is based on the dollar, although the minimum betting denomination is $2 and other bets are in multiples of $5 (which would not entail odd pennies in the payoff). Thus, if a calculated payoff amounts to $1.89 per dollar, the break to $1.80 is applied first and then the result is doubled to arrive at $3.60 as the actual payoff on a $2 wager, resulting in a loss to the bettor of 18 cents instead of eight cents. A proper payoff on a $10 wager would be $18.90, which would require no rounding off to arrive at an even ten cents, but breakage nevertheless is applied at the $1 level and multiplied by ten, resulting in a payoff of only $18.00.

Variety bets. In addition to win, place, and show wagering, pari-mutuel systems offer a variety of other bets: doubles (selecting winners of two races), twin doubles (winners of four races), quinielas or quinellas (the first two horses to finish regardless of the order between them), exactas, perfectas, or forecasts (the first two horses in precise order), and so forth. More elaborate wagers are offered at some pari-mutuel tracks and payoffs figured according to the same considerations that govern regular wagering, although separate equipment is necessary. As is to be expected in view of the odds against success, payoffs on such wagering usually exceed those of regular wagering; and in the case of the more elaborate wagers the payoffs occasionally exceed $100,000.

Off-track betting. Off-track betting through the totalizator in Australia, New Zealand, and South Africa, for example, has been very beneficial to racing as it has increased revenue for prize money and development of racetracks. In France, the most popular form of off-track wagering is the *tierce*, in which the bettor attempts to select the first three finishers in precise order. Outlets for wagering are cafés and tobacco shops, which entail little extra expense; and the *tierce*, which is conducted on about 75 races a year, is largely responsible for the enviable economic health of racing in France.

In England, off-track betting shops (which offer wagering on other sports besides horse racing) have in general been beneficial to the shop operators but of no benefit to racing. Closure of a number of smaller racecourses has been attributed to competition from off-track betting shops.

In the United States, off-track betting made its debut in New York City in 1971. Although the announced objective was to tie the off-track betting machinery directly to the track totalizator system, it was not immediately achieved. One stumbling block was the refusal of racetracks and horse owners to cooperate with the off-track betting corporation on the ground that the share of off-track wagering allocated to them as compensation for loss of patronage at the tracks was woefully inadequate: only 1½ percent, compared to more than 7 percent of on-track wagering, was returned to the sport. As of mid-1971, modifications of the New York City system and various other off-track betting plans were being studied in a number of states.

Racing in the United States and Canada stood at a great crossroads. Depending upon how the extremely delicate and very complex matter of off-track wagering was handled, the sport could go on to even greater heights—or it might be killed or diminished beyond recognition.

Theatre racing, in which patrons could wager on televised events, had been proposed. If such a system were adopted, the possibility exists that patrons would prefer to wager on the most famous horses running in the richest races, to the exclusion of local events, and the sport would dwindle to comparatively few horses competing for enormous purses in each of a few racing centres. Whether the sport could survive without its traditional grass roots is questionable. (W.H.P.R.)

BIBLIOGRAPHY. *Stud Books* of various countries provide records of pedigrees. *Racing Calendars, Turf Registers*, or, in North America, the *American Racing Manual*, include seasonal racing records of individual horses, conditions of races, winners of important fixtures, and other statistical data. The

Thoroughbred Record (weekly), publishes as one issue each year an annual statistical review of North American racing. In addition to detailed reports and statistics concerning racing in England and Ireland, the *Bloodstock Breeders' Annual Review* offers an excellent diary, year by year, of racing throughout the world. T.H. BROWNE, *History of the English Turf, 1904–1930*, 2 vol. (1931), is a continuation of a similarly entitled series by THEODORE A. COOK (3 vol., 1901–04), authoritatively written and excellently illustrated. W.S. VOSBURGH, JOHN L. HERVEY, and ROBERT F. KELLEY, *Racing in America*, 5 vol. (1922–60), is a comprehensive record of American sport through 1959; WILLIAM H.P. ROBERTSON, *The History of Thoroughbred Racing in America* (1964), is a one-volume work. C.M. PRIOR, *History of the Racing Calendar and Stud-Book* (1926), traces the evolution of English racing. SIR CHARLES LEICESTER, *Bloodstock Breeding* (1958), summarizes various breeding theories and discusses in detail pedigrees of English Derby winners, but many of the author's observations are of universal application.

Harness racing: JOHN L. HERVEY, *The American Trotter* (1947), is the most authoritative and definitive history of the Standardbred horse and world harness racing. TOM AINSLIE, *Complete Guide to Harness Racing* (1970); and PHILIP A. PINES, *The Complete Book of Harness Racing* (1970), are excellent guides to the sport today, the former stressing race-going and wagering and the latter an overall view of the sport. JAMES C. HARRISON *et al.*, *Care and Training of the Trotter and Pacer* (1968), covers every aspect of training, racing, and stable management. DWIGHT AKERS, *Drivers Up*, 2nd ed. (1947), is an interesting history of the sport's evolution. MARIE HILL, *Adios: The Big Daddy of Harness Racing* (1971), is a well-documented history of the most successful racehorse stallion of all time. DONALD P. EVANS, *Big Bum: The Story of Bret Hanover* (1969), reviews the career of the world champion pacer; and P.W. MOSER, *The Story of Greyhound* (1940), does the same for the trotter that held the world record for 31 years. MARGUERITE HENRY, *Born to Trot* (1950), is a good introductory book for younger readers. The UNITED STATES TROTTING ASSOCIATION, *Hoof Beats*, is the official monthly magazine of the sport and the same association's *Trotting and Pacing Guide* is a complete record book, published annually, containing full statistics of every phase of the sport.

(W.H.P.R./S.F.B.)

Horticulture

Horticulture is the branch of plant agriculture dealing with garden crops, generally fruits, vegetables, and ornamentals. The word is derived from the Latin *hortus*, garden, and *colere*, to cultivate. The English word horticulture is first found in 1678 in Edward Phillips' *The New World of English Words;* the Latin *horticultura* first appears in 1631 in Peter Lauremberg's book by that name. The word *hortus* is found in classical Latin.

The concept of garden culture as distinct from field culture dates back to medieval farming systems that used extensive areas for grains, forages, and pastures (commons); small intensive kitchen gardens; and wild lands for timber and game. Modern agriculture has maintained these traditional distinctions in the divisions of agronomy (extensively cropped grains, forages, fibre crops), horticulture (intensively cultivated garden crops for food and ornament), and forestry (forest trees and products; also wildlife).

The field of horticulture is traditionally divided into food crops (pomology and olericulture) and ornamentals (floriculture and landscape horticulture). Pomology deals with fruit and nut crops. Olericulture deals with herbaceous (nonwoody) plants, including carrot (edible root), asparagus (edible stem), lettuce (edible leaf), cauliflower (edible flower), tomato (edible fruit), and pea (edible seed). Floriculture deals with production of flowers and ornamental plants: generally, cut flowers, potted plants, and greenery. Landscape horticulture is a broad category that includes plants for the landscape including turf but particularly nursery crops such as shrubs, trees, and vines.

Horticulture also may be separated into scientific, agricultural, or artistic disciplines. These include genetics and breeding, physiology and cultural practices, soils and fertilization, economics and marketing, storage and processing, flower design and landscape architecture. Thus horticulture is an art and a science, an avocation and an industry. As an art, horticulture involves unique skills involving careful timing of ancient practices, many of which are empirically derived and perfected. It may involve an aesthetic sense of beauty and form. As an avocation, horticulture and gardening are outlets for recreation and pleasure for millions of people. Horticulture is also a huge business involved in all production phases from planting to processing. In addition, horticulture can be considered a body of technology involving many associated sciences and a scientific discipline in its own right. The science has provided methods and resources to explain the art and has become the guiding force for the improvement and refinement of horticulture.

This article is primarily concerned with the general aspects of horticultural practices, particularly those in relation to ornamental crops. For small, noncommercial gardens see GARDENING, for fruits and vegetables see FRUITS AND FRUIT FARMING and VEGETABLES AND VEGETABLE FARMING.

Horticultural practices

PROPAGATION

Propagation, the controlled perpetuation of plants, is the most basic of horticultural practices. Its two objectives are to achieve an increase in numbers and to preserve the essential characteristics of the plant. Propagation may be achieved sexually by seed or asexually by utilizing specialized vegetative structures of the plant (tubers or corms) or employing such techniques as cutting, layering, grafting, and tissue culture. (A detailed discussion of the methods of controlling sexual propagation may be found in the article PLANT BREEDING.)

Seed propagation. The most common method of propagation for self-pollinated plants is by seed. In self-pollinated plants, the sperm nuclei in pollen produced by a flower fertilize egg cells of a flower on the same plant. Propagation by seed is also used widely for many cross-pollinated plants (those whose pollen is carried from one plant to another). Seeds are usually the most inexpensive and often the only method and offer a convenient way to store plants over long periods of time. Seeds kept dry and cool normally maintain their viability from harvest to the next planting season. Some can be stored for years under suitable conditions. Seed propagation also makes it possible to start plants free of most diseases. This is especially true with respect to virus diseases, since it is almost impossible to free plants of virus infections and since most virus diseases are not seed transmitted. There are two disadvantages to seed propagation. First, genetic segregation occurs in seed from cross-pollinated plants because they are heterozygous. This means that the plant grown from seed may not exactly duplicate the characteristics of its parents and may possess undesirable characteristics. Second, a long time is required in some plants from seed to maturity. Potatoes, for example, do not breed true from seed and do not produce large tubers the first year. These disadvantages are overcome by vegetative propagation.

The practice of saving seed to plant the following year has developed into a specialized part of horticulture. Seed technology involves all of the steps necessary to assure production of seed with high viability, freedom from disease, purity, and trueness to type. These processes may include specialized growing and harvesting techniques, cleaning, and distribution.

The location of the seed-producing area varies with the crop. When the seed or the associated fruit is the commercial part of the crop, seed is usually produced within the crop-growing area. When the seed is not a usable part of the crop, as in ornamentals and many vegetable crops, seed production tends to concentrate in areas separate from crop-producing areas.

Relatively little tree and shrub seed is grown commercially; it is generally harvested from natural stands. Rootstock seed for fruit trees is often obtained as a by-product in fruit processing industries. Seed growing and plant improvement are related activities. Thus many seed-producing firms actively engage in plant-breeding programs to accomplish genetic improvement of their material.

F_1 hybrids

F_1 is the designation of the first generation resulting from crossing two different true breeding or homozygous strains. The plant so produced may exhibit better characteristics than either parent. The advantages of hybrids also include uniformity and increased vigour (heterosis or hybrid vigour). But a continuance of the same desirable characteristics cannot be obtained simply by crossing F_1 with F_1, because the offspring will exhibit the whole range of characteristics inherent in their heredity. Thus, seed of hybrids must be remade each year. Possession of the unique set of parents gives a seedsman even more effective control of a hybrid than does the patent process.

Harvesting of dry seeds is accomplished by threshing. Seeds from fleshy fruits are recovered through fermentation of the macerated (softened by soaking) pulp or directly from screening. Machines have been developed to separate and clean seed, based on size, specific gravity, or surface characteristics. Extended storage of seed requires low humidities and cool temperature.

Trade in seeds requires quality control. For example, U.S. government seed laws require detailed labelling showing germination percentage, mechanical purity, amount of seed, origin, and moisture. Seed testing is thus an important part of the seed industry.

While most vegetable seed germinate readily upon exposure to normally favourable environmental conditions, many seed plants that are vegetatively propagated fail to germinate readily because of physical or physiologically imposed dormancy. Physical dormancy is due to structural limitations to germination such as hard impervious seed coats. Under natural conditions weathering for a number of years weakens the seed coat. Such seeds may be artificially worn or weakened to render the seed coat permeable to gases and water by a process known as scarification. This is accomplished by a number of methods including abrasive action, hot water, or acid treatment. Physiologically imposed dormancy involves the presence of germination inhibitors. Germination in such seed may be accomplished by treatment to remove these inhibitors. This may involve cold stratification, storing seed at high relative humidity and low temperatures, usually slightly above freezing. Cold stratification is a prerequisite to the uniform germination of many temperate zone species such as apple, pear, and redbud.

Scarification

Vegetative propagation. Asexual or vegetative reproduction is based on the ability of plants to regenerate tissues and parts. In many plants vegetative propagation is a completely natural process; in others it is an artificial one. There are many advantages to vegetative propagation. These include the unchanged perpetuation of naturally cross-pollinated or heterozygous plants and the possibility of propagating seedless progeny. This means that a superior plant may be reproduced endlessly without variation. In addition, vegetative propagation may be easier and faster than seed propagation, because seed-dormancy problems are eliminated and the juvenile nonflowering stage of some seed-propagated plants is eliminated or reduced.

Vegetative propagation is accomplished by use of (1) apomictic seed; (2) specialized vegetative structures such as runners, bulbs, corms, rhizomes, offshoots, tubers, stems, or roots; (3) layers and cuttings; (4) grafting and budding; and (5) tissue culture.

Apomixis. Apomixis, the development of asexual seed (seed not formed via the normal sexual process), is a form of vegetative propagation for some horticultural plants including Kentucky bluegrass, mango, and citrus. Virus-free progeny may be produced in oranges from a seed that (as in many types of citrus) is formed from the nucellus, a maternal tissue.

Vegetative structures. Many plants produce specialized vegetative structures that may be used in propagation. These may be storage organs such as tubers that enable the plant to survive adverse conditions or organs adapted for natural propagation—runners or rhizomes—so that the plant may rapidly spread.

Bulbs (shortened stems with thick, fleshy leaves) are found in such plants as the onion, daffodil, and hyacinth. Bulbs commonly grow at ground level, though bulblike structures (bulbils) may form on aerial stems in some lilies or in association with flower parts, as in the onion. Buds in the axils (angle between leaf and stem) of the fleshy leaves may form miniature bulbs (bulblets) that when grown to full size are known as offsets. Corms are short fleshy underground stems without fleshy leaves. The gladiolus and crocus are propagated by corms. Corms may produce new cormels from fleshy buds. Rhizomes are horizontal, underground stems that may be compressed, as in iris, or slender, as in turf grasses. Runners are specialized aerial stems, a natural agent of increase and spread for such plants as the strawberry, strawberry geranium, and bugle weed (*Ajuga*). Tubers are fleshy enlarged portions of underground stem. The edible portion of the potato, the tuber, is also used as a means of propagation.

A number of plants form lateral shoots from the stem, which when rooted serve to propagate the plant. These are known collectively as offshoots but are often called offsets, crown divisions, ratoons, or slips.

Roots may also be structurally modified as propagative and food-storage organs. These tuberous roots, fleshy swollen structures, readily form shoots that are called adventitious because they do not form from nodes. The sweetpotato and dahlia are propagated by tuberous roots. Shoots that rise adventitiously from roots are called suckers. The red raspberry is propagated by suckers.

Layerage and cuttage. Propagation may be accomplished by methods in which plants are induced to regenerate missing parts, usually adventitious roots or shoots. When the regenerated part is attached to the plant the process is called layerage; when the regenerating portion is detached from the plant the process is called cuttage.

Layerage often occurs naturally. Drooping black raspberry stems tend to root in contact with the soil. The croton, a tropical plant, is commonly propagated by wrapping sphagnum enclosed in plastic around a stem cut to induce rooting. After rooting, the stem is detached and planted. Though simple and effective, layerage is not normally adapted to large-scale nursery practices.

Cuttage, the use of a detached plant part to regenerate a missing part or parts to form a complete plant, is one of the most important methods of propagation. Many plant parts may be used; thus cuttings can be classified as root, stem, or leaf. Stem cuttings are the most common.

The ability of stems to regenerate missing parts is variable; consequently plants may be easy or difficult to root. The physiological basis for cuttings to form roots is due to an interaction of many factors. These include transportable substances in the plant itself (plant hormones such as auxin), carbohydrates, nitrogenous substances, vitamins, and substances not yet identified. Environmental factors such as light, temperature, humidity, and oxygen are important, as are age, position, and type of stem.

Although easy-to-root plants such as willow or coleus may be propagated merely by plunging a stem in water or moist sand, the propagation of difficult-to-root species is a highly technical process. To achieve success with difficult-to-root plants special care is taken to control the environment and encourage rooting. A number of growth regulators stimulate rooting. A high degree of success has been achieved with indolebutyric acid, a synthetic auxin that is applied to the cut surface. A number of materials known as rooting cofactors have been found that interact with auxin to further stimulate rooting.

Humidity control is particularly important to prevent death of the stem from desiccation (drying out) before rooting is complete. The use of an intermittent-mist system in propagation beds has proved to be an important means of improving success in propagation by cuttings. These operate by applying water to the plant for a few seconds each minute.

Grafting. Grafting involves the joining together of plant parts by means of tissue regeneration. The part of the combination that provides the root is called the stock; the added piece is called the scion. When more than two

parts are involved, the middle piece is called the interstock. When the scion consists of a single bud the process is called budding. Grafting and budding are the most widely used of the vegetative propagation methods.

Grafting has uses in addition to propagation. The interaction of rootstocks may affect the performance of the stock through dwarfing or invigoration, and in some cases may affect quality. Further, the use of more than one component makes it possible to affect disease resistance and hardiness of the combination. Other uses include variety change (topworking) and repair (inarching, bridging, and bracing).

Grafting as a means of growth control is used extensively with fruit trees and in ornamentals such as roses and junipers. Fruit trees are normally composed of a "scion" cultivar grafted onto a rootstock. Sometimes an interstock is included between the scion and stock. The rootstock may be grown from seed (seedling rootstock) or asexually propagated (clonal rootstock). In the apple, a great many clonal rootstocks are available to give a complete range of dwarfing; rootstocks are also available to invigorate growth of the scion cultivar.

How grafting works

Stock cambium and scion cambium respond to being cut by forming masses of cells (callus tissues) that grow over the injured surfaces of the wounds. The union resulting from interlocking of the callus tissues is the basis of graftage. In dicots (*e.g.*, most trees) cambium—a layer of actively dividing cells between xylem (wood) and phloem (bast) tissues—is usually arranged in a continuous ring; in woody members, new layers of tissue are produced annually. Monocot stems (*e.g.*, lilacs, orchids) do not possess a continuous cambium layer or increase in thickness; grafting is seldom possible.

The basic technique in grafting consists in placing cambial tissues of stock and scion in intimate association, so that the resulting callus tissue produced from stock and scion interlocks to form a living continuous connection. A snug fit may be obtained through the tension of the split stock and scion or both. Tape, rubber, and nails may be used to achieve close contact. In general, grafts are only compatible between the same or closely related species. Success in grafting depends on skill in achieving a snug fit. Warm temperatures (80–85° F, 27–30° C) increase callus formation and improve "take" in grafting. Thus grafts using dormant material are often stored in a warm, moist place to stimulate callus formation.

In grafting or budding, the rootstock may be grown from seed or propagated asexually. Within a year a small amount of scion material from one plant can produce hundreds of plants. Some methods of grafting and budding are illustrated in Figure 1.

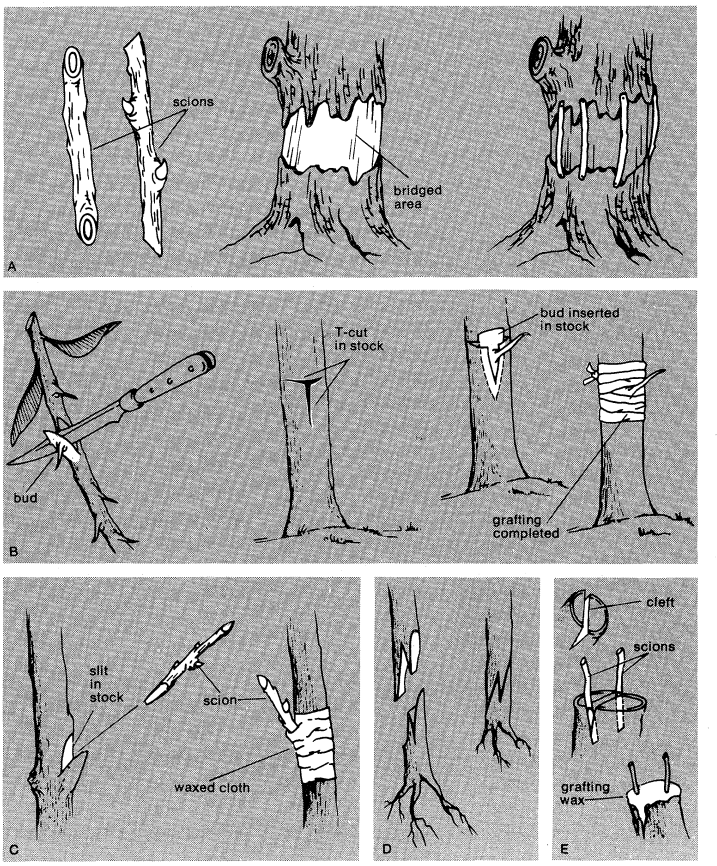


Figure 1: *Methods of grafting and budding.*
(A) Bridge graft of a damaged tree. (B) Budding. (C) Side graft. (D) Whip-and-tongue graft. (E) Cleft graft.
From *Plant Science: An Introduction to World Crops* by Jules Janick, Robert W. Schery, Frank W. Woods, and Vernon W. Ruttan. W.H. Freeman and Company. Copyright © 1969

Tissue culture. Tissue-culture techniques utilizing embryos, shoot tips, and callus can be used as a method of propagation. The procedure requires aseptic techniques and special media to supply inorganic elements; sugar; vitamins; and, depending on the tissue, growth regulators and organic complexes such as coconut milk, yeast, or amino acid extract.

Embryo culture has been used to produce plants from embryos that would not normally develop within the fruit. This occurs in early ripening peaches or in some hybridization between species. Embryo culture can also be used to circumvent seed dormancy.

A shoot tip, when excised and cultured, may produce roots at the base. This technique is employed for the purpose of producing plants free of disease. Certain orchids are rapidly multiplied by this method. Cultured shoot tips form an embryo-like stage that may be sectioned indefinitely to build up large stocks rapidly. These bulblike bodies left unsectioned develop into small plantlets. A similar procedure is used with the carnation, in which the shoot tip forms a cell mass that may be subdivided.

Callus-tissue culture—a very specialized technique that involves growth of the callus, followed by procedures to induce organ differentiation—has been successful with a number of plants including carrot, asparagus, and tobacco. Used extensively in research work, callus culture has not been considered a practical method of propagation. Callus culture produces genetic variability because in some cases cells double their chromosome number. Recently, in rice and tobacco, mature plants have been obtained from callus formed from pollen. These plants have half the normal number of chromosomes, the haploid number, instead of the diploid number.

ENVIRONMENTAL CONTROL

The intensive cultivation practiced in horticulture relies on extensive control of the environment for all phases of plant life. The most basic environmental control is achieved by location and site: sunny or shady sites, proximity to bodies of water, altitude, and latitude.

Structures. Various structures are used for temperature control. Cold frames, used to start plants before the normal growing season, are low enclosed beds covered with a removable sash of glass or plastic. Radiant energy passes through the transparent top and warms the soil directly. Heat, however, as long wave radiation, is prevented from leaving the glass or plastic cover at night. Thus heat that builds up in the cold frame during the day aids in warming the soil, which releases its heat gradually at night to warm the plants. When supplemental heat is provided, the structures are called hotbeds. At first, supplemental heat was supplied by respiration through the decomposition of manure or other organic matter. Today, heat is provided by electric cables, steam, or hot-water pipes buried in the soil.

Greenhouses

Greenhouses are large hotbeds, usually steam heated. While they were formerly made of glass, plastic films are now extensively used. Modern greenhouse ranges usually have automatic temperature control. Summer temperatures may be regulated by shading or evaporative "fan and pad" cooling devices. Air-conditioning units are usually too expensive except for scientific work. Greenhouses with precise environmental controls are known as phyto-

trons. Other environmental factors are controlled through automatic watering, regulation of light and shade, addition of carbon dioxide, and fertility regulation.

Shade houses are usually walk-in structures with shading provided by lath or screening. Summer propagation is often located in shade houses to reduce excessive water loss by transpiration.

Temperature control. A number of temperature control techniques are used in the field, including hot caps, cloches, plastic tunnels, and mulches of various types. Hot caps are cones of translucent paper or plastic placed over the tops of plants in the spring. These act as miniature greenhouses. In the past small glass sashes called cloches were placed over rows to help keep them warm. Polyethylene tunnels with wire hoops are now used for the same purpose. As spring advances the tunnels are slashed to prevent excessive heat buildup. In some cases the plastic tunnels are constructed so they may be opened and closed when necessary. This technique is widely used in Israel for early production of vegetables.

Mulches, insulating substances spread over the surface of the soil, regulate soil temperature, conserve soil moisture, and control erosion and weeds. In many cases mulches are used for their pleasing effect as background for flowers and shrubs in ornamental plantings. Plastic, either clear or black, may be used. The temperature-stabilizing effect of summer mulch is due to insulation, heat absorption, and shading. The surfaces of bare, dark-coloured soils may be 30° F (16.7° C) higher than air temperature. Mulches may also be applied in the fall to temper low winter temperatures. The insulation of the mulches conserves ground heat and also stabilizes and buffers soil temperature to prevent recurring freezing and thawing that rips and injures plant roots through soil heaving. Winter mulches protect roses and strawberries in high latitudes.

The storage of perishable plant products is accomplished largely through the regulation of their temperature to retard respiration and microbial activity. Excess water loss may be prevented by controlling humidity. Facilities that utilize the temperature of the atmosphere are called common storage. The most primitive types take advantage of the reduced temperature fluctuations of the soil by using caves or unheated cellars. Above-ground structures must be insulated and ventilated. Complete temperature-regulated storages utilizing refrigeration and heating are now common for storage of horticulture products. The regulation of oxygen and carbon dioxide levels along with the regulation of temperature is known as controlled-atmosphere storage. Rooms are sealed so that gaseous exchange can be effectively controlled. Many horticultural products, such as fruit, can be kept fresh for as long as a year under these controlled conditions.

Frost control. Frost, a thin layer of ice crystals deposited on soil and plant surfaces when temperatures drop to freezing, may destroy many horticultural crops and plants. Frost is especially damaging to perennial fruit crops in the spring—because flower parts are sensitive to freezing injury—and to tender transplants. The two weather conditions that produce freezing temperatures are rapid radiational cooling at night and introduction of a cold air mass with temperatures below freezing. Radiation frost occurs when skies are clear and calm; air mass freezes occur when skies are overcast and windy.

Frost control methods involve either reduction of radiational heat loss or conservation or addition of heat. Radiational heat loss may be reduced by hot caps, cold frames, or mulches. Removing weeds that shade the soil increases the amount of heat stored during the day. Clean cultivated areas are less susceptible to frost than sodded or mulched areas. Heat may also be added from the air. Wind machines that stir up the air, for example, provide heat when temperature inversions trap cold air under a layer of warm air. These have been used extensively in citrus groves. Heat may be added directly by using heaters, usually fuelled with oil. Sprinkler irrigation can also be used for frost control. The formation of ice is accompanied by the release of large amounts of heat, which maintains plants at the freezing temperature as long as the water is being frozen. Thus continuous sprinkling during frosty nights has been used to protect strawberries from frost injury.

Frost injury to transplants may be prevented through processes that increase the plant's ability to survive the impact of unfavourable environmental stress. This is known as hardening off. Hardening off of plants prior to transplanting may be accomplished by withholding water and fertilizer, especially nitrogen. This prevents formation of succulent tissue that is very frost tender. Gradual exposure to cold is also effective for hardening. Induced cold resistance in crops such as cabbage is considerable; unhardened cabbages show injury at 28° F (−2.2° C), while hardened plants withstand temperatures as low as 22° F (−5.6° C).

Light control. Light has a tremendous effect on plant growth. It provides energy for photosynthesis, the process by which plants, with the aid of the pigment chlorophyll, synthesize carbon compounds from water and carbon dioxide. Light also influences a great number of physiological reactions in plants. At energy values lower than those required for photosynthesis, light affects such processes as dormancy, flowering, tuberization, and seed-stalk development. In many cases these processes are affected by the length of day (photoperiodicity).

The control of light in horticultural practices involves increasing energy values for photosynthesis and controlling day length. Light is controlled in part by site and location. In the tropics day length approaches 12 hours throughout the year, whereas in polar regions it varies from 0 to 24 hours. Light is also partly controlled by plant distribution and density.

Supplemental illumination in greenhouses increases photosynthesis. Cost of power, however, makes this impractical for all but crops of the highest value. Fluorescent lights are the most efficient for photosynthesis; special lights, rich in the wave lengths required, are now available.

Extension of day length through supplemental illumination and shading is common practice in the production of greenhouse flower crops, which are often induced to flower out of season. Artificial lengthening of short days, or interruption of the dark period, promotes flowering in long-day plants such as lettuce and spinach and prevents flowering of short-day plants such as chrysanthemums. Similarly, during naturally long days, shading to reduce day length prevents flowering of long-day plants and promotes flowering of short-day plants. The manipulation of day length is standard practice to control flowering of greenhouse chrysanthemums throughout the year. Tungsten lights have proven effective for extending day length because they are rich in the red end of the spectrum that affects the photoperiodic reaction. Extending the day length is an economical practice because only a low light intensity is required. The same effects can be obtained through interruption of the dark period, even with light flashes. Decreasing day length is usually accomplished by covering the plants with black shade cloth.

TRAINING AND PRUNING

Training, the orientation of the plant in space, is achieved by physical techniques that direct the shape, size, and direction of plant growth. Training may be accomplished by providing supports to which plants can be bent, twisted, or fastened. Training is often associated with pruning, the judicious removal of plant parts. Pruning, however, is performed for other purposes: for example, to adjust fruit load and to regulate size and quality. Altering form or size can improve appearance or usefulness. Trees and shrubs can be trained to a variety of shapes limited only by the imagination.

Topiary and bonsai

The training of plants to grow in unnatural shapes for ornamental purposes is called topiary. In Roman times and during the Renaissance, when topiary was in high fashion, plants were trained to unusual and fantastic shapes such as beasts, ships, and building facades. Though this type of topiary is no longer in vogue, hedges and

shrubs are still trained to geometric shapes in formal gardens.

Another extreme form of training is the Japanese art of bonsai, the creation of dwarfed potted trees by a combination of pruning (both roots and tops) and restricted nutrition. Living trees over a hundred years old and only a few feet high are grown in special containers arranged to resemble the natural landscape.

Particular spatial arrangements may increase light utilization, facilitate harvesting or disease control, or improve productivity and quality. Thus, training and pruning form an essential part of fruit growing throughout the life of the plant. Special attention is given in the formative years to obtain desired shape and structure. The key to training is the point on the main stem from which branches form. In the central-leader system of training the trunk forms a central axis with branches distributed laterally up and down and around the stem. In the open centre or vase system the main stem is terminated and growth forced through a number of branches originating close to the upper end of the trunk. An intermediate system is called the modified-leader system. In espalier systems, plants are trained to grow flat along a wire or trellis (see Figure 2). Properly executed espaliers are extremely

Pruning systems

By courtesy of (B,C) Henry Leuthardt Nurseries; from (A) *Arnoldia* (August 29, 1969)

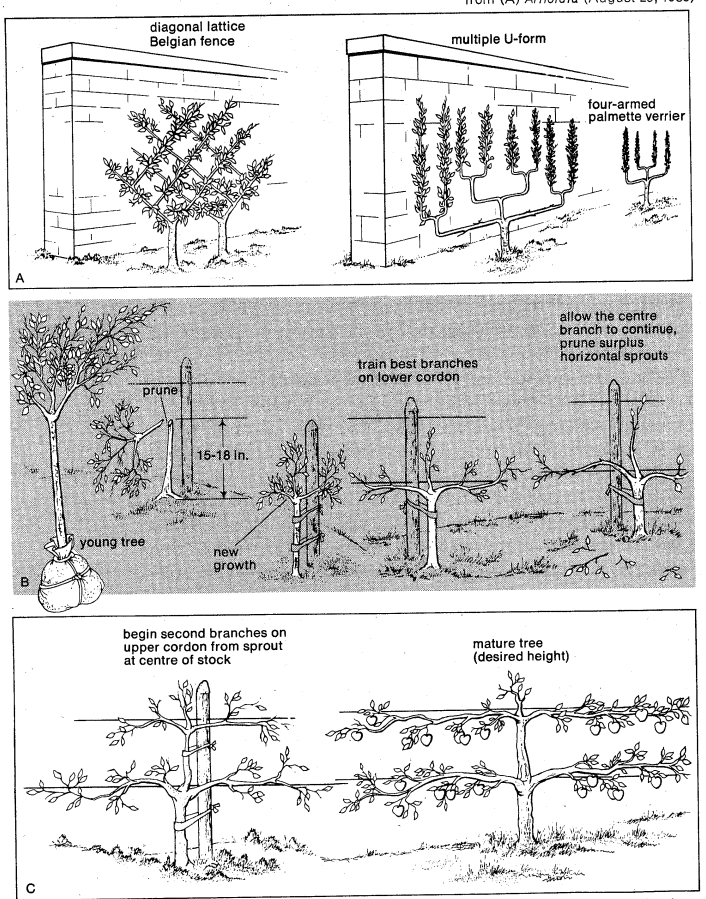


Figure 2: *The espalier technique for training fruit trees and shrubs.* (A) Examples of espalier patterns. (B) Planting and training an espalier. (C) Final shaping and (right) tree in its finished form, a double horizontal cordon.

attractive as ornamentals. Espaliers in combination with dwarfing rootstocks allow high density orchards that are very productive on a per unit area basis with the fruit close to the ground for easy harvest. Extensive pruning is required annually to maintain the system.

There are a number of physiological responses to training and pruning. Orientation of the plant may have a marked effect on growth and fruiting. Thus fruit trees planted on an inclined angle become dwarfed and flower earlier; training branches in a horizontal position produces the same effect. This effect is achieved naturally when a heavy fruit load bends a limb down. The main effects of pruning are achieved by altering the root–shoot balance. Thus an explosion of vegetative growth normally occurs after extensive shoot pruning. Severely pruned plants, especially if they are in the juvenile stage of growth, tend to remain vegetative. Similarly the slowdown of vegetative growth by root pruning encourages flowering.

The two basic pruning cuts are known as "heading back" and "thinning out." Heading back consists in cutting back the terminal portion of a branch to a bud; thinning out is the complete removal of a branch to a lateral or main trunk. Heading back, usually followed by the stimulation of lateral budbreak below the cut, produces a bushy, compact plant, suitable for a hedgerow. Thinning out, which encourages longer growth of the remaining terminals by reducing lateral branches, tends to open up the plant, producing a longer plant.

GROWTH REGULATION BY CHEMICALS

Control of plant growth through growth-regulating materials is a modern development in horticulture. These materials have resulted from basic investigations into growth and development, as well as systematic screening of materials to find those that affect differentiation and growth. This field was given great impetus by the discovery of a class of plant hormones known as auxins, which affect cell elongation.

Auxins have been correlated with inhibition and stimulation of growth as well as differentiation of organs and tissues. Such processes as cell enlargement, leaf and organ separation, budding, flowering, and fruit set and growth are influenced by auxin. In addition, auxin has been associated with the movement of plants in response to light and gravity. Auxin materials are used in horticulture for the promotion of rooting, fruit setting, fruit thinning, and fruit-drop control.

Gibberellins are a group of related, natural-occuring compounds of which only one, gibberellic acid, is commercially available. Gibberellins have many effects on plant development. The most startling is the stimulation of growth in many compact or dwarf plants. Minute applications transform bush to pole beans or dwarf to normal corn. Perhaps the most widespread horticultural use has been in grape production. The application of gibberellin is now a regular practice for the culture of the "Thompson seedless" cultivar ("Sultanina") of grapes to increase berry size. In Japan applications of gibberellic acid are used to induce seedlessness in certain grapes.

Cytokinins are a group of chemical substances that have a decisive influence on the stimulation of cell division. In tissue culture high auxin and low cytokinin give rise to root development; low auxin and high cytokinin give rise to shoot development.

Dormins have recently been established as a class of inhibitory compounds. Abscisic acid is the only naturally occurring dormin known, but similar synthetic substances have been reported.

Ethylene, a hydrocarbon compound, acts as a plant hormone to stimulate fruit ripening as well as rooting and flowering of some plants. An ethylene-releasing compound, 2-chloroethylphosphonic acid, has many horticultural applications, of which the most promising may be uniform ripening of tomato and the stimulation of latex flow in rubber.

Many compounds that inhibit growth hormones have application in horticulture. For example, a number of materials that inhibit formation of gibberellins by the plant act to dwarf plants. These include chlorinated derivatives of quaternary ammonium and phosphonium compounds. Many of these have applications in floriculture. Growth retardants such as succinic acid-2,2-dimethylhydrazide, a gibberellin supressor, have applications in horticulture from a wide array of effects that include dwarfing and fruit maturity. The growth inhibitor maleic hydrazide has been effective in preventing the sprouting of onions and potatoes.

The flowering regulating hormone (florigen) is known to

exist but as yet remains undetected. The isolation of this growth regulator will have great impact on horticulture.

SOIL MANAGEMENT

The soil must be managed for fertility (the ability to provide plant nutrients) and physical condition. Nutrients necessary for growth must be provided and released in forms readily available to the plant. Soil management is also concerned with careful and sustained land use through control of erosion. Thus conservation becomes an essential part of soil management.

Sixteen elements are essential for plant growth. Three of these, carbon, oxygen, and hydrogen, are provided through water and air; the other 13 are provided through the soil. The elements required in relatively large amounts are called major elements—nitrogen, phosphorus, potassium, calcium, magnesium, and sulfur. The minerals required in small quantities are called trace elements—iron, boron, manganese, zinc, molybdenum, copper, and chlorine.

Fertilizers

Soil fertility, or its capacity to supply nutrients, involves the amount and availability of plant nutrients. Addition of nutrients to plants is called fertilization; the materials that supply nutrient elements are called fertilizers. Fertilizers may be classified as natural organics or chemical. In the past most fertilizers were waste organic materials such as manure, crop residues, blood, or fish scraps. Many fertilizers today are chemicals synthesized from inorganic materials. Fertilizers are usually applied to the soil but may also be applied through leaves or stems. Carbon dioxide added to the air in greenhouses is also considered to be a fertilizer. Fertilizers that supply nitrogen and phosphorus and potassium, the elements that are usually the most limiting to plant growth, are referred to as "complete."

The nutrient requirement of plants can be determined by correlating plant responses with the mineral content of plant and soil through leaf or tissue analysis and soil testing. Although severe shortage of certain nutrients can be diagnosed through characteristic deficiency symptoms, in good commercial practice nutrient shortages are not permitted to become so severe.

Fertilizers may be applied in solid, liquid, or gaseous form. Some are dissolved in irrigation water or applied directly to the foliage. Nitrogen can be added to the soil as ammonia gas. Proper placement and timing are important factors in efficient fertilization practices. Timing (*e.g.*, season) may be critical. Overfertilizing is a problem that may lead to excessive salt concentrations and destruction of tender plant parts (fertilizer burn).

Soil reaction, the degree of acidity or alkalinity expressed as pH, has tremendous indirect effects on plant growth. Abnormally high or alkaline pH (above 9) or low pH (below 4) is in itself toxic to plants. Between these extremes, the effect is usually on nutrient availability. Thus, at high pH, iron deficiency often results because insoluble iron compounds form that are unavailable to plants. In addition, plants vary in their reaction to pH. Most plants thrive best in conditions close to neutral (pH 7). Plants that prefer acid soil include rhododendrons, gardenias, azaleas, camellias, cranberries, and blueberries.

Soil reaction may be modified. Soils can be made more alkaline (pH increased) by adding calcium, magnesium, sodium, or potassium. Calcium is the most efficient; it has other beneficial effects and is an essential element. Addition of calcium or compounds of calcium and magnesium to reduce acidity is called liming. Soil may be made more acid by the addition of hydrogen. This is accomplished by adding substances that produce strong acids in the soil. Addition of compounds that contain sulfur is the most efficient method. Thus fertilizers containing sulfur such as ammonia sulfate or superphosphate are used to lower pH.

In addition to the nutrient level and soil reaction, the structure of the soil plays a significant role in soil management because of its effect on water-holding capacity and aeration of the soil. Soil structure also influences root growth and soil micro-organisms, which are important in making nutrients available to plants. The physical condition of soil is largely determined by organic matter formed by decomposition of plant and animal residues. Organic matter decays when added to soil but leaves fractions that resist decomposition to form a dark, colloidal substance called humus. Humus has high adsorbent and absorptive properties for nutrients and moisture that greatly affect soil condition and fertility. Thus practices that increase or maintain soil organic matter are essential to sound horticulture. Organic matter can be applied directly by addition of natural organic fertilizers such as manures or through the plowing down of plant material such as a growing crop or grass sod (green manure). The extended use of legumes or grass sod is the best way to increase soil organic matter. The physical properties of potting soils can be controlled by direct incorporation of organic matter such as peat or through addition of inert substances such as vermiculite (expanded mica) or perlite (expanded volcanic lava).

Hydroponic gardening

Although soil is the most common medium for the growth of plants and thus of vital concern to horticulturists, it is not essential to plant growth. Plants can be grown quite well in nutrient solutions alone or in sand or gravel, to which nutrient solutions are added. Such soilless cultivation is known as hydroponics. Hydroponic gardens have been used to provide fresh vegetables where soil is unavailable, chiefly in coral islands. Hydroponic gardens were so used during World War II to provide military personnel with fresh vegetables. Though hydroponics is still practiced on a limited scale, in most situations soil has proven the more versatile and economic medium. Hydroponic culture is used extensively in experiments if precise nutritional control is required.

Water management. Water management in horticulture is concerned with the regulation and use of water in plant growth. This involves irrigation, the addition of supplemental water, and drainage, the removal of excess water. Because of the close association between water and soil, the control of moisture is an integral part of soil management.

There are a number of general methods of land irrigation. In surface irrigation water is distributed over the surface of soil. Sprinkler irrigation is application of water under pressure as simulated rain. Subirrigation is the distribution of water to soil below the surface; it provides moisture to crops by upward capillary action. A new method called trickle irrigation involves the slow release of water to each plant through small plastic tubes. This technique is adapted both to field and to greenhouse conditions.

Removal of excess water from soils may be achieved by surface or subsurface drainage. Surface drainage refers to the removal of surface water by development of the slope of the land utilizing systems of drains to carry away the surplus water. In subsurface drainage open ditches and tile fields intercept ground water and carry it off. The water enters the tiling through the joints, and drainage is achieved by gravity feed through the tiles. (See also IRRIGATION AND DRAINAGE.)

PEST CONTROL

Horticultural plants are subject to a wide variety of injuries caused by other organisms. These injuries are called diseases when caused by some infectious agent (the pathogen) living in intimate association with the plant for an extended period of time. Injuries caused by insects and animals are often classed as damage rather than disease, and in such instances the term predator rather than pathogen is used. Harmful effects of other plant species or weeds result from competition for essential growth factors such as light, nutrients, or water.

Plant pests include viruses, bacteria, fungi, higher plants, nematodes, insects, mites, birds, and rodents. The control of plant pests can be considered a discipline in itself. In Europe it is called plant protection. In the U.S. pest control is usually divided into three disciplines: plant pathology, entomology, and weed control.

There are various methods used to control plant pests. The most successful treatments are preventive rather than curative, aiming to deter injury rather than cure affected plants. Successful control often depends on the proper use of many methods. Expensive methods of pest control are only feasible with high-value products. Thus horticulture bears a disproportionate share of the cost of pest control.

Control of pests is achieved through practices that prevent harm to the plant or methods that affect the plant's ability to resist or tolerate intrusion by the pathogen. These can be classified as cultural, physical, chemical, or biological.

Cultural practices that reduce effective pest population include the elimination of diseased or infected plants or seeds (roguing), cutting out of infected plant parts (surgery), removal of plant debris that may harbor pests (sanitation), and alternating crops unacceptable to the pest (rotation). Any of a number of techniques may be employed to render the environment unfavourable to the pest, such as draining land, flooding, pH change. Safeguards against the spread of pests between geographic areas may be enforced through quarantines.

Physical methods may be used to protect the plant against intrusion or to eliminate the pest entirely. Physical barriers range from the traditional garden fence to bags that protect each fruit, a common practice in Japan. Heat treatment is used to destroy some seed-borne pathogens and is a standard soil treatment in greenhouses to eliminate soil pests such as fungi, nematodes, and weed seeds. Cultivation and tillage are standard practices for weed control.

The horticultural industry is now dependent upon chemical control of pests through pesticides, materials toxic to the pest in some stage of its life cycle. Commercial growers of practically all horticultural crops rely on complete schedules utilizing many different compounds. Pesticides are usually classed according to the organism they control: for example, bactericide, fungicide, nematicide, miticide, insecticide, rodenticide, and herbicide.

Pesticide selectivity

Selectivity of pesticides, the ability to discriminate between pests, is a relative concept. Some nonselective pesticides kill indiscriminately; most are selective to some degree. Most fungicides, for example, are not bacteriocidal. The development of highly selective herbicides makes it possible to destroy weeds from crops selectively. Selectivity may be achieved through control of dosage, timing, and method of application.

Plant pests may also be controlled through the manipulation of biological factors. This may be achieved through directing the natural competition between organisms or by incorporating natural resistance to the whole plant. The introduction of natural parasites or predators has been a successful method for the control of certain insects and weeds. Incorporation of genetic resistance is an ideal method of control. Thus breeding for disease and insect resistance is one of the chief goals of plant breeding programs. A major obstacle to this method of control is the ability of pathogens (disease-producing organisms) to mutate easily and attack heretofore resistant plants. (See also PEST CONTROL; WEED CONTROL.)

PLANT BREEDING

The isolation and production of superior types known as "cultivars" is the very keystone of horticulture. Plant breeding, the systematic improvement of plants through the application of genetic principles, has placed improvement of horticultural plants on a scientific basis. The raw material of improvement is found in the great variation that exists between cultivated plants and related wild species. The incorporation of these changes into cultivars adapted to specific geographical areas requires both a knowledge of the theoretical basis of heredity and art and skill to discover, perpetuate, and combine these small but fundamental differences in plant material.

The goal of the plant breeder is to create superior crop varieties. The cultivated variety, or cultivar, may be defined as a group of crop plants having similar but distinguishable characteristics. The term cultivar has various meanings, however, depending on the mode of reproduction of the crop. With reference to asexually propagated crops, the term cultivar means any particular clone considered of sufficient value to be graced with a name. With reference to sexually propagated crops, the concept of cultivar depends on the method of pollination. The cultivar in self-pollinated crops is basically a particular homozygous genotype, a pure line. In cross-pollinated crops the cultivar is not necessarily typified by any one plant but sometimes by a particular plant population, which at any one time is composed of genetically distinguishable individuals. (See PLANT BREEDING.)

Ornamental horticulture

Ornamental horticulture consists of floriculture and landscape horticulture, each of which is divided into the plant growing and marketing and the associated activities of flower arrangement and landscape design. The turf industry is also considered a part of ornamental horticulture.

Although flowering bulbs and flower seed represent an important component of agricultural production for the Low Countries of Europe, ornamentals are relatively insignificant in world trade. As a result world statistics on production are unavailable.

FLORICULTURE

Floriculture is the growing of cut flowers, potted plants, and associated greenery, including their subsequent arrangement. Floriculture has long been an important part of horticulture, especially in Europe and Japan, and accounts for about half of the nonfood horticultural industry in the U.S. Because cut flowers and potted plants are largely produced in plant-growing structures in temperate climates, floriculture is largely thought of as a greenhouse industry; there is, however, considerable outdoor culture of many flowers. Floriculture is a competitive and highly technical business requiring knowledge, skill, and large amounts of capital. Greenhouse floriculture is in many respects the most sophisticated part of plant agriculture.

The industry is usually very specialized with respect to its crop; the grower must provide precise environmental control and be vigilant in the constant struggle against pests and diseases. Exact scheduling is imperative since most floral crops are seasonal in demand. Because the product is perishable, transportation to market must function smoothly to avoid losses.

The floriculture industry involves the grower, who mass-produces flowers for the wholesale market, and the retail florist, who markets to the public and contributes services such as arrangement and delivery. The grower is often a family farm, but, as in all modern agriculture, the size of the growing unit is increasing. There is a movement away from urban areas of high taxes and labour costs to locations having lower tax rates and a rural labour pool and also toward more favourable climatic regions (milder temperature and more sunlight). The development of air freight has emphasized interregional and international competition. Flowers can be shipped long distances by air and arrive in fresh condition to compete with locally grown products.

The retail florist is usually separate from the grower. Typically he buys wholesale cut flowers, plants, greenery, and associated products and makes up floral arrangements. Most typically he is a designer and businessman rather than a breeder or grower.

LANDSCAPE HORTICULTURE

The industry of landscape horticulture is divided into growing, maintenance, and design. Growing of plants for landscape is called the nursery business, although a nursery refers broadly to the growing and establishment of any young plants before permanent planting. The nursery industry involves production and distribution of woody and herbaceous plants and is often expanded to include ornamental bulb crops—corms, tubers, rhizomes,

and swollen roots as well as true bulbs. Production of cuttings to be grown in greenhouses or for indoor use (foliage plants), as well as the production of bedding plants, is usually considered part of floriculture, but this distinction is fading. While most nursery crops are ornamental, the nursery business also includes fruit plants and certain perennial vegetables used in home gardens, for example, asparagus and rhubarb. Christmas tree production is also included, though the growth of other seedling trees is usually considered part of forestry.

Next to ornamental trees and shrubs, the most important nursery crops are fruit plants, followed by bulb crops. The most important single plant grown for outdoor cultivation is the rose. The type of nursery plants grown depends on location; in general (in the Northern Hemisphere) the northern areas provide deciduous and coniferous evergreens, whereas the southern nurseries provide tender broadleaf evergreens.

The nursery industry includes wholesale, retail, and mail-order operations. The typical wholesale nursery specializes in relatively few crops and supplies only retail nurseries or florists. The wholesale nursery deals largely in plant propagation, selling young "lining out" stock of woody material to the retail nursery, which cares for the plant until growth is complete. Many nurseries also execute the design of the planting in addition to furnishing the plants.

Bulb crops. The bulb crops include plants such as the tulip, hyacinth, narcissus, iris, day lily, and dahlia. Included also are nonhardy bulbs used as potted plants indoors and summer outdoor plantings such as amaryllises, anemones, various tuberous begonias, caladiums, cannas, dahlias, freesias, gladioli, tiger flowers, and others. Hardy bulbs, those left in the soil over winter, include various crocuses, snowdrops, lilies, daffodils, tulips, and many others.

Many bulb crops are of ancient Old World origin, introduced into horticulture long ago and subjected to selection and crossing through the years to yield many modern cultivars. One of the most popular is the tulip. Tulips are widely grown in gardens as botanical species but are especially prized in select forms of the garden tulip (which arose from crosses between thousands of cultivars representing several species). Garden tulips are roughly grouped as early tulips, breeder's tulips, cottage tulips, Darwin tulips, lily-flowered tulips, triumph tulips, Mendel tulips, parrot tulips, and others. The garden tulips seem to have been developed first in Turkey but were spread throughout Europe and were adopted enthusiastically by the Dutch. The Netherlands has been the centre of tulip breeding ever since the 18th century, when interest in the tulip was so intense that single bulbs of a select type were sometimes valued at thousands of dollars. The collapse of the "tulipmania" left economic scars for decades. The Netherlands remains today the chief source of tulip bulbs planted in Europe and in the U.S. The Netherlands has also specialized in the production of related bulbs in the lily family and provides hyacinths, narcissus, crocus, and others. The Dutch finance extensive promotion of their bulbs to support their market. Years of meticulous growing are required to yield a commercial tulip bulb from seed. Thorough soil preparation, high fertility, constant weeding, and careful record keeping are part of the intensive growing in The Netherlands, employing much hand labour. Bulbs sent to market meet specifications as to size and quality, which assure at least one year's bloom even if the bulb is supplied nothing more than warmth and moisture. The inflorescence (flower) is already initiated and the necessary food stored in the bulb. Under less favourable maintenance than prevails in The Netherlands, a subsequent year's bloom may be smaller and less reliable; it is not surprising that tulip-bulb merchants suggest discarding bulbs after one year and replanting with new bulbs.

Tulips

Herbaceous perennials. Garden perennials include a number of herbaceous species grown for their flowers or occasionally used as vegetative ground covers. Under favourable growing conditions the plants persist and increase year after year. The biggest drawback to perennials as compared with annuals is that they must be maintained throughout the growing season but have only a limited flowering period. Typical perennials are hollyhocks, columbines, bellflowers, chrysanthemums, delphiniums, pinks, coralbells, phlox, poppies, primroses, and speedwells.

Perennials are often produced and sold as a sideline to other nursery activities; some are sold through seedhouses. Perennial production could be undertaken on a massive scale, with attendant economies, but the market is neither large enough nor predictable enough (except for the greenhouse growing of such cut flowers as chrysanthemums and carnations) to interest most growers. In northern climates, many varieties are mulched in autumn to protect the plant from winter kill.

Shrubs. Production of ornamental shrubs is the backbone of the nursery trade in Europe and the United States. The nursery business is about equally divided among the production of (1) coniferous evergreens such as yew, juniper, spruce, and pine; (2) broad-leafed evergreens such as rhododendron, camellia, holly, and boxwood; (3) deciduous plants such as forsythia, viburnum, barberries, privets, lilacs, and flowering vines; and (4) roses, which will be discussed separately.

Fields of specialization have evolved within the ornamental shrub industry. Some firms confine activity mostly to production of "lining out" stock—seedlings and rooted cuttings for sale to growers for field plantings that must be tended several years before reaching salable size. Such firms will generally offer not only the bare root liners themselves but small plants started in plant bands or slightly larger ones in containers. They may also offer small grafted plants of expensive or rare cultivars. Lining out stock of a typical evergreen, such as the pfitzer juniper, is relatively inexpensive; it is more economical for a grower to purchase these ready to plant in the field than to undertake specialized propagation himself.

The field grower may, in turn, specialize in mass growing for the wholesale trade only. The field plantings are tended until they attain marketable size, having been shaped and sheared as necessary, provided with pest control, irrigated, culled, and so on. Because of the time required to produce a marketable crop and because of rising labour costs, this phase of the nursery industry involves economic hazards. But wholesale growing escapes the high overhead of retail marketing in urban areas, and although many growers do sell stock at the nursery, they generally avoid the expensive merchandising required of the typical urban-area garden centre. Growers are especially interested in labour-saving technology and are turning to herbicidal control of weeds and shortcut methods for transplanting (as a substitute for hand digging and burlapping).

Most shrubs are marketed "balled and burlapped" (B & B). The plants are dug with soil and wrapped with burlap or other suitable material. Recently, a trend has developed toward container-grown stock—nursery stock grown in the container in which it is sold. This allows year-round sales of plant material. Deciduous shrubs may be marketed bare root. There are obvious economies in not having to dig and wrap the plant carefully and handle added weight and soil volume. Fairly satisfactory techniques have been developed for holding bare-root ornamental shrubs in viable condition until marketed. Usually this involves digging after leaf fall in autumn and storing the plant through winter in cold cellars with regulated humidity for spring sale. Unfortunately, plants are often handled carelessly. Though the chief cause of loss during storage is desiccation, respiration also plays a part when temperatures rise. But even if the plants are properly cared for during cold storage, they often suffer from warming and drying at the sales outlet.

Storage and marketing

Roses. The production of roses is probably the most specialized of all shrub growing; the grower often deals solely in rose plants. Most are bud grafted onto rootstocks (typically *Rosa multiflora*). This is the only way

to achieve rapid and economical increase of a new selection to meet market demands. Large-scale production of roses has tended to centre in areas where long growing seasons make rapid production possible.

Because the budding operation calls for skilled hand labour and because field maintenance is expensive, few economies can be practiced in the production of roses. But techniques have been developed for the distribution of roses that do offer certain economies, such as coated paper or plastic bags instead of damp moss to retain humidity and wax coating of stems of dormant stock to inhibit desiccation.

Trees. Ornamental shade trees are usually grown and marketed in conjunction with shrubs. As part of the mid-20th-century migration to the suburbs in many countries and the construction of houses on cleared land, shade trees have become an increasingly important part of the nursery picture. As interest in shade and ornamental trees increased, creation of improved cultivars followed. There is still some activity in transplanting native trees from the woodlot, and some trees are still grown from genetically unselected seed or cuttings; but more and more trees, like roses and shrubs before them, are vegetatively propagated as named cultivars, many patented.

There are some obvious drawbacks to tree growing as an industry. The crop takes a fairly long time to reach marketable size, and trees not salable within a year or two become too large for convenient handling. The investment in a tree is high and consequently so is the price. Risks are entailed in trying to anticipate what the market will be for a crop planted today but not usable for five or six years. Yet so indispensable is an ornamental tree to attractive landscaping that this branch of horticulture continues to flourish.

The digging, handling, and transporting of trees has always been difficult. Except in the sapling stage, trees are seldom sold bare root. A good deal of experienced manpower is needed to dig a tree and wrap its root system in burlap, the traditional way of moving larger trees. Special machines with slings and hoists have been developed, and recently in the U.S. a split scoop was put into production that first digs a hole where the tree is to be planted, digs the tree so that the root ball is the exact size of the previously dug hole, and then transports the tree to the hole. Thus there is no need for wrapping or separate handling. Such a device makes it possible to move relatively large trees easily, but of course the tree must be contracted for in the field rather than selected from a nursery display.

LANDSCAPE DESIGN

The design and planning of landscapes has become a distinct profession that in many cases is only incidentally horticultural. Landscape architecture in its broadest sense is concerned with all aspects of land use for man. As a horticulturist, the landscape architect uses living plant materials along with other landscape materials—stone, mortar, wood—as elements of landscape design. Unlike the materials of the painter or sculptor, plants are not static but change seasonally and with time. The colour, form, texture, and line of plants are used as design elements in the landscape. Plant materials are also manipulated as functional materials to control erosion, as a surface material, and for enclosures to provide protection from sun and wind.

Landscape architecture originated in the design of great estates, and home landscape is still an integral part of landscape architecture. More recently, however, landscape architecture has begun to include larger developments such as urban and town planning, parks both formal and "wild," public buildings, industrial landscaping, and highway and roadside development. (See GARDEN AND LANDSCAPE DESIGN.)

TURF

The turfgrass industry produces and maintains specialized grasses and other ground covers for utility, recreation, and beautification. Grass is an ubiquitous feature of many urban and suburban facilities such as airports, cemeteries, educational institutions, golf courses, highways, residential and commercial buildings, and parks. In the U.S., annual turf maintenance expenditures exceed $4,000,000,000, of which 70 percent is spent on residential lawns.

The turfgrass industry reaches a large market throughout the affluent countries of the world. The maritime climate of the British Isles is particularly suited to beautiful turf. Although lawns are usually not the major home landscape feature, as in the U.S., public areas are even more emphasized.

The major turf species in Europe is *Poa annua*. In the U.S. the major species grown in the northern two-thirds of the country are three cool-season grass species: Kentucky bluegrass, the fine or red fescues, and bentgrass. There are a number of grass species adapted to the South: Bermuda grass, bahia grass, carpet grass, centipede grass, Zoysia grass, and St. Augustine grass.

Horticulture education and research

Scholarly works in horticulture appear continuously in scientific literature. Specific institutions devoted to horticultural research, however, go back to the beginning of the experiment-station system, the first being a private laboratory of John Bennet Lawes, with the later collaboration of Joseph Henry Gilbert, in Rothamsted, England (1834). Horticultural education and research in the U.S. was given great impetus by Morrill of the Land Grant–College Act (1862), which provided educational institutions in agricultural and mechanical arts for each state. Presently state experiment stations and the federal experiment stations of the U.S. Department of Agriculture, with its centre at Beltsville, Maryland, carry out systematic research efforts in horticulture. Although much research is carried out on horticultural food crops, there has been an increasing emphasis on ornamentals. Horticultural research is also conducted by private companies: for example, the seed industry, canning and processing firms, and finally private foundations and botanical gardens. Throughout the world there are now more than 1,000 institutes with 7,000 professional staff members conducting horticultural research.

Horticultural education is an established part of professional agricultural education worldwide. Training in horticulture up to the Ph.D. degree is offered in universities. There are relatively few schools devoted to the training of gardeners and horticultural technicians in the U.S., although a number of state universities have two-year programs in horticulture. Vocational horticultural training is more highly developed in Europe.

There are a great number of national and international societies devoted to horticulture. These include community organizations such as garden clubs, specialty organizations devoted to a particular plant or group of plants (*e.g.*, rose and orchid societies), scientific societies, and trade organizations. The first society devoted to horticulture originated in 1804 with the establishment in England of the Royal Horticultural Society. There are similar organizations in other European countries. The American Pomological Society was formed in 1848. The American Horticultural Society, established in 1945, is devoted largely to ornamentals. The American Society for Horticultural Science was established in 1903 and is now the most widely known scientific society devoted to horticulture. The International Horticultural Society for Horticultural Science, formed in 1958 with permanent headquarters in The Hague, The Netherlands, sponsors international congresses every four years. Most societies and horticultural organizations publish periodicals.

There are thousands of publications in the world devoted to some aspect of horticulture. The scientific and technical horticultural literature since 1930 is abstracted in *Horticultural Abstracts*, prepared by the Commonwealth Bureau of Horticulture and Plantation Crops, East Malling, Kent.

BIBLIOGRAPHY. J. JANICK, *Horticultural Science*, 2nd ed. (1972), an introduction to the scientific and technological as-

pects of modern horticulture; E.L. DENISEN, *Principles of Horticulture* (1958), a traditional approach to the subject; R.W. SCHERY, *Plants for Men*, 2nd ed. (1972), an economic text covering many crop plants; H.T. HARTMANN and D.E. KESTER, *Plant Propagation: Principles and Practices*, 2nd ed. (1968), the authoritative work in the field; L.R. HAWTHORN and L.H. POLLARD, *Vegetable and Flower Seed Production* (1954), on seed production technology; INTERNATIONAL SOCIETY FOR HORTICULTURAL SCIENCE, *Horticultural Research International*, 2nd ed. (1972), a world list of research stations and organizations; A. LAURIE, D.C. KIPLINGER, and K.S. NELSON, *Commercial Flower Forcing: The Fundamentals and their Practical Application to the Culture of Greenhouse Crops*, 7th ed. (1968), the standard floriculture text; A.A. HANSON and F.V. JUSKA (ed.), *Turfgrass Science* (1969), a comprehensive monograph on all aspects of the turf industry.

(J.J.)

Hospital

A hospital is an institution that is built, staffed, and equipped for the identification—diagnosis—of disease and the treatment, both medical and surgical, of the sick and the injured; for their housing during this process; and for certain other procedures, such as health examinations and the management of childbirth, that are ordinarily directed more to the preservation of health and physical well-being than to the cure of disease or of physical or mental abnormality.

Hospitals have long existed in every civilized country. In the developing countries, which contain much of the world's population, there are not enough hospitals, equipment, and trained staff, and, by the standards of the highly industrialized countries, the hospitals that are present are poorly equipped to handle the great volume of sick persons that need their care.

In many parts of the world, numbers of people fail to receive the benefits of modern medical science, do not enjoy public-health measures and hospital care, suffer through untreated illnesses, and die untimely deaths.

At the same time, in the developed countries, the hospital as an institution becomes continually more complex. Each year brings an acceleration of new medical knowledge requiring whole new hospital departments that were unknown a few years ago. The general hospital is taking on added services, some of which a few years ago were thought to be the prerogative of hospitals for persons with chronic disease. Patients are hospitalized for shorter periods in hospitals, and the cost per patient-day is rising. That general hospitals increasingly play the role of community health centres is indicated by the extraordinary increase in hospital emergency-department services and the tendency for more general hospitals to offer psychiatric, rehabilitation, and extended-care service. As hospitals become more complex and more involved treatment and surgery are undertaken, not only does the ratio of staff to patient increase but a more highly trained staff is required; during recent years a combination of medicine and engineering has produced a vast array of new instrumentation, much of which requires a hospital setting for its use. Such now relatively common procedures as open-heart surgery, for example, were unknown a few years ago.

HISTORY OF HOSPITALS

Care of the sick is a fundamental need of community life, and arrangements to deal with illness are present in all societies in some form. As it is an elemental need, care of the sick has always been closely linked with the economic and social development of people.

Hospitals in antiquity

It is known that hospitals existed in Ceylon (modern Sri Lanka) in 437 BC and were established in India somewhat earlier, in the time of Buddha. Eighteen institutions built by Aśoka, emperor of India in the 3rd century BC, are said to have had some characteristics similar to those of modern hospitals in that cleanliness was stressed, the patients were treated with kindness, and diet therapy was practiced. Temples were used as hospitals in Greek and Roman times.

The advent of Christianity gave impetus to the establishment of hospitals, which became integral parts of the church organization. From a decree of Constantine issued in AD 335, Christian hospitals developed at Rome, Constantinople, Ephesus, and other parts of the Roman Empire. The Hôtel-Dieu of Lyons was opened in 542 and the Hôtel-Dieu of Paris in 660. In these hospitals more attention was given to the well-being of the patient's soul than to curing his bodily ailments. The manner in which monks looked after their own sick became a model for the laity. The monasteries had an *infirmitorium*, a place to which their sick were taken for treatment. The monasteries possessed a pharmacy and frequently a garden with medicinal plants. In addition to caring for sick monks, the monasteries opened their doors to pilgrims and other travellers.

Religion continued to be the dominant influence in the establishment of hospitals during the Middle Ages. The growth of hospitals accelerated during the Crusades, which began at the end of the 11th century. Pestilence and disease were more potent enemies than the Saracens in defeating the crusaders. Military hospitals came into being along the travelled routes; the Knights Hospitallers of the Order of St. John in 1099 established in the Holy Land a hospital that could care for some 2,000 patients. This order has survived through the centuries as the St. John's Ambulance Corps.

While most hospitals of the Middle Ages were associated with monasteries, a few were built by cities. Three people living in the 19th century—Florence Nightingale, Louis Pasteur, and Lord Lister—had much to do with the phenomenal growth and acceptance of hospitals in modern times. Florence Nightingale organized nursing and made it a dignified profession with a high esprit de corps; nursing in modern hospitals is a direct result of her efforts. Without the work of Louis Pasteur in the development of germ theory and Lord Lister's application of it, surgery as it is now known would have been unlikely. Nineteenth-century developments in anesthesia made it possible to perform longer and more difficult operations.

Early hospitals in North America

The first hospital in North America was built in Mexico City in 1524 by Cortéz; the structure still stands. The French established a hospital in Canada in 1639 at Quebec city, the Hôtel-Dieu du Précieux Sang, which is still in operation although not at its original location. In 1644 Jeanne Mance, a French noblewoman, built a hospital of axe-hewn logs on the island of Montreal; this was the beginning of Hôtel-Dieu de St. Joseph, out of which grew the order of the Sisters of St. Joseph, now considered to be the oldest nursing group organized in America. The first hospital in the territory of the present-day United States is said to have been a hospital for soldiers on Manhattan Island, established in 1663.

The early hospitals were primarily almshouses, one of the first of which was established by William Penn in Philadelphia in 1713. The first incorporated hospital in America was the Pennsylvania Hospital, in Philadelphia, which obtained a charter from the crown in 1751.

Such hospitals had little in common with the hospital of today. The phenomenal advance in medical science in the 20th century, and its outstanding achievements in the prevention and cure of disease, the amelioration of human suffering, and the lengthening of the span of human life—all have depended in large part on the modern medical care dispensed in the general hospital.

THE MODERN HOSPITAL

Hospitals may be classified in various ways; by ownership and control; by type of service rendered; by length of stay; by size; or by facilities and organization provided. Terms in general use include the general as distinct from the special hospital; the short-stay hospital; and the long-term hospital.

Ownership, control, and financing. *Ownership and control.* In most countries outside of North America all or nearly all hospitals are governmentally owned and operated. In Great Britain, for example, except for a small number run by religious orders or serving special groups, as in the case of some Italian and Jewish hospitals and some facilities for the chronically ill, most

hospitals are within the National Health Service, and the local hospital management committee answers directly to the regional hospital board and ultimately to the Department of Health and Social Security. In the United States and Canada most hospitals are nonprofit and neither owned nor operated by governmental agencies. Many are associated with universities; others were founded by religious groups or by public-spirited individuals. Mental hospitals have been traditionally the responsibility of the state or provincial governments, while military and veterans hospitals have been provided by the federal government. In addition, there are some municipal and county general hospitals.

Financing. Almost universally, hospital construction costs are met at least in some part by governmental contributions. Operating costs are taken care of in a variety of ways. Ultimately, a large part of the expenses not covered by private endowments or gifts is met by contributions from the general funds of some unit of government or out of funds collected by insurance carriers from subscribers. In countries in which hospital insurance is not universal or complete, some of the operating costs are met by charges on uninsured or inadequately insured patients.

Carriers of hospital insurance

The carriers of the hospital insurance in a particular country may be governmental agencies, private corporations or agencies, or both. In Britain, for example, under the National Insurance Act, the government is the carrier. All persons who have reached the minimum school-leaving age and are not full-time students, beyond the age of retirement, in prison, or receivers of benefits from the insurance and who do not have less than a certain minimum income are contributors under the plan whether employed by others, self-employed, or nonemployed. Employers also contribute.

In the United States persons who are employed by others or are self-employed make compulsory contributions toward a form of national hospital insurance, Medicare, which pays a large portion of the hospital costs of persons aged 65 or over. Employers make matching payments. A majority of the persons ineligible, by reason of age, for benefits under the Medicare program are enrolled in some other form of hospital insurance. This may be Blue Cross, a group enrollment plan for employees that was sponsored by the American Hospital Association; one of the plans of approximately 900 commercial insurance companies; or one of the approximately 800 independent plans, including community and community-controlled plans and those operated by unions, employers, welfare funds, and private medical clinics.

In the United States, even with federal participation under Medicare and Medicaid (a program for persons under 65 who are unable to pay), the payment for health-care services on an insurance basis, either voluntary or governmental, is much less advanced than it is in many other parts of the world. In Europe, particularly, the financial support of services in hospitals tends to be much more collectivized. Less than 10 percent of the costs of hospital operation in Europe is covered by payments made directly by patients. Details vary somewhat from country to country: in the Union of Soviet Socialist Republics the program is based entirely on funds from a single source, while in Great Britain the funds for total hospital operation are appropriated by the Ministry of Social Security to each regional hospital board, which in turn distributes them to the local hospital groups. In Sweden some 90 percent of hospital operating costs are provided by local or provincial units of government from public revenue; the remaining 10 percent comes from payments made by insurance funds on behalf of the patient. In general, in most European countries, hospital operating costs are paid out of insurance funds; this is true in France, Germany, Italy, The Netherlands, Norway, and elsewhere.

The community general hospital

The general hospital. A community general hospital with about 200 beds has an organized medical staff, a professional nursing staff, and much expensive diagnostic equipment. In addition to the essential services of a first-class hotel, it has a pharmacy, a laboratory, departments for X-ray diagnosis and treatment and for physical therapy, probably a maternity division (ordinarily including a nursery and a delivery room), operating rooms, recovery rooms, an outpatient department, and an emergency department.

In a somewhat larger hospital one would expect to find additional facilities: dental services; a nursery for premature infants; a bank of organs for use in transplantation; a department of renal dialysis (removal of wastes from the blood by passing it through semipermeable membranes, as in the artificial kidney); equipment for inhalation therapy; an intensive-care unit; a self-care unit; a volunteer-services department; and, possibly, a home-care program. If the hospital is affiliated with a medical school, it may have facilities for closed-circuit television.

The advances made after World War II include the use of antibiotics, a vast new array of laboratory procedures, many new surgical techniques, new materials and equipment for radiation therapy, and an increased emphasis on physical therapy and rehabilitation. One trend that has developed has been the increasing use of the emergency department of the general hospital to meet medical needs that may not be critical emergencies. This is done by all classes of society and is particularly noticeable at night and during weekends.

The legally constituted governing body of the hospital, with full responsibility for the conduct and efficient management of the hospital, is usually a hospital board. The board establishes policy and, on the advice of a medical advisory board, appoints a medical staff and an administrator. It exercises control over expenditures and has the responsibility for seeing that professional standards are maintained.

The administrator is the chief executive officer, responsible to the board alone. In a large hospital there are many separate departments, each controlled by a department head. The largest department in any hospital is nursing, followed by the dietary department and housekeeping. Other departments, all important to the functioning of the hospital, include laundry, engineering, stores, purchasing, accounting, pharmacy, physical therapy, social service, pathology and X-ray, and medical records.

In many hospitals interdepartmental committees function with excellent results, and many difficulties are solved. Committees on pharmaceuticals, food service, medical records, hospital forms, laundry, and purchasing are among those that may be organized. A committee on laundry and linen, for example, might have as members the administrator or his delegate and representatives from nursing, purchasing, laundry, engineering, and the sewing room.

Organization of medical staff

The medical staff is organized into such departments as surgery, medicine, and obstetrics. The degree of departmentalization of the medical staff depends on the specialization of its members and not primarily on the size of the hospital, although there is usually some correlation between the two. The chiefs of the medical-staff departments, along with the radiologist and the pathologist, make up the medical advisory board, which usually holds monthly meetings on medical–administrative matters. The professional work of the individual staff members is reviewed by such medical-staff committees as those on medical records, medical audit, tissue (which verifies that tissue removed by a surgeon should in fact have been removed), credentials, and utilization. In a large hospital the committees may report to the medical advisory board; in a smaller hospital, to the medical staff directly, at regular staff meetings.

Specialized health- and medical-care facilities. Hospitals that specialize in one type of illness or one type of patient are common in Europe and in America, although modern opinion holds that, except in large university centres where postgraduate teaching is carried out on a large scale, the special hospital should become a department of a general hospital. Hospitals for the treatment of mental diseases and of tuberculosis are prominent examples of specialized institutions; also common

are hospitals exclusively for children; women; nervous diseases; ear, nose, and throat diseases; orthopedics; eye diseases; and cancer. Many of these hospitals are famous as centres of research and teaching.

Changing conditions or changing modes of treatment have lessened the need for some types of specialized institutions or reduced the number of such institutions that are required. This trend may be seen in tuberculosis and mental hospitals.

Tuberculosis hospitals. During the last two decades of the 19th century and the first three decades of the 20th century, much attention was given to providing tuberculosis hospitals. They were frequently situated in rural areas, where rest, relaxation, special diets, and fresh air were provided. Even if the tuberculosis was in an early stage, a stay of upward of two years was thought necessary to effect a healing of the disease; a permanent cure was not considered entirely feasible. The procedure did not change much until the discovery of antibiotics; the use of antibiotics, along with advances in chest surgery and routine X-ray programs, has changed the picture completely. Since the mid-20th century the decline in the use of tuberculosis sanatorium beds has been marked, and many of the sanatoriums in North America and Europe have been converted to other types of care. The compulsory pasteurization of milk has played a large part in eliminating bone tuberculosis, which formerly was often transmitted from cattle.

Mental hospitals. Psychiatric patients traditionally have been housed in long-stay mental hospitals, formerly called asylums. In the past few decades, however, a fundamental change has occurred, not only in the treatment of the mentally ill but also in the public attitude toward them. Most large general hospitals now have a psychiatric unit.

Changes in treatment of mental patients

The hospital stay of many persons with chronic mental disease has been shortened by modern medication and more understanding on the part of the public. Mental patients have been given an opportunity to participate in many activities, first within the hospital setting and later in the community, either with trial visits at home or with placement of selected patients in foster homes. A recent innovation has been night care for psychiatric patients, allowing the patient to work in the community during the day and return to the hospital at night for medication and other treatment.

Hospitals for women and for children. Specialized hospitals for women are common and are often staffed completely by women doctors. Many of these hospitals are privately owned. In England they are known as nursing homes (not to be confused with the North American use of the term); on the continent of Europe they are called clinics. Primarily, these hospitals are for surgical conditions, but medical and obstetrical cases are usually accepted.

There are many hospitals for children in the Union of Soviet Socialist Republics, and several cities in North America have their own specialized hospitals for children. Most hospitals accept children up to age 14, but some specify up to age 16. In cities without such hospitals the local general hospital usually provides a special ward for children.

Military hospitals. Many of the advances in medicine are the outcome of experience gained in war, and, in times of peace, military physicians and surgeons have made notable contributions in the fields of hygiene, preventive medicine, medical statistics, and hospital construction.

The largest hospital on record, the Chimorazo Hospital of Richmond, Virginia, was a military hospital of some 9,000 beds established by Confederate forces during the United States Civil War.

Advances during World War I

During World War I it was discovered that good medical care could be provided in improvised facilities, including public buildings, private homes, ships, barges, trains, and ambulances; another development was the mobile field hospital, with physicians, nurses, and supplies following the movement of troops for emergency treatment. That war also saw the mass movement of convalescent patients over long distances and by means previously considered impossible. The control of contagious diseases was much further advanced than in any previous war, and many of the new methods were later adapted to peacetime civilian practice.

One of the outstanding features of both World War II and the Korean war, as compared with earlier wars, was the much better chance a wounded soldier had of survival. This was the result not only of better medical and surgical treatment but also of much more efficient and faster evacuation of the wounded from the battle area to a sorting area, from which the more seriously injured were shipped out by plane to base hospitals several miles back. These facilities were well equipped and staffed, but the important factor was that casualties were treated soon after the wounds occurred, before infection was too far advanced.

Hospital ships, used to transport the wounded home for prolonged treatment in permanent hospitals on land, follow a long tradition; the Spanish Armada, in 1588, was accompanied by a hospital ship. Modern hospital ships are floating hospitals with equipment and medical and nursing staffs matched only by the most modern land hospitals.

Civilian hospital ship. In 1958 the organization People-to-People Health Foundation was formed to establish the first peacetime hospital ship, the "Hope." This ship, a converted naval hospital ship, set out on its maiden voyage on September 22, 1960, from California. Its purpose was to carry medical help to any country that needed and requested it. Its mission was to be threefold: to teach, to heal, and to be a pilot for what was hoped would later grow into a great fleet. The physicians who served were volunteers, while the nurses and other personnel were paid only token wages. The first voyage, which consumed a year, was to the Far East, to Indonesia, and to Vietnam.

Portable disaster hospitals

Packaged hospitals for large-scale disasters. Portable disaster hospitals have been developed, containing equipment for setting up receiving and sorting units, operating rooms, wards, a central sterile supply, a pharmacy, a general laboratory, an X-ray laboratory, and a unit for general stores. Each of the portable hospitals can be used to expand the facilities of an existing hospital or as a separate 200-bed unit in a preselected building. The equipment for a typical portable hospital of this type is contained in about 660 boxes and crates, weighs 45,000 pounds, and requires 7,500 cubic feet of storage space.

Regional planning. Sweden and the Union of Soviet Socialist Republics provide examples of advanced planning in the integration of hospital networks into coordinated health services. In both nations the government has assumed the responsibility of providing health care to all citizens. In Sweden financing is in part by compulsory health insurance.

Sweden is divided into health-service regions; each region includes several counties and has a central hospital. Each county within a region has a county hospital with up to 1,000 beds and with specialized and outpatient facilities to serve a population of about 300,000. The counties, in turn, are divided into districts, each of which has a population of about 75,000 and is served by a district hospital, which usually has 300 or more beds. Smaller communities have health centres or ambulatory service centres that are not administered as part of the hospital system.

The Soviet Union takes a somewhat different approach. In its thinly populated rural areas, general hospitals, called *uchastok* hospitals, may serve populations as small as 2,000 to 15,000 persons. These 15- to 100-bed general hospitals occupy the same premises and employ the same staff as general clinics—polyclinics—that provide general and specialized care. The hospital–clinic staff includes a general physician, a surgeon, and a dental surgeon. In larger *uchastok* centres there may also be a radiologist and a pathologist.

The hospitals next in size, the district hospitals, have 250–500 beds and usually have divisions for surgical, medical, obstetrical, and pediatric services and provide

care for infectious diseases; they may also include departments for eye, ear, nose, and throat disorders and for orthopedic surgery. Patients who cannot be treated adequately in the district hospitals are referred to the next higher level, the regional hospital, which serves a population of 1,000,000–5,000,000 people and contains up to 1,250 beds.

The republic hospital occupies the highest level in the Soviet system. Such a hospital, or complex of hospitals, serves as a referral centre and has the responsibility of undergraduate medical education. It may also be associated with one or more research institutes.

Regional planning in North America is less advanced. One regional pattern being proposed is a satellite system, centred on a metropolis and applying the principle of progressive patient care. The system would be focussed on the efficient provision of comprehensive health care to the residents of the region. Less serious cases would be handled in the outer, more accessible health facilities of the system; the more serious would be referred to the inner hospitals of the ring or to the research and teaching hospital at the core.

Metropolitan planning council

The term metropolitan planning council is used frequently to denote an advisory planning group that tries to coordinate services among member hospitals in a metropolitan area and to decide such questions as where additional beds are to be built. In North America, however, the hospitals tend not to be government operated, and it is difficult to achieve close cooperation among voluntary groups.

Extended health care and facilities. *Convalescent care.* Modern treatment of persons recovering from an illness, surgery, or injury gives attention not only to the patient's physical recovery but also to his emotional and social well-being. Many persons receive care outside the general hospital during convalescence—perhaps in (1) a convalescent hospital; (2) a nursing home, usually a smaller institution with medical attention available on call; (3) a chronic-disease hospital, providing long-care remedial treatment; or (4) the patient's own home, with the aid of services and equipment supplied by a general hospital.

Long-term hospital care. Long-term hospitals have traditionally been associated with the treatment of chronic disease requiring a limitation of activity. Among such conditions most often reported are heart disease, arthritis, rheumatism, back impairments, high blood pressure, mental and nervous disorders, and impairment of vision. The modern tendency to replace the special chronic-disease hospitals with chronic-disease units within general hospitals allows better coordination and integration of services and ensures continuity of care.

Progressive-care concept. With the advance in medical science and the ever-increasing cost of hospital operation, the progressive-care concept, a step-by-step progam for total care, is gradually gaining ground, both for outpatient and inpatient care. Progressive care can be divided into five categories: (1) intensive care; (2) intermediate care; (3) self-care; (4) long-term care; and (5) organized home-care programs. Two of the categories, self-care and home-care programs, are relatively new departures from past practice and consequently deserve special attention.

Categories of progressive care

Self-care facilities are organized into a separate unit in which ambulatory patients who require only diagnostic or convalescent care are given accommodations of the hotel type. The patients are free to wear street clothes and to go to the hospital cafeteria. It is evident that such a ward or wing of a general hospital requires much less costly equipment than the intensive- or intermediate-care units and can be staffed with far fewer nurses and aides.

Home-care programs are for selected patients who need some care but not all of the treatment facilities of a hospital. The patients are provided with a range of individualized medical, nursing, social, and rehabilitative services in their own homes, coordinated through one central agency. Patients can be considered ready for home care when: (1) diagnosis and a plan for treatment have been established; (2) inpatient hospital facilities are no longer required for proper care; (3) no more than two visits per week by physicians are required; (4) the nursing service has found that the physical environment of the home is such that the patient receives adequate care; (5) the patient is too ill to visit an outpatient clinic but does not need hospital care; (6) the family environment would have a therapeutic effect, and family members or others can be taught to provide the necessary care; and (7) the family and the patient prefer that he be cared for at home. Although home care is a means of conserving highly expensive acute-care beds and represents a significant extension of hospital service and, although most patients on home care do as well as they are expected to and some do much better, home-care programs have not been adopted on a large scale.

BIBLIOGRAPHY. The literature on hospitals is quite extensive. The best general reference book is MALCOLM T. MACEACHERN, *Hospital Organization and Management*, 3rd ed. (1969), the standard text for hospital boards, administrators, students, or anyone wishing to learn about hospitals. JOSEPH KARLTON OWEN (ed.), *Modern Concepts of Hospital Administration* (1962), has a long list of knowledgeable contributors, each of whom writes on his specialized field. For general statistical information regarding hospitals in the United States or Canada, the best source is the *Guide Issue of the A.H.A.* (published in August of each year). *Trustee* is a monthly journal (A.H.A.) for hospital governing boards but contains much information of use to the general reader with some knowledge of hospitals. A very readable history of medicine and to some extent hospitals is HOWARD W. HAGGARD, *Mystery, Magic, and Medicine* (1933; republished under the title *Devils, Drugs and Doctors*, 1937). The literature on hospitals in wartime is quite extensive. EDGAR ERSKINE HUME, *Victories of Army Medicine* (1943), sketches the United States Army Medical Department from 1775 to 1943. *Janes Fighting Ships* for 1944–45 portrays some hospital ships and gives tonnage. For the story of the hospital ship *Hope*, see WILLIAM B. WALSH, *A Ship Called Hope* (1964), *Yanqui, Come Back* (1966), and *Hope in the East* (1970). Probably the most important sources of information about hospitals may be obtained from the lending libraries of the American Hospital Association and the Canadian Hospital Association. Information on hospitals abroad may be obtained through selected publications of the World Health Organization.

(W.D.P.)

Hotel and Motel Industry

A hotel is a building or institution in which are provided lodging, meals, and other services for the travelling public. A motel, or motor hotel, generally performs the same functions in a format designed primarily for guests arriving by automobile.

Hotels have long been important elements in the economies of many countries. The enormous increase in tourism of the 20th century has caused the industry to outgrow national boundaries and become international. Growth has been especially large in certain resort areas, such as the Mediterranean and Caribbean seas. The modern industry has profited from the introduction of such new techniques as computerized reservation and credit systems and new forms of financial organization, such as the international hotel chain and the franchise motel. In the early 1970s more than 750 new hotels were planned or under construction throughout the world.

History to World War II. Inns were in existence in very ancient times to serve merchants and other travellers, though accounts of their operations are sketchy. In Roman times hostelries called *mansiones* were situated along the Roman road system, principally to accommodate travellers on government business. The commercial revival of the Middle Ages stimulated a wide development of inns and hostels. Many were operated by monasteries and other religious institutions. A famous example is the hospice in the Great St. Bernard Pass in the Swiss Alps, founded in the 10th century by St. Bernard of Montjoux (Bernard of Menthon) and still operated by the community of Augustinian monks.

Inns and hostels of the Middle Ages

In 13th-century China, Marco Polo found an extensive system of relay houses in existence to provide accommodations for travellers and way stations for the Mongol postal service.

Privately operated inns intended primarily for the use of merchants were widespread in both Islāmic and western European countries in the High Middle Ages. In many of them the innkeeper served in a semi-official capacity as part of a marketing or fair organization, safeguarding merchants' funds and acting as a broker.

The Industrial Revolution stimulated much progress in innkeeping, especially in England, whose inns became a standard for the world. Cleanliness, comfort, and good food became the hallmarks of the English inn. Meanwhile, across the Atlantic American innkeepers were setting a standard for size; by 1800 the inns of the United States were the largest in the world. The first American hotel to gain wide fame for its size was the 73-room City Hotel, described at its opening in 1794 as "an immense establishment." The American trend toward large size continued into the 20th century; hotels of 1,000 or more rooms are fairly common in the United States, while 600-room structures are rare elsewhere.

As travel for pleasure gained vogue in Europe, a new class of resort hotels was built in many countries. Along the French and Italian Riviera from Marseilles to Pisa resort hotels were constructed to serve wealthy vacationers, who frequently came for the entire summer or winter season. Luxury hotels soon made their appearance in the cities; in 1889 the Savoy Hotel in London set a new standard with its own electricity, a theatre, a private chapel, and a printing press.

Railroad hotels

The railroad had great impact on hotel development. Faster travel eliminated the need for establishments that had served the old coach routes in England, and many of them were forced out of business. On the other hand, new hotels were profitably built along rail lines. To accommodate the new travelling class, hotels were constructed as close as possible to the train station. This influence continued well into the 1930s, with the railroad companies themselves often owning hotels or leasing land for their construction.

A major development in hotel history occurred in 1908 with the opening in Buffalo, New York, of the hotel Statler (later named The Buffalo). Many innovations in services and conveniences were introduced by Ellsworth Milton Statler, the builder and operator, primarily for the convenience of the large and growing class of business travellers. From the Buffalo Statler grew the Statler Company, the first great chain operation in hotelkeeping.

World War I was followed by a period of tremendous hotel construction and growth. More new hotels were built from 1920 to 1930 than in any ten-year period before or since. Hotels increased not only in number but also in size. The Stevens Hotel (later the Conrad Hilton Hotel) in Chicago opened with 3,000 rooms and retained the title of world's largest until the late 1960s, when the Hotel Rossiya (Russia) opened in Moscow. Hotel-building declined sharply during the Depression of the 1930s and did not resume until after World War II.

Postwar developments. *The motel.* Prior to 1950, hotels were classified into the three categories of transient, residential, and resort. A fourth category—motel–motor hotel—is now necessary. It is not difficult to explain the phenomenal growth and success of this segment of the industry. The character of the hotel industry has always been directly related to the principal mode of transportation of the times. By 1950 in the United States and by the 1960s in Europe and Japan, the family automobile had become a principal mode of travel. Tourists furthermore preferred informality to the old hotel system involving formal dress, lobby parades, tipping, and parking problems. Finally, motel rooms were new, with modern furniture, wall-to-wall carpeting, and television. The swimming pool proved another popular attraction.

Although motels originated in the United States as roadside cabins, usually independently operated by a husband-and-wife team, their average size has increased to about 100 units today, and professional management has largely taken over. Though motels are still more common in the United States and Canada than elsewhere, Great Britain, Europe, the Caribbean, and Japan have all witnessed significant building in recent years.

Chain operation. Another important postwar development involves the trend to chain operation. While a number of hotel chains date from the early 1900s, the great growth period began after World War II. Although the chain-hotel operation is more predominant in the United States, the concept is becoming more common in most parts of the world. Chain operation permits increased efficiency, particularly in purchasing, personnel, promotion and sales, reservations, and financing. The four giant corporations in the hotel field in the early 1970s were Hilton Hotels Corporation; Sheraton Hotels and Motor Inns; Inter-continental Hotels; and, in the United Kingdom, Trust Houses, Ltd., later Trust Houses Forte, Ltd. Other prominent and rapidly growing chain operations are Marriott, Western International, Steigenberger, and Loew's hotels. The chain concept has become popular in the motel field as well. Holiday Inns Inc. is the single largest chain in the world, with more guest rooms than even the hotel giants.

Hotel corporations

Closely allied to chain operation is the franchising concept, which has been the most rapid development in the industry since 1950. Franchising allows an individual or group of investors to go into business for themselves while enjoying most of the advantages of chain operation. The investor purchases a franchise from a major company, pays a license fee as well as some type of daily fee per room to cover advertising, and leases the sign displayed in front of the motel. The investor owns the land, the building, and the equipment. From the franchising company he receives assistance in architectural planning, in training his staff, and in developing accounting procedures, purchasing procedures, and standard operating procedures. Most important, he benefits from the referral system of the franchiser, by which guests are guided to his motel by other franchise holders through a joint reservation system. Often the investor may purchase equipment and supplies at a considerable saving from the franchise chain.

Motel franchising

Several motel chains that pioneered the franchising concept have achieved phenomenal growth through its use. In each case, company-owned motels account for only a small portion of the total. A recent development has been the growth through the franchising concept of chains within the chains. Groups of investors form a corporation and obtain the franchise rights for a certain geographical area from the franchiser. As they build and operate the properties, they become a chain within the larger franchising chain.

The rapid growth of tourism (over 10 percent annually) exerts a strong influence on hotel operation internationally. Not surprisingly, many airlines have entered the hotel business.

Resort hotels. Resorts today fit many categories: seashore, mountain, winter, summer, year-round. Traditionally, a resort hotel provides comfortable rooms, excellent cuisine, and a location with scenic, historic, recreational, or therapeutic advantages. The majority of resorts today attempt to operate all year. The popularity of skiing and other winter sports opened a new season for northern resorts formerly restricted to summer operation.

Among the greatest growth areas have been the Mediterranean, the Caribbean, and Hawaii. Miami Beach, Florida, has 48 hotels per square mile. Jamaica, Barbados, Puerto Rico, the Virgin Islands, the Bahamas, and Trinidad and Tobago all count tourism as one of their most important industries. In Hawaii the island of Oahu has no more space available for hotels, and the outer islands are witnessing tremendous hotel growth. The Riviera and Black Sea, among the most famous of resort areas, have maintained their popularity in the face of increased competition from around the world. Spain has seen tremendous growth in its tourism, and Portugal is also becoming very popular. Resort areas in the Pacific are increasing in popularity; they are seen by many as the next great region for development in tourism.

Residential hotels. The residential hotel is essentially an apartment building offering maid service, a dining room, room meal service, and sometimes a cocktail

lounge. Residential hotels range from the luxurious, offering full suites, to the moderate, with single rooms.

The condominium

The condominium may be considered a form of residential hotel. In this concept, the apartments or suites are actually purchased by the occupant and the ownership is his. In resort areas, people often purchase condominiums as an investment. The owner may spend his vacation there and offer the condominium for rent by transients at other times, under the supervision of the condominium management. This concept is growing in popularity.

One of the most recent developments is a combination of condominiums, resort hotel, apartments, conference centre, and private homes. Many people feel that this combination is the blueprint for the resort complex of the future.

Management techniques. Today's hotel manager employs modern management techniques, to forecast and control labour and materials costs, and modern marketing methods. The electronic computer has achieved great success in hotel accounting and reservations. Research is now in progress for a completely computerized hotel operation.

The problem of employee costs versus productivity has become serious in hotel management in recent years. Since the productivity of hotel labour has remained constant or even declined, hotel architects have sought to design structures in which maximum service can be rendered with minimum personnel. Techniques for automating functions now burdened with high labour costs are being studied by hotels, manufacturers of equipment, and schools of hotel management. Automation of a service industry remains a challenge, however, with problematical results.

BIBLIOGRAPHY. For the reader who desires information on the overall nature of the hotel-motel field, description of operating departments and their functions within the enterprise, as well as the history, growth, and development of the industry, GERALD W. LATTIN, *Modern Hotel and Motel Management*, 2nd ed. (1968); and DONALD E. LUNDBERG, *Inside Innkeeping* (1956), are two excellent accounts. For information on careers in the industry and how to prepare for them, GERALD W. LATTIN, *Careers in Hotels and Restaurants* (1967), is recommended.

Technical books describing principles, practices, and theories of hotel-motel operation tend to specialize in certain function areas—*e.g.*, accounting, food and beverage, planning, sales, legal, and promotion. The most recent and very comprehensive book on accounting for the hospitality industry is CLIFFORD T. FAY, JR., RICHARD C. RHOADS, and ROBERT L. ROSENBLATT, *Managerial Accounting for the Hospitality Service Industries* (1971). Within the industry, the most widely used text on food and beverage management is JOSEPH BRODNER, HOWARD M. CARLSON, and HENRY T. MACHAL (eds.), *Profitable Food and Beverage Operation*, 4th ed. rev. (1962). All the technical procedures for performing this complex function are extremely well presented. C. DEWITT COFFMAN, *The Full House: A Hotel/Motel Promotion Primer* (1964), is by far the best book on hotel marketing, sales, and promotion. Little has been written on resort hotels, but E. ABRABEN, *Resort Hotels: Planning and Management* (1965), provides a good overview of the general procedures for planning and management. A sound technical approach to the legal problems of the industry is furnished by HENDRIK ZWARENSTEYN, *Legal Aspects of Hotel Administration* (1962). For the reader interested in a comprehensive coverage of all phases of motel operation, GEORGE O. PODD and JOHN D. LESURE, *Planning and Operating Motels and Motor Hotels* (1964), is highly recommended. HERBERT K. WITZKY, *Modern Hotel-Motel Management Methods* (1964), deals primarily with the personnel and human relations aspects of the hotel-motel business although some general management material is also covered.

A number of books have been written on the careers of famous hotelmen. Among the best are FLOYD MILLER, *Statler* (1968); CONRAD HILTON, *Be My Guest* (1957); ERNEST HENDERSON, *The World of "Mr. Sheraton"* (1962); and WHITNEY BOLTON, *The Silver Spade: The Conrad Hilton Story* (1954).

(G.W.La.)

Houseplant

Many exotic plants native to warm, frost-free parts of the world are grown indoors in portable containers or in miniature gardens, especially in colder climates. Houseplants have thus become a significant part of everyday life in economic as well as aesthetic terms and are the objects of a considerable business enterprise that supplies growing plants for interior decor.

The range of houseplants

Most houseplants are derived from plants native to the tropics and near tropics. Those that make the best indoor subjects are the species that adjust comfortably to the rather warm, dry conditions that generally prevail in indoor living spaces.

Although many plants can be grown successfully indoors, there are certain groups that, because of their attractiveness and relative ease of maintenance, are generally considered the best houseplants. These include the aroids, bromeliads, succulents (including cacti), ferns, begonias, and palms, all of which have long been favourites. Somewhat more demanding are those that are grown primarily for their flowers—the African violets, camellias, gardenias, geraniums (*Pelargonium* species), and orchids.

Historical development of the houseplant. Paintings and sculptures make clear that the practice of indoor gardening can be traced at least to the early Greeks and Romans, who grew plants in pots and perhaps brought them into their homes. The older civilizations of Egypt, India, and China also made use of potted plants but usually in outdoor situations, often in courtyards that were extensions of the house, and for centuries the Japanese have carried on the dwarfing of trees and other plants for room ornaments. But the popular art of growing houseplants did not receive much comment until the 17th century, when, in *The Garden of Eden* (printed in 1652), Sir Hugh Platt, an English agricultural authority, wrote of the possibility of cultivating plants indoors. Shortly thereafter, glasshouses (greenhouses) and conservatories, which were used during Roman times to force plants to flower, appeared in England and elsewhere to house exotic plants. In mid-19th-century England and France, books began to appear on the growing of plants in private residences, and the use of enclosed portable cases of plants (Wardian cases, or terraria) became popular. Since that time interest in houseplant cultivation has become so great that few houses in civilized countries today are without some form of plant life.

Some leading types of houseplants. There are thousands of tropical and subtropical plants that can adapt to growing indoors. Although some fancy exotic species do well only in a humid conservatory or a glass-enclosed terrarium, a great many species have been introduced that endure the adverse conditions of dry heat and low light intensity that prevail in many houses. A selection of the more widely favoured houseplants follows, under two sections: foliage plants, some of which also bear interesting flowers; and flowering plants, species kept primarily for their flowers.

Foliage plants. In the aroid family, which has provided a range of long-lived houseplants, most prominent are the philodendrons. These are handsome tropical American plants, generally climbers, with attractive leathery leaves, heart-shaped, and often cut into lobes. *Monstera deliciosa*, or *Philodendron pertusum*, the "Swiss cheese" plant, has showy, glossy, perforated leaves slashed to the margins.

Dieffenbachia, the dumb canes, in a number of attractive species, are handsome tropical foliage plants usually with variegated leaves; they tolerate neglect and thrive even in dry rooms. *Aglaonema*, the Chinese evergreens, are fleshy tropical Asian herbs of slow growth, with leathery leaves often bearing silvery or colourful patterns; they are very durable and tolerant of indoor conditions. *Scindapsus*, popularly known as pothos, or ivy-arums, are tropical climbers from the Malaysian monsoon area; their attractive variegated leaves usually are small in the juvenile stage. They do well in warm and even overheated rooms. *Spathiphyllum*, the peace lilies (not a true lily), are easy-growing, vigorous tropical herbs forming clumps; they have green foliage, and a succession of flowerlike leaves (spathes), usually white. *Anthurium* with colourful spathes, such as the flamingo flower, does best under humid conditions. *Caladium* are tropical American tuberous herbs producing fragile-look-

ing but colourful foliage; they keep surprisingly well if protected from chills and wintry drafts.

Begonias, with their often very decorative leaves, are long-time favourites among houseplants, but, with few exceptions, they require more humidity and fresh air than the modern home provides. *Begonia metallica*, with its olive-green, silver-haired foliage; *B. masoniana*, with beautiful green, puckered leaves splotched brown; and *B. serratipetala*, with small foliage spotted pink, are examples of types more resistant to dry rooms. Double-flowered varieties of *B. semperflorens*, the rose begonia, are popular plants for the windowsill.

Small foliage plants

There are many small foliage plants, often with strikingly patterned foliage, which are native to the tropical forest floor; some of them have become remarkably good houseplants. Among them are several prayer plants (*Maranta* species), which fold their attractive leaves at night; and the exquisite *Calathea makoyana*, or peacock plant, with translucent foliage marked with a feathery peacock design. *Pilea cadierei*, or aluminum plant, is easy to grow; it has fleshy leaves splashed with silver. *Codiaeum* species, or crotons, are multicoloured foliage plants that need maximum light and warmth to hold their leaves and coloration well. Although primarily thought of as bedding plants, the varicoloured coleuses, or painted nettles, can decorate a sunny window with a brilliant array of leaf patterns. *Peperomia* species form miniature rosettes or vines, with waxy foliage that are corrugated and decorated either with silver or creamy white.

Bromeliads constitute a plant family peculiar to the Western Hemisphere; they dwell on trees and rocks (as epiphytic plants) or on the forest floor (as terrestrial plants) and usually form rosettes of leathery, concave leaves, many with bizarre design or striking variegations. Their flowers may be hidden deep in the centre of the rosette, surrounded by a cup of brilliant crimson inner leaves, as in *Neoregelia* and *Nidularium*. Species of *Aechmea* and *Guzmania* form colourful spikes or heads of long-lasting leathery bracts or bright berries. *Billbergia* species are tubular in shape; their showy flower stalk, with blue flowers, is often pendant. Most forms of *Tillandsia* and *Vriesea* have spear-shaped, flattened, colourful flower spikes. The terrestrial *Cryptanthus*, the earth stars, are more or less flattened rosettes with striking leaf design, mottled, striped, or tiger-banded in silver over greens and bronzes.

Succulents. Cacti, native to the Western Hemisphere, have developed a special capacity to store water in thick, fleshy bodies. They thrive in much sun and need very little water. There are many often curious forms: the tiny button cactus, *Epithelantha;* the myriad pincushion species of *Mammillaria; Parodia*, or Tom Thumb cactus; and *Rebutia*, the pygmy cactus. The last two bloom when young and tiny. Other forms include *Gymnocalycium*, or chin cactus; *Notocactus*, or ball cactus; *Echinocactus*, known as barrel cactus; various *Opuntia* species, including bunny ears and chollas; and *Cephalocereus*, or old-man cactus, with its glistening white hair. Larger cacti include *Cereus* and its relatives, often night-blooming, and the giants of the desert, such as *Carnegiea*, the saguaro, with branching columns up to 60 feet (18 metres) in height. Cacti of tropical forests include the epiphytic *Rhipsalis* and the near-epiphytic orchid cacti, *Epiphyllum*, which blooms in many colours; both kinds are suitable for growing in baskets.

Succulents other than cacti have also contributed favourite subjects for indoor growing. Typical stem succulents are *Euphorbia*, with their often angled candelabra-like columns resembling those of cacti. Leaf succulents are represented by *Aloe*, famous since ancient times as the medicine plants; *Echeveria*, or hen and chickens; *Kalanchoe tomentosa*, the panda plant; *Crassula*, the jade plant; and *Haworthia*, which have rosettes with pearly dotted leaves. Durable pot plants include the strap-leaf snake plants, or *Sansevieria;* they are remarkable for tolerating much neglect and growing in locations in which few other plants can succeed.

Dracaena, the dragon trees, includes such good houseplant subjects as *D. marginata*, from Madagascar, which forms clusters of twisted stems topped by rosettes of narrow, leathery leaves. Other good examples are *D. deremensis* "Warneckei," with its handsome, symmetrical rosette of sword-shaped, milky-green leaves striped white; and *D. sanderiana*, the ribbon plant, a diminutive and slender, highly variegated species that can be grown in water. Similar in appearance is *Pandanus veitchii*, which has a rosette of leathery, sword-shaped leaves—glossy green and banded white—arranged in spirals.

Sub-tropical evergreens

Several subtropical evergreens can be grown in cooler locations indoors. Pre-eminent among them is the Norfolk Island pine (*Araucaria heterophylla*, or *A. excelsa*) —not a true pine—an undemanding graceful conifer with tiered branches of fresh green needles; it is long-lived even in dim corners, in any temperature above freezing. *Podocarpus*, the somber Buddhist pine, forms dense pyramids of dark-green needlelike leaves; it also prefers cooler locations.

Among the many broad-leaved woody evergreens used as houseplants are *Brassaia actinophylla*, the Queensland umbrella tree, better known as schefflera. Its spreading crowns of palmately divided, glossy green leaves, do best in a light and warm location. Another picturesque plant is *Polyscias fruticosa*, the Ming aralia, with willowy, twisting stems densely clothed toward their tops with fernlike, lacy foliage.

The so-called rubber trees of the genus *Ficus* are widely used in homes and offices. All require good light to hold their foliage well. Best known are the large-leafed *F. elastica* "Decora," but perhaps even more attractive, because of their very graceful habit, are several small-leaved kinds, such as *F. benjamina*, *F. retusa*, and *F. nitida*. The giant violin-like, leathery leaves of *F. lyrata*, better known as *F. pandurata*, make the plant an attractive indoor "tree." *Coccoloba*, the sea grape, is another sturdy, woody plant, somewhat resembling *Ficus*, with leathery, rounded leaves and crimson veining.

Because of their majestic beauty and distinctive decorative appeal many palms are grown indoors. Best known of the feather palms is the paradise palm (*Howeia*, or *Kentia*), a noble plant that combines natural grace with sturdiness; its thick, leathery leaves can stand much abuse. The parlour palms and bamboo palms of the genus *Chamaedorea* have dainty fronds on slender stalks; they keep well even in fairly dark places. Similar in appearance is the areca palm (*Chrysalidocarpus*) with slender yellowish stems carrying feathery fronds in dense clusters. The pygmy date (*Phoenix roebelenii*), a compact palm with gracefully arching, dark-green leaves, is an excellent houseplant if kept warm and moist.

Ferns

Ferns, which come in a wide variety of forms, provide many popular houseplants. Among the best smaller parlour ferns is the sword fern, *Nephrolepis*, with bushy rosettes of leafy fronds; the holly fern (*Cyrtomium*), which has glossy dark leathery leaves; and the leather fern (*Rumohra*), with its leathery but lacy fronds. The bird's-nest fern (*Asplenium nidus*) forms a rosette of parchment-textured, fanlike, light-green fronds. Long-lasting *Polypodium*, often known as hare's-foot ferns because of their pawlike, woolly rhizomes (rootlike structures), have feathery leaves on slender stalks. Among the attractive humid-loving ferns are the several species of dainty maidenhairs (*Adiantum*). The so-called table ferns are a varied group of manily *Pteris* and *Pellaea* species; some are frilly, others variegated; and in their younger stages they are ideal subjects for terraria, enclosed containers. *Platycerium*, or staghorn ferns, always have aroused great curiosity because of their unusual shapes. Growing as epiphytes on trees, their sterile fronds cling snugly to the bark or, in cultivation, to a wire basket or wooden block; their much divided fertile fronds resemble the antlers of deer. One of the best of the palmlike tree ferns is the Hawaiian *Cibotium*, with a stout, fibrous trunk that bears a crown of light-green fronds.

Popular fernlike plants include *Asparagus* species that have plumy fronds. Species of *Selaginella*, called sweat plants or moss ferns, are strictly warm terrarium subjects; their delicate fronds greedily soak up moisture from the atmosphere to keep from shrivelling.

Vines and trailers. Vines and trailers, weeping plants with stems too weak to support themselves, occur in most plant families. Best known are many varieties of ivy (*Hedera*). Generally, they prefer a cool location, but some small-leaved or variegated varieties do well on the windowsill. Several *Cissus* species, such as *C. rhombifolia*, the grape ivy, with metallic foliage, and the leathery *C. antarctica*, or kangaroo vine, are excellent subjects for planter boxes or room dividers. Intriguing is the slow-growing *Hoya*, or wax plant, with leathery foliage and waxy, wheelshaped blooms. By contrast, the inch plants and wandering Jew, species of *Tradescantia* and *Zebrina*, are rapid growers with watery stems and varicoloured leaves; these long-beloved houseplants are used widely in window shelves or hanging baskets. The spider plants (*Chlorophytum*, or *Anthericum*) are houseplant favourites, forming clusters of fresh green ribbonlike leaves banded white; young plantlets develop from the tip of arching stalks.

Flowering plants. Most of the flowering potted plants seen at holiday times are not easy subjects for long-term indoor cultivation. They require high light intensity, careful watering, and day–night differences in temperature that are not usually available in the home; greenhouses would offer better chances for successful cultivation. There are exceptions, however; one of the most successfully adapted houseplants is the African violet (*Saintpaulia*), with hundreds of named varieties, with blossoms from violet blue through rose to white and single- and double-flowered forms. Window bloomers, such as *Abutilon*, the parlour maples, have bell-like flowers resembling Chinese lanterns. *Impatiens*, or busy-Lizzie, are succulent herbs producing a succession of spurred flowers in gay colours. *Hibiscus*, the rose mallows, have short-lived giant blossoms in brilliant colours. Geraniums (botanically *Pelargonium*) have long been popular flowering plants in the sunny window; the foliage is often variegated or scented, and flower clusters may be in reds, pinks, and white.

A number of bulbous plants do well in lighted windows: *Hippeastrum*, better known as amaryllis; *Clivia*, the Kafir lily; *Haemanthus*, the bloodflower; *Neomarica*, the apostle plant; and *Veltheimia*, the forest lily.

Orchids present a more difficult and specialized subject for successful home cultivation, usually because of their requirements for light, controlled temperature, and sufficient humidity and ventilation. There are some kinds, however, that promise good results with ordinary care: epiphytic *Epidendrum* species, with waxy, usually fragrant, often greenish blossoms, and *Oncidium* species, or butterfly orchids, with brightly coloured, long-lasting yellow flowers marked with brown and often produced in large sprays.

Small fruit plants

Small flowering plants that produce edible fruit can be grown on a windowsill. With sufficient light and ventilation, success may be had with the Calamondin orange (*Citrus mitis*), the dwarf Chinese lemon (*C. limon* "Meyeri"), and the American-wonder lemon (*C. limon* "Ponderosa"). If space permits, the fig tree (*Ficus carica*) can be grown to yield edible fruit as can the Chinese dwarf banana (*Musa nana*) and the dwarf pomegranate (*Punica granatum nana*), the pineapple (*Ananas comosus*), and the coffee tree (*Coffea arabica*).

The needs of houseplants. This article is intended only to present basic information about houseplants; each type of plant will respond particularly well when given the special care that it needs.

The growing of houseplants can be simplified greatly by providing as close an approximation as possible to the original habitats of the various kinds. Plants that originate in the humid jungles of Malaysia or the rain forests of the Amazon, for example, generally require moist, shady conditions; others found in the semi-arid regions of Mexico or southern Africa generally respond well to dry, sunny conditions.

Light. Some of the shade-loving foliage plants (aspidistras, India-rubber plants, ivies, ferns, and some philodendrons) can subsist for months and even thrive in such darkened locations as wall pockets, niches, mantels, or alcoves. Even in these cases some bright light—not necessarily direct sunlight—during a part of the day is beneficial. Artificial light is often sufficient for some of these plants. Flowering plants, however, need much more light if they are to flower. Blooming of some kinds of plants is regulated by the duration of exposure to light, natural or artificial (see PHOTOPERIODISM).

Most houseplants do best when grown near a window. Southerly or westerly exposures are preferred for sun-loving plants, including flowering sorts (except African violets and begonias, which do better in easterly exposures) and cacti and succulents; northerly or easterly exposures are best for shade-loving plants. Many bulb plants (hyacinths, narcissuses, tulips) can be started in dim or even totally dark places, then moved into the light.

Temperature and humidity. The proper levels of these two factors, so essential to success with houseplants, are the most difficult to obtain in modern houses, especially during the winter months when most rooms are too hot and dry for the majority of houseplants. When moisture in the air is constant, a rise in temperature is accompanied by a drop in relative humidity. Temperatures as low as possible, consistent with comfort, will benefit most house plants: 60°–70° F (15°–21° C) is a good average during the day, with a drop of 5°–10° F (3°–6° C) at night. African violets, poinsettias, and most foliage plants fare best in the high part of the range; cinerarias and cyclamens respond best in the lower part. Most succulents, including cacti, prefer warm days and cool nights.

Soil. Bowls and dishes of all descriptions serve as containers for growing houseplants. Unglazed pottery with drainage holes are generally preferred to glazed containers without drainage holes because the former afford better aeration of the soil.

When, after a time, the soil mixture becomes depleted or encrusted with an accumulation of salts, or the plant becomes pot-bound (*i.e.*, the roots require more space to grow), repotting is beneficial. Decorative boxes filled with sand, gravel, or moss ensure less drying of the containers; the boxes may be filled with a soil mixture, and, allowing for necessary drainage, the plants can be grown directly in them.

Desert plants usually thrive in a gritty soil composed of clay with sand or other coarse material, whereas acid-loving plants benefit from a greater amount of leafmold and peat moss. Most houseplants generally grow well in a standard mixture composed of good garden loam, peat moss, and regular (coarse) builder's sand in equal amounts. Hydroponics, or soilless culture, in which water and minerals (and often a soil substitute such as sand) provide the support and nourishment for plants, is sometimes practiced in rooting cuttings taken from established plants; it is a common commercial method of growing many plants including vegetables.

Soilless culture

Water. Plants grown indoors in soil require regular watering, the amount depending largely upon the type of plant, the kind and size of the container, the soil mixture, and the surrounding atmosphere. The technique of watering is undoubtedly the most difficult part of indoor gardening. The soil in pots of desert plants, such as the cacti, is allowed to dry out before the next watering; that in pots of jungle plants is always kept moist but not soggy. In hot dry rooms, frequent watering and application of mist is beneficial to plants derived from moist tropical areas.

Watering by total immersion of pots is advocated when the number of plants is small. The plants can be observed for a few weeks to determine how quickly they dry out, and watering can be regulated accordingly.

Watering by immersion is not always practicable, however. A meagre supply of water applied to the surface does not penetrate deeply to the roots; watering, when done, should be thorough. Water-filled saucers placed under the pots are also beneficial but slower acting; water kept in the saucers all the time, however, may waterlog the soil and cause root rot. The microscopic root hairs, active in absorption of water and minerals, should not wither and dry; in addition, a dry ball of earth is a good

environment for mealybugs and root lice (see below *Ailments*).

Nutrients. Houseplants supplied with the necessary minerals produce the food and energy needed for healthy growth. These minerals, initially supplied in the original potting soil, are gradually depleted. Although many organic fertilizers, such as well-rotted animal manures, bone meal, and cottonseed meal, are good sources of nutriment, they are not as convenient as many commercial preparations for use indoors.

Foliar feeding, spraying nutrients on the leaves, has been practiced by florists for many years; this is one reason for the luxuriant foliage plants on display in their shops. This type of feeding, which is relatively inexpensive, can be likened to the mild feeding that tropical plants receive in their native habitats through daily rain showers. This is particularly true of all epiphytes (*e.g.*, bromeliads, certain orchids and ferns, gesneriads, and others that cling to trunks and branches of trees). Rain brings to these plants food-laden dust of the atmosphere, and the plants absorb the nutrients through their leaves. Foliar feeding is not meant to entirely replace root feeding but rather to supplement it, being applied occasionally to bolster the plants or to revive them from a dormant period.

Ailments. Houseplants are not usually troubled with fungous or bacterial diseases, probably because of the warm, dry conditions that commonly prevail in heated rooms, but they are prone to attacks by insects and mites.

Pests. Few houseplants are immune to attack by mealybugs, soft-bodied insects with a protective, cottony, white covering. Mealybugs have microscopic needlelike mouthparts that enable them to pierce plant tissues and extract sap. If the infestation is heavy, the affected plant part turns yellow and drops off. Colonies of mealybugs can be eradicated by touching them with a small brush or a cotton swab dipped in grain or rubbing alcohol. If a plant is heavy-textured like a cactus, India-rubber plant, or philodendron, a dousing of water under pressure often eliminates the bugs. Insecticides marketed in aerosol spray cans provide convenient and effective control.

Mealybug infestation

Besides the mealybugs, scale insects of both the soft and armoured variety may cause problems generally much harder to eradicate. Scale insects are characterized by a shield or scale above the body, and they become tightly attached to the plant from which they suck sap. A light-oil-emulsion insecticide is effective, especially when applied while the young insects are still present.

Plant lice, or aphids, usually appear on tender growth. About the size of a pinhead, soft-bodied, and yellow, black, or green in colour, they suck plant juices, causing stunted growth and curling of leaves. They prevent normal opening of the flowers but are easily controlled by nicotine sulfate sprays.

Thrips, small, slender blackish insects that hop when disturbed, injure plants by causing blotching and streaking of leaves; they feed by rasping the plant tissues to suck up the sap and can usually be detected by the small black dots of excrement they leave behind. For their control, a nicotine sulfate solution prepared with soap is effective, as are some of the newer commercial products.

Root lice, common in dry soil, gather about the roots, from which they suck the plant juices, causing wilting and loss of leaves. They can be eliminated by washing the roots under running water and then replanting in fresh soil. Organic pesticides are also effective; they are applied either directly to the soil and watered in immediately or dissolved in water used to water the plants. This treatment also eliminates worms, ants, and other pests that may be living in the soil.

Red spider, a mite, is another troublesome pest. Ornamental plants suffer from them greatly, especially if they are kept in a warm, dry atmosphere. These tiny mites spin fine strands of silk, particularly on the underside of the leaves, where they are not easily detected until the infestation is severe. Their damage causes a speckled condition on leaves, which eventually drop off. A daily syringing with water helps to eliminate these pests; chemical substances called miticides also aid in their control.

Lack of vigour. Plants are often disfigured by a drying or browning of portions of the leaves. This may result from a food deficiency; from the presence of artificial gas or coal gas; from sudden exposure to brilliant sunlight or cold drafts; from improper application of fertilizer or insecticide; or from high room temperatures, very low humidity, and either an excess or deficiency of water.

All houseplants require sunlight in some degree, otherwise they cannot manufacture food to grow. When a plant shows no signs of growing or the growth becomes spindly from too subdued light conditions, gradual exposure to more light may revive it. Artificial lighting of the proper quality and intensity can be substituted for natural light.

It is common for all plants to enter a period of rest or dormancy, the season and length of which vary with the species. Many foliage plants, notably ferns and palms, produce fewer new leaves during winter; cacti and succulents are often quiescent in early autumn; and some flowering plants after blooming appear on the verge of death. During such dormant periods, houseplants require much less water, warmth, and attention.

BIBLIOGRAPHY. General popular books dealing with houseplants include A.B. GRAF, *Exotic Plant Manual*, 2nd ed. (1972), with 4,200 photographs; ARNO and IREN NEHRLING, *An Easy Guide to House Plants* (1958); MONTAGUE FREE, *All About House Plants* (1946); and EIGIL KIAER, *The Complete Guide to Indoor Plants* (1958). Three more complete referential sourcebooks are A.B. GRAF, *Exotica 3* (1970), a monumental pictorial encyclopaedia (12,000 photographs), with a key to plant care in five languages and an extensive bibliography; ROYAL HORTICULTURAL SOCIETY, *Dictionary of Gardening*, 4 vol. (1951; suppl. 1956, 2nd ed. 1969), an extensive work on all aspects of gardening, with much information on houseplants; and F. ENCKE (ed.), *Pareys Blumengaertnerei*, 2nd ed., 2 vol. (1958–61), one of the best works on cultivated plants, in German, well illustrated and thorough.

(A.B.Gr.)

Hsia Kuei

One of China's greatest masters of landscape painting, Hsia Kuei (Xia Gui in Pin-yin romanization) was influential in developing during the period of the late Sung dynasty (1126–1279) a style of painting characterized by lyric expressiveness, subtle mood, and a quality of nature mysticism. Neatly placed headlands, evenly fading mists, soaring peaks, broad spaces, and tiny people presented a landscape picture that was definite as well as poetic and suggestive—an effect achieved by combining quick powerful brushstrokes with telling areas of ink wash.

Hsia Kuei's birth and death dates are unknown. According to most Chinese sources, he served in the Imperial Academy (Yü hua-yüan) under the emperor Ning Tsung (reigned 1195–1224), eventually attaining the rank of *tai-chao*, or "painter in attendance," and being awarded the highest honour a court painter could receive, the Golden Belt. The earliest source of information on him, however, a collection of painters' biographies compiled in 1298 by Chuang Su and titled *Hua-chi pu-i*, states that he was active in the academy under the reign of the emperor Li Tsung (reigned 1225–64). Perhaps his service in the Academy overlapped these two reigns and can provisionally be dated around 1200–40.

He was born in or near the city of Hangchow in Chekiang Province, then the capital of China, to which the Imperial Court had been moved in the early 12th century when the north of China was invaded by the Chin (Jurchen) Tatars. Because of this move, the latter half of the Sung dynasty is known as the Southern Sung period (1126–1279). During this period, the centre of painting, and the concentration of major artists, was in the Imperial Painting Academy. Accordingly, the leading masters of the 12th and first half of the 13th centuries, with a few exceptions, were the court painters of Hangchow. Most famous among them are Hsia Kuei and his contemporary Ma Yüan, whose school of painting came to be known as the Ma-Hsia school.

Influence of Li T'ang

No information exists on Hsia Kuei's early artistic training or the identity of his teacher, but most sources agree that he followed the stylistic tradition of an earlier landscapist in the academy, Li T'ang. Li T'ang's career extended from the late 11th century until sometime after 1138, when the court resettled in Hangchow; he had followed it there on its southward exodus and rejoined the academy when peace allowed its resumption of activity, living long enough after that to set the new direction that its painting took in the Southern Sung period. Li T'ang transformed the monumental, imposing landscape type of the 11th century into a new image of nature, seen closer up, less awesome, allowing easier response by the viewer. Where early Sung landscape paintings had been carefully constructed microcosms, complex in design and full of detail, with small figures, buildings, and so forth, scattered throughout the scene, Li T'ang tended to simplify his compositions and limit his materials, achieving thereby a more sharply focussed and immediate effect. He thus, with his followers, created the new mode of landscape sometimes called the lyric style. His ideas were developed and exploited by academy landscapists later in the 12th century, practically all of whom were to some degree his followers.

Hsia Kuei was no exception. The painting that is stylistically earliest among his signed, reliable works—a fan-shaped album leaf in the National Palace Museum, Taipei, representing two men in a house in a valley gazing out at a waterfall—follows fairly closely the models of earlier 12th-century academy masters who imitated Li T'ang but brings the scene even closer than they had, further limiting the range of view and focussing interest on the two figures. The drawing is more conventional than in Hsia Kuei's later works, although a few features of style that anticipate these are to be seen.

Album leaves

The favourite form of the Southern Sung academy painters was the album leaf, which sometimes took the round or oblate shape that indicates it was originally mounted on a flat fan but more often was square, or nearly square, in shape. Most of Hsia Kuei's surviving works are album leaves. Two preserved in Japan—a signed landscape in the famous collective album called *Hikkoen* ("Garden Plowed by the Brush") and another, unsigned, in the Tokyo National Museum—along with a signed, fan-shaped leaf in the Boston Museum of Fine Arts representing a sailing boat on a river during a storm, with houses and windswept trees on the shore—are widely accepted as genuine productions of his hand. All are painted on silk, the two in Japan in ink monochrome, the Boston leaf with a few touches of light colour. All follow the same simple but effective scheme: the composition is diagonally divided, upper left to lower right, with the landmass in the lower segment and the upper occupied only by distant hills (in the Hikkoen picture, even these are missing). In each, a central group of leafy trees projects into the empty area, and a house is seen below, partly hidden. In both the leaves in Japan, a figure of a farmer returning homeward enters across a bridge in one lower corner; in the Boston leaf, a sailing boat at left serves the same function as a secondary focal point. All are exquisitely calculated, perfectly balanced, conveying with great precision a scene glimpsed through haze, sharply focussed at a few points but obscured at others. In all, the void plays as positive a role as the solid masses.

Hand scrolls

Hsia Kuei was also a master at composing the hand scroll, or horizontal scroll, which one viewed by unrolling from one end to the other, rerolling the portion already viewed as one moved. The effect was that of gazing at a continuous panorama, as in an imaginary journey through the scenery of nature. The finest example of that form is his great "Pure and Remote View of Streams and Mountains" in the National Palace Museum, Taipei. Originally signed, it is now missing a final section that bore the signature. A colophon by a later writer, dated 1378, ascribes it to Hsia Kuei. The remaining portion is about 27 feet long. Executed on paper, in rich ink tones ranging from the deepest blacks to subtle, seemingly evanescent, washes of pale gray, it is probably the most brilliant performance in the ink-monochrome technique in the whole of Chinese painting. Strong tonal contrasts are used to give powerful bulk to the rocks by distinguishing their sunlit and shadowed faces. Broad, broken strokes of ink render their rough-textured surfaces. This technique is derived from Li T'ang's "small ax-cut" texture strokes, so-called because they suggest a surface hewn with an ax; the broader form, seen in works by Ma Yüan and Hsia Kuei, are called large ax-cut strokes.

These and other features of the drawing agree closely with descriptions of Hsia Kuei's style by Chinese writers. They speak of his use of a "split brush" (*i.e.*, the brush tip divided so as to make two or more strokes at once) in painting tree foliage, and of his freehand drawing of architecture, bridges, and so forth, "without employing a ruler." They say that he liked to use a brush with a worn tip, the purpose of this presumably being to avoid the over-refined, "slick" drawing of other Southern Sung academy artists. They speak also of his working with great facility and spontaneity, so that the painting was "finished suddenly, in a strange and mysterious way."

By courtesy of the National Palace Museum, Taipei, Taiwan, Republic of China



Detail, length 68.58 cm, "Pure and Remote View of Streams and Mountains," hand scroll in ink and paper by Hsia Kuei, early 13th century (Southern Sung). In the National Palace Museum, Taipei, Taiwan.

The composition of the "Pure and Remote View" offers dramatic juxtapositions of near and far, solid and void. Distant parts of the picture are painted in dimmer ink, a basic device for achieving a sense of space in Chinese landscape that is known as atmospheric perspective. The whole impression is of a spacious, mist-hung, peaceful terrain, seen toward twilight. The standard characterization of Ma-Hsia landscape as "romantic" seems well justified in this, the supreme surviving work of the school.

Several others of Hsia Kuei's hand scroll compositions survive in original or copy. The "Twelve Views from a Thatched Hut" exists in several versions, of which the best by far, preserving, however, only four of the views, is in the William Rockhill Nelson Gallery of Art in Kansas City. Another fine composition is preserved in a copy in the Shanghai Museum, which originally bore an inscription and the date 1204. A famous and often-recorded work of Hsia Kuei, his "Ten Thousand Miles of the Yangtze River," now survives only in a copy that is probably of the 16th century, in the National Palace Museum, Taipei.

Hanging scrolls

Hanging scroll compositions based on Hsia Kuei's paintings are less common. The "Landscape: A Fisherman's Abode After the Rain" in the Boston Museum of Fine Arts is ascribed to him and is in his style. A painting of "A Misty Gorge" in the Freer Gallery of Art is in his style and of his period. It reportedly once bore his signature, which was lost in remounting. Several paintings in Japanese collections may be copies of his works; the most famous and often-reproduced is the "Rain Storm" formerly in the Kawasaki Collection, once regarded as a major work of the master but now recognized to be a copy.

For a time after his death, Hsia Kuei may have suffered the same low esteem as other Southern Sung academy painters in the eyes of proponents of the new literati, or scholar-amateur, school of painting, for whom professionalism was equated with empty technique and academicism. Chuang Su wrote in 1298 (*Hua-chi pu-i*):

> He painted landscapes and figures, all very vulgar and debased. In the late Sung period, when the way the world had fallen into decline, and people's minds were restless and upset, Hsia Kuei won a name for himself by pursuing excess. But of real substance there is none.

Hsia Wen-yen, in *T'u-hui pao-chien* (1365), wrote of him in more positive terms:

> His works have an exciting (stimulating) quality. His ink tones give the effect of colours; his brushwork is mature and controlled; the ink washes are applied rich and wet, a remarkable achievement. His snow scenes are completely based on those of [the 11th-century landscapist] Fan K'uan. Of all the Academy masters of landscape after Li T'ang, none was his equal.

Tung Ch'i-ch'ang (1555–1636), great artist-critic of the late Ming period, who was generally scornful of the entire professional-academy tradition, exempted Hsia Kuei somewhat from his sweeping condemnation.

> Hsia Kuei followed Li T'ang but added the element of *simplicity* to his style. What he did was like what clay-workers call "reduced modeling." In his conceptions and intent, he is quite devoid of the "shortcuts" of imitativeness. With some forms hidden, some sunk [into mist], he has the "ink-plays" of the two Mi's at his brush-tip.

The last sentence, obscure as it may sound, is especially significant. "Ink-plays" was a term used by the scholar-amateur artists for their unassuming, noncommercial pictures. Two of them who were particularly fond of so designating their paintings were Mi Fei, also known as Mi Fu (1051–1107), and his son Mi Yu-jen (1086–1165), both highly respected by Tung Ch'i-ch'ang and other literati critics for their spontaneity and inspired, intuitive mode of painting. To relate Hsia Kuei to them was to credit him with the same qualities, in which academy artists were generally held to be deficient. While it is difficult to see much clear resemblance between paintings by or after the two Mi's and those of Hsia Kuei, Tung Ch'i-ch'ang's observation is not an empty one: in his use of subtle ink washes to render atmosphere, in his hiding of transitions in mist and reduction of fine-drawn detail, Hsia Kuei does appear to have learned something from Mi Yu-jen. Such affinities set him apart from his co-workers in the academy, including Ma Yüan, whose sleeker and more elegant style, used for carefully constructed, often mannered compositions, seems quite incompatible with the literati painters' ideals.

Ma Yüan's style proved relatively easy to imitate, and hundreds of later painters did so. The style of Hsia Kuei, on the other hand, produced its brilliant effects only in the hands of a master; direct attempts at imitation by lesser painters of later times are on the whole pallid and uninteresting. Nevertheless, his influence was considerable. One painter in the Yüan period (1279–1368), Sun Chün-tse, did creditable pictures in a mixed Ma-Hsia manner, and some artists of the Che school under the Ming dynasty, in the 15th and 16th centuries, played entertaining variations on the style, while robbing it of most of its depth, both pictorial and expressive. Through these Ming epigones and their flatter, more decorative landscapes, the Ma-Hsia school was transplanted in Japan, where it served as the basis for most of the ink-monochrome landscapes of the Muromachi period (1338–1573). Two of the major works of the Japanese painter Sesshū (1420–1506), both hand scrolls, appear to be modelled on paintings by Hsia Kuei. As late as the early 17th century in China, a few painters were still producing landscapes avowedly "in the manner of Hsia Kuei." But the direction that painting took in the Ch'ing dynasty (1644–1911) did not allow such tonal subtleties or such concern with space and distance, and Hsia Kuei's style was scarcely reflected in paintings of that period, nor was the artist himself mentioned in its literature. It is only in modern times that he has come to be recognized as one of the leading masters of Chinese landscape painting and one of the great interpreters of nature in world art.

Influence on Japanese painting

MAJOR WORKS

Some examples of Hsia Kuei's style and work (dating from the 13th century or considered to be faithful copies); "Pure and Remote View of Streams and Mountains" (National Palace Museum, Taipei, Taiwan; other versions and variations of this work may be found in the Asano Collection, Odawara, Japan; the Metropolitan Museum of Art, New York; and in the Freer Gallery of Art, Washington, D.C.; "Rain Storm" (formerly Kawasaki Collection, Kōbe, Japan); "Twelve Views from a Thatched Hut" (William Rockhill Nelson Gallery of Art, Kansas City, Mo.; a copy is held in the Moore Collection of Yale University Art Gallery, New Haven, Conn.); "Hikko-en," album, 2 leaves attributed to Hsia Kuei (Nakamura Collection, Tokyo); "Ten Thousand Miles of the Yangtze River," attributed to Hsia Kuei (National Palace Museum, Taipei, Taiwan); "Landscape with Two Men Gazing at a Waterfall," fan-shaped painting (National Palace Museum, Taipei, Taiwan); "River Landscape with a Windswept Tree" (Museum of Fine Arts, Boston); "River View with a Boat at Anchor" (Tokyo National Museum; a copy in the Iwasaki Collection, Tokyo).

BIBLIOGRAPHY. One of the best English language sources of information on Hsia Kuei is the article on him by S.E. LEE in the *Encyclopaedia of World Art*, vol. 7, col. 650–654 (1963). See also OSVALD SIREN, *Chinese Painting: Leading Masters and Principles*, vol. 2, pp. 119–124 (1956); LAURENCE SICKMAN and ALEXANDER SOPER, *The Art and Architecture of China*, pp. 129–137 (1956); JAMES CAHILL, *Chinese Painting*, pp. 79–87 (1960), and *Chinese Painting, XI-XIV Centuries*, pp. 14–15, 34–35 (1960).

(J.F.C.)

Hsi Chiang (River)

The Hsi Chiang (Xi Jiang in Pin-yin romanization), the major river of southern China, flows generally eastward for 1,216 miles (1,957 kilometres) from the Yunnan highlands to the South China Sea and drains, along with the Pei Chiang (*chiang*, "river") and the Tung Chiang, a basin 173,000 square miles (448,000 square kilometres) in extent. It is shorter than the other important Chinese rivers—the Yangtze, the Huang Ho (Yellow River), and the Sungari, but it delivers an enormous quantity of water, and its rate of flow is second only to that of the Yangtze.

Ports such as Canton, Hong Kong, and Macau are located in the delta region, one of the heavily populated

areas of China. The delta's fertile soil is intensively cultivated, and as many as three rice crops a year can be produced. The river's name, applied only to its lower course, means West River (*i.e.*, flowing from the west). The Hsi Chiang is also an important commercial waterway, navigable as far inland as Wuchow—170 miles upstream from Canton—and beyond (see CANTON; HONG KONG; MACAU).

The river basin. The Hsi Chiang itself drains an area of about 127,000 square miles of southern China and North Vietnam. It shares a delta with the Pei Chiang and the Tung Chiang. Over half of the basin is mountainous and lies between 9,900 and 1,650 feet above sea level; more than 40 percent of the rest of the basin is occupied by hills between 330 and 1,650 feet high. The lowlands of the delta account for less than 5 percent of the total drainage area. Most of the mountains and hills are composed of limestone, and the river has cut a cavernous valley through them. The riverbed is broken by rapids and gorges, and its walls are often high and steep; the landscape is of the type known as karst, in which the limestone rocks are honeycombed with tunnels and openings, so that much of the drainage runs underground, and deep sinks or swallow holes abound.

The Nan-p'an Chiang

The river course. The main headstream is generally considered to be the Nan-p'an Chiang (South Pan River), rising in the Yunnan–Kweichow plateau region at an altitude of about 6,900 feet but dropping a total of 5,900 feet in the first 530 miles of its course, which flows in a southeasterly direction through Yunnan Province. It then forms part of the border between Kweichow Province and the Kwangsi Chuang Autonomous Region for a distance of about 535 miles. Southeast of the town of Ts'e-heng, the river receives the Pei-p'an Chiang (North Pan River) and is then known as the Hung-shui Ho, or Red River. This section of the river flows for about 400 miles through a narrow valley with high, mountainous banks dropping down about 850 feet. Its bed—between 165 and 1,000 feet wide—is broken by rocky rapids that are less than three feet deep and difficult to navigate. For the first 75 miles of its course the Hung-shui continues to form part of the Kweichow–Kwangsi border, flowing in an easterly direction until just before it reaches the town of Chiu-t'ien-o, where it makes a great bend to the south and flows through Kwangsi. At Kungchuan it turns northward and then resumes its easterly direction.

At Shih-lung the river receives the Liu Chiang, its most important left-bank tributary, and is then called the Ch'ien Chiang. This shortest section of the river is no more than 75 miles in length, the river dropping about 50 feet in this distance. The channel grows dramatically, occasionally achieving depths of 280 feet. For almost half its length, the Ch'ien Chiang flows through the narrow, rock-strewn Tat'eng Gorge between the cities of Wu-hsüan and Kuei-p'ing, at the end of which it receives its most important right-bank tributary, the Yü Chiang, and is then known as the Hsün Chiang. This section flows for about 120 miles in an easterly direction, dropping a further 55 feet and receiving the Jung Chiang on the right bank at T'eng-hsien, and the Kuei Chiang on the left bank at Wuchow (Ts'ang-wu) on the Kwangtung Province border. The widest point of the Hsün Chiang is 1,120 feet, at Lungtangsha (gorge), while its maximum depth reaches almost 200 feet.

Below Wuchow, where it enters Kwangtung Province, the river is known as the Hsi Chiang. Its valley consists of a series of winding gorges and wide hollows. The Sanjung and Lingyang gorges narrow to widths of 230 to 260 feet and are about 250 feet deep. Throughout its 130-mile length the Hsi Chiang drops only about 30 feet, flowing to the east until it joins with the Pei Chiang at San-shui. It then turns south through the vast Chu Chiang, or Canton, Delta before emptying into the South China Sea west of Macau.

Rivers that form the delta

The delta. The delta is formed by three main rivers —the Hsi Chiang, the Pei Chiang (North River), and the Tung Chiang (East River). At San-shui, the Hsi Chiang and Pei Chiang are joined by a short channel and then divide. The larger branch, the Hsi Chiang, bends to the south and forms the western border of the delta, while a lesser branch, the Fo-shan Chih-liu, flows eastward into the delta itself. The Tung Chiang flows from the east and enters the delta's main channel, the Chu Chiang (Pearl River), which begins just above Canton. Canton lies west of the point where the Tung Chiang flows into the main channel; Hong Kong stands to the east and Macau to the west of the entrance to the Chu Chiang estuary, which is about 18 miles wide.

Covering an area of about 1,500 square miles in southeastern Kwangtung Province, the delta is a complex network of river branches and channels, divided by islands of alluvial soil and by hills that were once coastal islands. The fertile islands are only slightly above sea level and are protected from the sea by a system of flood dikes.

Climate. Because of its great size and its varied terrain, the Hsi Chiang Basin experiences a wide climatic range. In general, temperatures and rainfall increase from west to east. In the west at K'un-ming, in Yunnan Province, the annual temperature averages 62° F (17° C); the annual precipitation of 41 inches falls mainly during the months of July and August. At Canton, in the east, the annual average temperature is 73° F (23° C), and the rainfall is 64 inches, most of it falling in May and June. Far to the north of the basin, at Kuei-lin, in Kwangsi Chuang Autonomous Region, the yearly average temperature is 62° F (17° C), and rainfall averages 74 inches a year, falling mostly in May and June.

Vegetation. Forests cover much of the mountainous region of the basin, especially in the north and west, the stretch along the border of Kweichow Province being the most heavily forested. The more important trees are the pine, the fir, the camphor, the tung, and bamboo. In the eastern part of the basin, in the low places of the maritime zone and in the valleys, the vegetation is mostly cultivated, and includes rice, peanuts (groundnuts), sugarcane, hemp, tobacco, and fruit.

Animal life. The river contains an abundance of freshwater fish. The waters of the Chu Chiang Delta are well utilized as fishponds. Monkeys, bears, and tapir frequent the mountainous regions to the west.

Human ecology. Throughout the mountainous part of its course, the river has but little relationship to the peoples who live in its vicinity. Most settlement occurs in small valleys or in plains between mountains, known as *patse*. The villages are isolated, compact agricultural units. River towns become more frequent in the hilly part of the Kwangsi region, where the river is an artery of commerce; towns include Kungchuan, Ch'ien-chiang, Lai-pin, Kuei-p'ing, T'eng-hsien, and Wuchow.

The Chu Chiang Delta is one of the most densely populated regions of China. It supports about 10,000,000 people and contains the cities of Canton, Fo-shan, Hsin-hui, and Chiang-men. The entire region is intensively cultivated. Rice is the most important crop, but in the cooler, drier climate of the west wheat, corn, sorghum, beans, and potatoes are also grown. An intricate irrigation system includes over 1,200 miles of flood dikes. The waters are also used as fishponds. The delta's many channels are vital to Canton's international commerce, as well as to trade with the interior.

Agricultural products

Hydrology. The river basin is nourished mainly by rainfall, and its water levels and volume vary according to the relative abundance of precipitation. The rate of flow more than doubles in the summer season. Most of the large flow results from the summer monsoon rains, when torrential floods may occur and frequently do cause catastrophic damage. The water is at its lowest during the winter dry period.

Fluctuations in the water level during the year may vary by as much as 80 feet at Wuchow; in the lower course and the delta variations in the level are less. Especially dangerous are the delta floods that are caused by the combined flooding of the Hsi Chiang, the Pei Chiang, and high tides. In the river basin, there were 29 floods in the 17th century, 26 in the 18th, 36 in the 19th, and 24 in the first half of the 20th century.

Navigation. The river is important for its use as a transportation artery. The basin contains over 9,000

miles of water routes, of which more than 6,800 miles are in use. Steamships can sail along more than one-third of the total length of the waterways, while Chinese junks and small craft ply all the navigable waters. Navigation is hampered by low water on many tributaries and by rapids on some sections of the river. In several places, as on the Yü Chiang, river craft are pulled over the rapids with hand-worked windlasses. During low-water periods, transportation ceases on some rivers, including the Tung Chiang and the Pei Chiang.

The water routes do not form an integrated system. Canton, the largest city in the basin, does not have direct access to either the Hsi Chiang or the Pei Chiang. The channels that connect the city to the water routes of the basin are winding and mud filled and are navigable only by shallow-draft boats. Because most of the river branches of the delta are shallow, oceangoing vessels cannot reach Canton, but must dock at Huang-pu-chiang (Whampoa), ten miles downstream.

Hydroelectric potential

Prospects for the future. The river basin has a large hydroelectric potential. Of the estimated potential supply of 28,500,000 kilowatts, about 25,000,000 can be generated by the Hsi Chiang itself. By the early 1970s, however, this resource was little developed. A series of major hydroelectric projects are planned at Sanjung, Tat'eng, Feilan, and Lungtang gorges, which will not only supply electricity but will also regulate the river's flow and provide protection from ruinous floods.

(A.A.So.)

Hsüan-tsang

In the 7th century AD Hsüan-tsang (Pin-yin romanization Xüan-zang), one of the most famous Buddhist monks in Chinese history, made a perilous journey from China to India and back, bringing home with him not only the sacred books of Buddhism but also invaluable geographical knowledge concerning the countries through which he passed on his journey. He is honoured by Buddhists in East Asia for his courage in undertaking the journey, for the volume and diversity of his translations of Buddhist texts into Chinese from Sanskrit, and for his founding in China of the Buddhist Ideation Only school, which stressed that the world is but a representation of the mind. Hsüan-tsang is further noted for his achievement as the transmitter of a great religious tradition from India to China and thence to Japan.

He was born in 602 in Ch'en-liu (modern K'ai-feng, in Honan Province), the youngest of four sons in a family in which there had been scholars for generations. He received a classical Confucian education in his youth, but under the influence of an older brother he became interested in the Buddhist scriptures and was soon converted to Buddhism. With his brother he travelled to Ch'ang-an and then to Ssu-ch'uan (modern Szechwan) to escape the turmoil that had resulted from the dynastic change from the Sui (581–618) to the T'ang dynasty (618–907). While in Ssu-ch'uan, Hsüan-tsang began studying Buddhist philosophy but was soon troubled by numerous discrepancies and contradictions in the texts. Not finding any solution from his Chinese masters, he decided to go to India to study at the fountainhead of Buddhism.

Journey to India

The biography of Hsüen-tsang by Hui-li and Yen-tsung, two Chinese monks, has preserved a detailed description of him as he set out all alone for India in 629. He was tall and handsome like a figure in a painting.

> His colouring was delicate, his eyes brilliant. His bearing was grave and majestic, and his features seemed to radiate charm and brightness. . . . His voice was pure and penetrating in quality and his words were brilliant in their nobility, elegance, and harmony, so that his hearers never grew weary of listening to him.

Later in India, his handsome features were the cause of his being seized by a group of river pirates, who were followers of the Hindu goddess Durgā and wanted to present him as a sacrifice to her. His protestations that he was only a devout pilgrim searching for the truth of Buddhism fell on deaf ears, whereupon he went into a deep meditation to prepare for his death. While he was absorbed in meditation, a violent storm arose that buffeted the boat. The pirates were so terrified that they freed the holy monk and asked for forgiveness and repentance.

Hsüan-tsang's journey took him across the frightful deserts of Central Asia and the towering, snowcapped mountains of northwest India before he arrived in the holy land of Buddhism, in 633. His journey across Central Asia was almost interrupted in Turfan in present-day Sinkiang Province, where the King was so impressed by Hsüan-tsang and his learning that he wanted to keep the Chinese monk there in his court as his spiritual preceptor. But Hsüan-tsang threatened a hunger strike, which weakened the King's will to detain him. The King proved to be the pilgrim's greatest benefactor, for he provided the Chinese traveller with letters of introduction to various ruling princes along the way, thus facilitating his travel to the very gates of India.

In India, Hsüan-tsang visited all the sacred sites connected with the life of the Buddha, and he journeyed along the east and west coasts of the subcontinent. The major portion of his time, however, was spent at Nālandā to perfect his knowledge of Sanskrit, Buddhist philosophy, and Indian thought.

Though Hsüan-tsang was attracted by the land of India and by Buddhist thought and was a man of universal interests, he remained true to his Chinese heritage. He spoke as a Chinese endowed with Confucian moderation and practical common sense when he severely criticized the ascetic excesses of the Hindu holy men, who wore garments soiled with dirt, ate decayed food, and moved around the land naked and filthy, with skin broken and feet horny and cracked. When he expressed his desire to return to China after his sojourn in Nālandā, his fellow Indian monks tried to dissuade him from leaving, saying that China was the land of the barbarians, because the Buddha had not been born there. This aroused Hsüan-tsang to a spirited defense of China, stating that in his native land the laws were universally observed, the ruling princes were virtuous, the subjects loyal, fathers affectionate, and sons obedient. He also pointed to the advances achieved by the Chinese in astronomy, music, and engineering.

While he was in India, the fame of Hsüan-tsang became so great that he was honoured and respected even by the powerful King Harṣa, ruler of North India. When Harṣa heard that the famous Chinese pilgrim had arrived near his camp, he hurried in person to the place where the pilgrim was staying so that he might pay his respects. Harṣa then convened a grand religious assembly presided over by Hsüan-tsang, on which occasion the latter successfully dispelled the reputed ignorance of the Theravāda (Way of the Elders) Buddhists and shattered the pride of the Brahmins (Hindu priestly caste). When Hsüan-tsang prepared to return to China, Harṣa loaded him with gifts and provided him with escorts.

Return to China

Hsüan-tsang returned to Ch'ang-an, the T'ang capital, in 645, after an absence of 16 years. His return contrasted markedly with his departure from China; he had had to leave by stealth and to travel only by night for fear of detection by the imperial guards stationed at the western passes of China to guard its borders. Now, however, he was accorded a tumultuous welcome at the capital, and, within a short time after his arrival, he was received in audience by the Emperor, who questioned him in detail about the climate, products, peoples, and manners of the countries he had visited. The T'ang legions had carried the might of the dynasty clear across Central Asia, and the Emperor thus was interested in the countries and peoples of that region. So impressed was he with what he heard from the Buddhist monk that he offered him a ministerial post. Hsüan-tsang, however, preferred to serve his religion, so he respectfully declined the imperial offer.

Hsüan-tsang spent the remainder of his life translating the Buddhist scriptures, numbering 657 items packed in 520 cases, that he brought back from India. He was able to translate only a small portion of this huge volume, about 73 items in 1,330 chapters. His main interest was centred on the teachings of the Ideation Only school, and in his massive work, *Ch'eng-wei-shih-lun* ("Treatise

on the Establishment of the Doctrine of Consciousness Only"), he summarized the teachings of the Indian masters of this tradition. Together with his chief disciple, K'uei-chi, he founded the Wei-shih school (Ideation Only school) in China. While Hsüan-tsang and K'uei-chi lived, the school achieved some degree of eminence and popularity, but its subtle and abstruse philosophy was alien to the Chinese tradition, and with the passing of the masters the school declined. Before this happened, however, a Japanese monk named Dōshō arrived in China in 653 to study under Hsüan-tsang, and, after he had completed his study, he returned to Japan to introduce the doctrines of the Ideation Only school in that country. During the 7th and 8th centuries, this school, called Hossō by the Japanese, became the most influential of all the Buddhist schools in Japan.

In addition to his translations, Hsüan-tsang composed a *Ta-T'ang Hsi-yü-chi* ("Records of the Western Regions of the Great T'ang Dynasty"), which preserved his account of the various countries passed through during his journey. The data this work contains has proved to be most useful for historians and archaeologists working in Central Asiatic and Indian history.

Hsüan-tsang died in 664. Out of veneration for this intrepid and devout Buddhist monk and pilgrim, the T'ang Emperor cancelled all audiences for three days.

BIBLIOGRAPHY. The works of SAMUEL BEAL, *The Life of Hiuen-tsiang*, new ed. (1911); and STANISLAS JULIEN, *Histoire de la vie de Hiouen-Thsang et de ses voyages dans l'Inde* (1853), are translations of a biography of Hsüan-tsang, compiled by two Chinese monks, HUI-LI and YEN-TSUNG. Beal's translation, however, covered only five of the ten chapters in the original Chinese; Julien's version is a complete translation. ARTHUR WALEY, *The Real Tripitaka*, pp. 11–130 (1952), is a popular version of the biography written in a lively and interesting style. It is not as complete as the account found in RENE GROUSSET, *Sur les Traces du Bouddha* (1929; Eng. trans., *In the Footsteps of the Buddha*, 1932), which relied mainly on Julien's translation as its source, and which discusses the life of the Chinese pilgrim against the background of T'ang history and Buddhist philosophy.

(K.K.S.C.)

Hsün-tzu

Hsün-tzu (in Pin-yin romanization Xun-zu) was one of the three great Confucian philosophers of the Classical period in China. He elaborated and systematized the work undertaken by Confucius and Mencius, giving a cohesiveness, comprehensiveness, and direction to Confucian thought that was all the more compelling for the rigour with which he set it forth; and the strength he thereby gave to that philosophy has been largely responsible for its continuance as a living tradition for over 2,000 years. Author of the book that bears his name, Hsün-tzu did not confine his philosophical energies to ethics and politics; he also made original contributions to philosophy in such divergent fields as semantics, aesthetics, education, logic, and the psychology of perception. He was also a distinguished Classical scholar with a rich knowledge of archaic Chinese myths, legends, ceremonies, and folklore. Many of his intellectual achievements came to be obscured as later Confucians focussed on the misanthropic view attributed to him—perhaps erroneously—that human nature is basically ugly or evil, and, beginning about the 12th century AD, his writings fell into a period of disfavour and neglect from which they are only now re-emerging.

Conjectural character of Hsün-tzu's life story

Personal life and official career. Hsün-tzu shares with many other influential figures of antiquity a hazy personal background and history; virtually nothing is known of his ancestry or of his personal life. Even his precise surname is disputable, with Sun occurring more regularly in early records than Hsün. The two characters that represent these names are written differently but are thought to have been pronounced the same in archaic Chinese. His given name was probably K'uang, his honorary cognomen *ch'ing*, a title for an official or nobleman. According to some sources he may actually have been descended from minor nobility, but his patrimony was without economic consequence, for no record shows him to have been a man of means. Like most other Chinese scholar-philosophers, he is generally accorded the honorific title *tzu*, and hence he has come down to posterity simply as Hsün-tzu—Master (or Philosopher) Hsün.

He was born in the ancient state of Chao, most of which lay in what is now Shansi Province. The several dates given for his birth span almost a half century, from an early estimate of 340 BC to a late one of 298 BC. The latter date is probably nearer the mark, but by any calculation his life falls squarely in the Warring States period (481–221), a period of almost unremitting political intrigue and internecine warfare.

Thus, he lived in a milieu of great change and instability, the antithesis of the harmonious and well-ordered traditional society he came to advocate as the philosophical ideal. Hsün-tzu was no hater of change, for the detailed descriptions of the just society found in his writings reflect a sophisticated appreciation of the manifold advantages in areas such as trade, social mobility, and technology that were accompanying the breakdown of the feudal order. At the same time, he could see that these societal transformations were also bringing to the Chinese the demise of their ancient socio-religious institutions, and he believed that the ritual practices (*li*) linked with those institutions were too important to be lost during secularization. For him, those ritual practices were important to the society because they were a culturally binding force for a people whose existence depended on cooperative economic efforts, and further, those ritual practices were important to the individual because they provided an aesthetic and spiritual dimension to the lives of the practitioners. By his fundamental insistence on the necessity of cultural continuity for both the physical and psychological well-being of his fellows, Hsün-tzu placed himself squarely in the ranks of Confucian philosophers.

Stress on value of ritual practices

Like many other intellectuals of the time, he first came to public attention in Chi-hsia, a famous centre of learning in the state of Ch'i, whose rulers traditionally provided subventions for scholars. That Hsün-tzu was already a mature thinker when he arrived there is suggested by the record of his having thrice officiated at a special court ceremony, an honour generally reserved for scholars of distinction. From his own writings it is clear that he visited the authoritarian state of Ch'in briefly—perhaps in 264—but the only date in his entire life on which all accounts agree is 255, when he was given the position of magistrate in the city of Lan-ling in the state of Ch'u. Unlike Confucius and Mencius, both of whom travelled extensively from state to state seeking a morally conscious ruler who would employ them and implement their ideas, Hsün-tzu chose a more modest official life. Except for one brief period, he continued as a magistrate until the death of his political patron, in 237. After leaving office he remained in Lan-ling until he died a few years later (*c.* 230).

Teaching and writing. Hsün-tzu is remembered for his teaching and his writing, not for his official or personal life. As a teacher, he had the distinction of being the tutor to not one but two of the most politically influential men in ancient China: Han-fei-tzu and Li Ssu. But the former became a Legalist philosopher opposed to the principles of his mentor, and the latter censured the Confucians and their writings when he became prime minister of China (*c.* 219 BC). Thus, both men earned the enmity of later Confucian historians, and the approbrium they have consistently received through the centuries has also affected the evaluation of their teacher. Ironically, his most famous students brought Hsün-tzu more reproof than renown.

Doctrine of man's evil nature

His writings were no less the recipient of moral disapproval than his teaching, owing in large measure to the oft-quoted essay "Man's Nature is Evil." Because Mencius believed that human beings were innately disposed toward moral behaviour, Hsün-tzu was perceived, as the author of this essay, to be attacking his illustrious predecessor. The Neo-Confucians of the Sung dynasty (AD 960–1279) adjudicated the controversy in favour of Mencius, who was canonized as the second Sage of Confucianism, while Hsün-tzu was declared heterodox. Some

contemporary scholarship suggests that the passages containing this doctrine are later interpolations into Hsün-tzu's text. Authentic or not, the doctrine has probably received more attention historically than it deserves, because it does not appear to be essential to the development of Hsün-tzu's overall philosophical position.

Whatever the final verdict on the significance of the "Man's Nature is Evil" passages, however, textual corruption is the exception and not the rule for Hsün-tzu's writings; unlike most other Chinese works of the same period his book managed to escape the forger's pen, and most of the material in the 32 essays that comprise the *Hsün-tzu* is considered to have come from his own hand. These essays are a milestone in the development of Chinese philosophy, because the anecdotal and epigrammatic style which had characterized earlier philosophical literature—*i.e.*, the *Analects*, *Tao-te Ching*, *Mencius*, *Chuang-tzu*—no longer sufficed to convey fully and persuasively the complex philosophical disputes of Hsün-tzu's day. He lived when the "hundred schools" of thought were competing for adherents in the Chinese marketplace of ideas, and, in order to meet the objections of his opponents and advance the Confucian position, he introduced a more rigorous style of writing that emphasized topical development, sustained reasoning, detail, and clarity.

Social and ethical stress

Hsün-tzu's primary concern was with social philosophy and ethics, as evidenced by the content of his essays: 18 of the 32 fall solely within these areas, and the remainder fall partly so; even the technical, linguistically oriented "Rectification of Names" is liberally sprinkled with comments about the adverse social consequences attending the abuse and misuse of language.

Among his other famous essays, "A Discussion of Music" became the classic work on the subject in China. Here, too, social issues are under consideration as Hsün-tzu discusses the importance of music as a vehicle for expressing human emotions without generating interpersonal conflict. Almost equally celebrated is "A Discussion of Heaven," in which he attacks superstitious and supernatural beliefs, and again the focus is social: one of the main themes of the work is that unusual natural phenomena (eclipses, etc.) are no less natural for their irregularity—hence are not evil omens—and therefore men should not be concerned at their occurrence; rather should men be concerned about what lies within their power to alter, namely hunger and sickness, which are portents of famine, disease, and death.

In the essay "A Discussion of Ritual" (*li*), Hsün-tzu elaborates the concept central to his entire philosophy. The *li* constituted the "Way" of Confucianism as interpreted by Hsün-tzu, being the ritualized norms governing the mores, manners, and morals of the people. Originally the behavioral expressions of early supernatural beliefs, the historical *li* were being abandoned by an increasingly agnostic intelligentsia during the Warring States period, and Hsün-tzu—an avowed atheist—provided an ethical and aesthetic philosophical basis for these ritual practices as their religious foundation was weakening.

The *li* are the basic stuff out of which Hsün-tzu builds his ideal society, and the scholar-officials who are to govern that society have as their primary function the preservation and transmission of these ritual practices. Like all early Confucians, Hsün-tzu was opposed to hereditary privilege, advocating literacy and moral worth as the determinants of leadership positions, rather than birth or wealth; and these determinants were to have as their foundation a demonstrated knowledge of the high cultural tradition—the *li*. No less significant politically than socially, the *li* were to be employed by scholars to ensure that everyone was in a place; and officials were to employ the *li* to ensure that there was a place for everyone.

Stature and influence. Hsün-tzu's model society was never put into practice, and, like Confucius and Mencius before him, he probably died believing himself to have been a failure. Yet the rationalism, religious skepticism, concern for man in society, historical and cultural sensitivity, and fondness for ancient lore and customs that pervade his writings also pervaded Chinese intellectual life for over two millennia. No one dealt with these issues more thoroughly than Hsün-tzu, and his passionate defense of the Confucian moral vision contributed substantively to lessening the distance between the philosophical ideal and the historical reality. He has been correctly described as the molder of ancient Confucianism. Traditional China, with its extensive lands and huge population, came to be largely a Confucian state—making Hsün-tzu one of the most influential philosophers the world has ever known.

BIBLIOGRAPHY. There is no complete translation of the *Hsün-tzu*. Competent partial works are HOMER H. DUBS, *The Works of Hsüntze* (1928); and BURTON WATSON, *Hsün Tzu: Basic Writings* (1963). The most detailed exposition of Hsün-tzu's philosophy is HOMER H. DUBS, *Hsüntze: Moulder of Ancient Confucianism* (1927). FUNG YU-LAN in *A History of Chinese Philosophy*, 2nd ed., 2 vol. (Eng. trans. 1952–53), presents the standard analysis and evaluation of Hsün-tzu's views against their Chinese background. Two recent works that de-emphasize the significance of natural human evilness in Hsün-tzu's philosophy are DONALD MUNRO, *The Concept of Man in Early China* (1969); and HENRY ROSEMONT, JR., "State and Society in the *Hsün Tzu*," in *Monumenta Serica* (1971).

(He.Ro.)

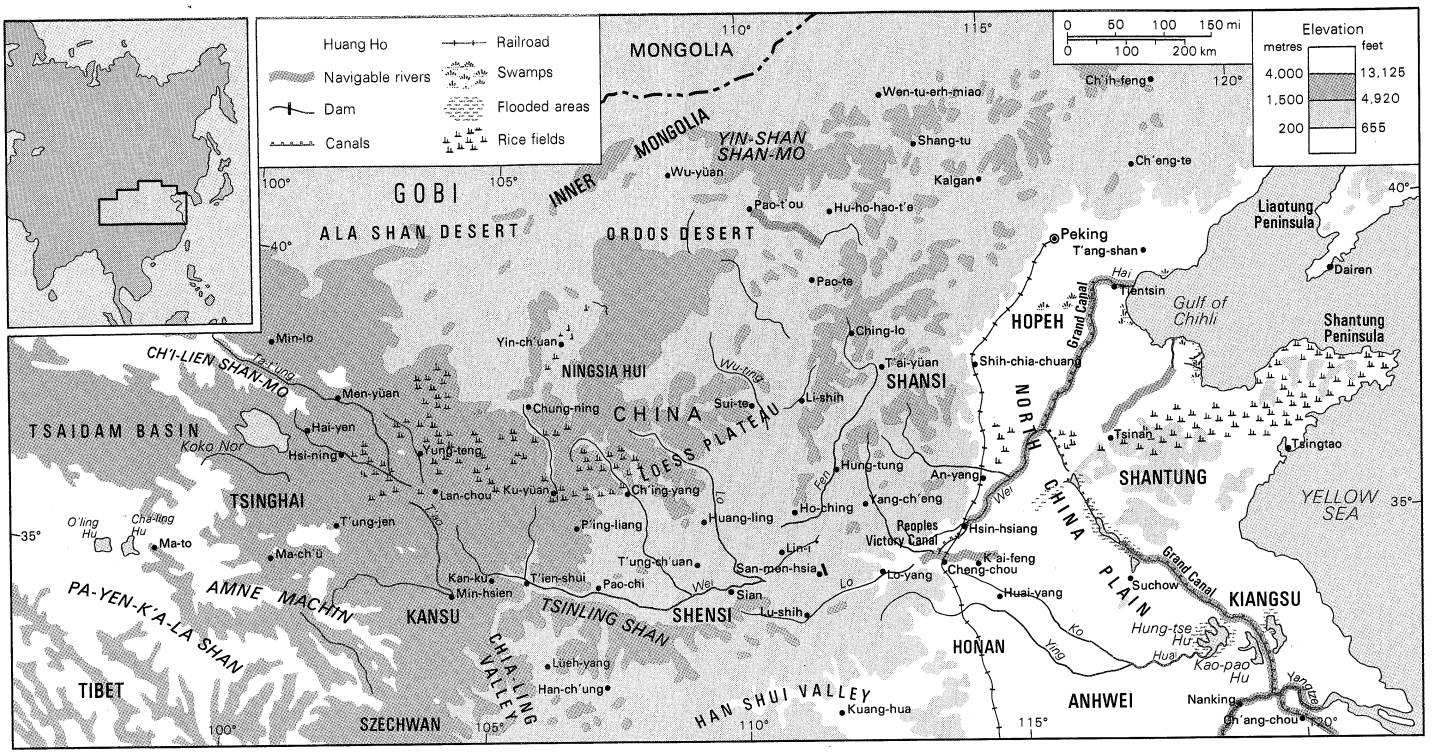
Huang Ho (River)

The Huang Ho (Huang He in Pin-yin romanization) or Yellow River, one of the most northerly and second longest of China's rivers, flows from the eastern highlands of the Tibetan Autonomous Region 3,011 miles (4,845 kilometres) to the Yellow Sea. It passes through the provinces of Tsinghai and Kansu, the autonomous regions of Ningsia Hui and Inner Mongolia, and then into the provinces of Shensi, Shansi, Honan, and Shantung as a shifting, turbulent, muddy stream that often overflows its banks and sends floodwaters across the North China Plain. For this it has been called "China's Sorrow" and "The Ungovernable." The name Huang (Yellow) comes from the fine, loess soil that it brings down with it. More than 110,000,000 people live in the basin of the Huang Ho, which flows past the cities of Lan-chou, Pao-t'ou, Sian, T'ai-yüan, Lo-yang, Cheng-chou, K'ai-feng, and Tsinan. It drains an area of 288,000 square miles (745,000 square kilometres).

Chinese civilization arose in the lower valley of the Huang Ho, which is mentioned in ancient Chinese writings of the 3rd millennium BC. During the reign of one Yao, runs the record, catastrophic flooding occurred on both the Huang Ho and Yangtze River, inundating all of the North China Plain. Regular records have been kept since the 6th century BC of major floods and also of changes in the river's course. Water levels have been studied since 1736. The first European to explore the upper reaches of the Huang Ho was a Russian traveller, Nikolay Przhevalsky, in 1879 and 1884. Since the 1950s the river has been studied intensively by Soviet scientists as well as by the Chinese.

The basin. The Huang Ho is divided into two distinct parts: the mountainous upper basin and the plains section, which is subdivided into the middle and lower basins. The broad highlands of the mountainous part rise 1,000 to 1,300 feet (300 to 400 metres) above the river and its tributaries. The highlands are built upon crystalline rocks, sometimes visible as eroded outcroppings on the surface. On the plateau these rock systems are covered with thick layers of friable deposits, mainly loess. The loess strata have thicknesses of 160 to 200 feet and in some places as much as 500 feet, extending eastward from the highlands of Tibet all the way to the North China Plain. Through this loose loam the river has cut deep valleys, carrying away with it huge quantities of silt; the easily eroded loess soil accounts for the instability of the river bed.

Downstream the Huang Ho Basin becomes fan shaped as it broadens out across the North China Plain, which is broken only by the low Shantung Hills. The plain consists of fine silt brought down by the river, over sand and gravel deposited when the sea receded in the geological past. Along the river are occasional areas of sand dunes 15 to 30 feet in height. The plain, which contains a number of old beds of the river, is China's rice granary.



The Huang Ho Basin and its drainage network.

The upper basin. The Huang Ho originates at an altitude of about 15,000 feet in the Pa-yen-k'a-la Shan (Pa-yen-k'a-la Mountains), in the eastern Tibetan Highlands. In its upper reaches the river crosses two large lakes, Cha-ling Hu and O-ling Hu. These are shallow lakes covering an area of about 40 square miles, rich in fish and frozen in winter. The Huang Ho in this region flows generally from west to east. It enters deep gorges winding its way along the southern slopes of the Amne Machin, somewhat above the city of Lan-chou. With many rapids, its fall exceeds ten feet per mile. Populated areas are few. Past the gorges it leaves the Tibetan Highlands and flows northward across the Ordos Desert, in northern China, having flowed about 726 miles from its source. The basin upstream from that point covers an area of about 48,000 square miles, consisting chiefly of inaccessible, highly mountainous terrain with a cold climate. The few inhabitants engage in cattle breeding.

Lakes Cha-ling Hu and O-ling Hu

The middle basin. The middle part of the Huang Ho consists of a great northward loop through the Ordos Desert and then a southward course through a long trench forming the border between Shensi and Shansi provinces to the city of Cheng-chou, a distance of more than 1,800 miles. Most of the middle basin is cut through the Loess Plateau at altitudes ranging between 3,000 and 7,000 feet. The plateau contains terraced slopes as well as alluvial plains and a scattering of peaks sometimes exceeding 1,500 feet in height. Within its large loop, the Huang Ho drains an area of about 23,000 square miles. The river at first flows in a northeasterly direction for about 550 miles among bare loess hills; it has many rapids, and in a number of places it narrows. It then turns eastward and flows for another 500 miles through alluvial plains; in this stretch it has many branches, and its fall is less than four inches per mile. If it continued eastward it would flow to Peking, but instead it turns sharply to the south and flows for about 445 miles through narrow gorges with steep slopes several hundred feet in height. The river's width usually does not exceed 150–200 feet in this section. The Huang Ho gradually widens and then turns sharply to the east for another 300 miles. Here it flows through inaccessible gorges between the eastern peaks of the Tsinling Shan. The average fall in this stretch is about 14 inches per mile and increasingly rapid in the last 100 miles before the river enters the North China Plain.

This middle basin of the river, between the Tibetan Highlands and the North China Plain, contains the two longest tributaries: the 537-mile Wei Ho of Shensi Province and the 430-mile Fen Ho of Shansi Province.

The two longest tributaries

The lower basin. The lower part of the Huang Ho is about 435 miles long with an average fall of about three inches per mile. This is an area of great floods because the river bed in many places lies above the surrounding land. In the section above the city of K'ai-feng, the low-water level is between six and ten feet above the surrounding countryside; the mid-water level is between 19 and 23 feet; and the high-water level sometimes as much as 33 feet above that of the land. From K'ai-feng to the Grand Canal (Yün Ho), the riverbanks rarely exceed three to six feet in height. Marshes are common. Below the Grand Canal the height of the banks increases to 13–16 feet and in some places to 25. Several centuries ago the Chinese built dikes along the river, which in places are broad enough to accommodate villages.

The delta of the Huang Ho begins approximately 160 miles from its mouth, spreading out over an area of about 950 square miles. Built up by silt brought down by the river, the delta is swampy and covered with reeds. A sand bar at the mouth impedes navigation by boats drawing more than four feet of water at low tide; at high tide the depth on the bar is eight or nine feet.

Hydrography. *Changes in course.* In the past 4,000 years several radical changes have occurred in the Huang Ho's course; at different times the river has entered the Yellow Sea at points varying by as much as 500 miles. From 2278 BC to 602 BC, when it occupied its northernmost course, it flowed through the city of Tientsin and entered the nearby Gulf of Chihli (Po Hai). From 602 BC to AD 70, both the river and its mouth shifted to the south of the Shantung Peninsula. From AD 70 to 1048 the Huang Ho again shifted north, taking up a course much along its present bed.

From 1048 to 1194 changes in the course of the river occurred farther inland, where the river enters the North China Plain. In 1194 the river occupied its southernmost course, the mouth having shifted to the southern edge of the delta. In that year, as a result of the rupturing of the protecting dikes, a second arm of the Huang Ho was created in the southern part of the Shantung Peninsula. From 1289 to 1324 the river took over the bed of the Ko Ho and a large part of the Huai Ho, later returning to its old bed. It was stable for more than 500 years until 1854,

when it again began to move farther to the north of the Shantung Peninsula, where it now is located.

Upstream alterations in course

In the upper and middle sections of the Huang Ho, alterations of the riverbed are comparatively slight; concave banks have been washed away, and the alluvial deposits have formed radiating convex banks, or arms, some of which are under cultivation. Numerous outcroppings of rock on the bottom of the river also cause changes in the bed. In the downstream areas, however, where the riverbed is higher than the surrounding land, changes in the bed sometimes lead to a sudden rupturing of the dikes and the flooding of extensive areas. This flooding, which occurs during periods of high water, covers the surrounding territory with huge amounts of silt.

History of floods. Breaks in the dikes have occurred throughout history. Between 960 and 1048 there were 38 breaks, and 29 from 1048 to 1194. In later years such breaches were less frequent as a result of systematic construction. The slackening of these efforts during the Taiping Rebellion led to a significant change in the course of the river in 1852–54.

In 1887 the Huang Ho burst the dikes near the city of K'ai-feng and began to flow into the Huai Ho, but engineering efforts succeeded in returning it to its former location in 1889. The flood of 1887 covered 30,000 square miles, completely burying many villages under silt. In 1889 another flooding injured nearly 1,000,000 persons and destroyed 1,500 villages. The next major flood, in 1921, wiped out hundreds of populated places, mainly near the mouth of the delta. In the flood of 1933 more than 3,000 populated places were submerged, 3,600,000 people injured, and 18,000 killed. Other floods occurred in 1938 and 1949.

The delta of the Huang Ho is one of the most active in the world. In 1870–1970 it pushed outward into the sea an average of more than 12 miles. Some outlying parts have been expanding even more rapidly: one area grew 6 miles during the period 1949–51, and another grew more than 15 miles in 1949–52.

Flowage and siltation. The Huang Ho carries an average annual volume of about 11.6 cubic miles (48.2 cubic kilometres) of water down to the sea—as much as 16.8 cubic miles in high-volume years and as little as 4.8 to 6.0 in low-volume years. There are also seasonal variations in its volume. More than half of the annual precipitation falls during the rainy season, July to October. The average annual precipitation for the entire basin is about 18½ inches, but its distribution is very uneven. In some years the bulk of the river volume comes from its tributaries. In the upriver areas the main source is snowfall in the mountains, with the high-water level (33 feet) occurring in the spring. The highest water levels in the middle and lower parts of the river (10 to 23 feet) occur in July and August. The maximum flow of water near Lan-chou is 7,000–8,000 cubic yards per second; near Lung-men, 13,000; and in the lower parts of the river, 47,000 (as recorded in 1943).

The Huang Ho is the world's muddiest river. It carries along about 57 pounds of silt per cubic yard of water, as compared with two pounds for the Nile, seven for the Amu Darya, and 17 for the Colorado. Floodwaters may contain up to 1,200 pounds of silt per cubic yard of water (70 percent by volume). The river carries down to the sea about 1,520,000,000 tons of silt a year, partly because much of the basin is composed of loess, which is loose and easily moved. Other factors are the steepness of the slopes, the rapidity of the current, and a lack of forested areas and reservoirs to check erosion and allow the silt to settle out. The irrigation system in the plain is not enough to slow the river current.

The Huang Ho freezes over in parts of its middle section for several months a year. On the great plain near K'ai-feng there are 15 to 20 ice-bound days a year but none at all farther downstream. Ice blockage is broken up with the help of bombs dropped from airplanes or sometimes artillery shelling.

Canals and dams

River development. The 4,000-year-old irrigation system of the Huang Ho, augmented by recently constructed irrigation and navigation canals, brings water to about 10,000 square miles of land. The Grand Canal, cutting across the lower river, runs for more than 1,100 miles from Peking in the north to Hangchow in the south. The People's Victory Canal, completed in 1953, runs for 30 miles parallel to the Peking–Hankow railroad and links the Huang Ho north of Cheng-chou with the Wei Ho at Hsin-hsiang. It raises the level of the Wei Ho, improving navigation between Hsin-hsiang and Tientsin and also providing a network of irrigation channels between Hsin-hsiang and the Huang Ho.

The government has plans for the development of "China's Sorrow" into a productive watercourse. A 1,000,000-kilowatt power station and dam has been built at San-men-hsia (Three-Gate Gorge) on the Honan-Shansi border, 130 miles west of Chengchou; the reservoir is 120 miles long. Before the construction of the reservoir, navigation by boat was restricted to a stretch of 100 miles or so on the lower reaches of the river. Current plans envisage the construction of new dams and reservoirs, the exploitation of its hydroelectric potential (estimated at 30,000,000 kilowatts), and extension of the navigable length of the river and its tributaries. (I.V.P.)

BIBLIOGRAPHY. HERBERT CHATLEY, *The Yellow River As a Factor in the Development of China* (1939); CH'UAN-CHUN WU *et al.*, *Economic Geography of the Western Region of the Middle Yellow River* (1958).

Hudson, Henry

Henry Hudson, an English navigator and explorer, made four historic voyages in the early 17th century, seeking a short route from Europe to Asia by way of the Arctic Ocean. The Hudson River, in what is now the eastern United States, and Hudson Strait and Hudson Bay, in what is now Canada, are all named for him. In 1607 and again in 1608, he attempted to discover a Northeast Passage to China. In the 1607 voyage he explored the coasts of the Svalbard (Spitsbergen) archipelago, to the north of Norway, to about latitude 80° N and rediscovered Jan Mayen Island to the east of Greenland (originally believed to have been discovered in the 15th century). His accounts of immense numbers of whales and walrus at Svalbard led the English and Dutch to establish a whaling industry there. On his third voyage in 1609, sailing for the Dutch, he laid the basis for that nation's claims to the Delaware and Hudson rivers in what is now the eastern United States. On his final voyage, from 1610 to 1611, he established an English claim to Hudson Bay that was later expanded to cover much of Canada. He is thought to have died in 1611.

The first and second voyages. Of Hudson's early life, nothing is known. Several Hudsons were associated with his sponsors, the Muscovy Company of London, a generation before his own time. A 1585 voyage by the English navigator John Davis, who sailed to the Arctic to make the first attempt to find a Northwest Passage from Europe to Asia, was planned in the home of a Thomas Hudson in Limehouse, now in the docks area of London's East End. Henry Hudson may have been present on that occasion and consequently developed a lifelong interest in Arctic exploration. It is certain that he was well informed about Arctic geography and that his competence as a navigator was such that two wealthy companies chose him to conduct hazardous explorations.

The search for the Northeast Passage

In the spring of 1607, sailing for the Muscovy Company, Hudson, his son John, and ten companions set forth "for to discover a Passage by the North Pole to Japan and China." Believing that he would find an ice-free sea around the North Pole, Hudson struck out northward. On reaching the edge of the polar ice pack, he followed it east until he reached Svalbard. From there he extended explorations made earlier by the 16th-century Dutch navigator Willem Barents, who had also sought a Northeast Passage to Asia.

A year later, the Muscovy Company again sent Hudson to seek a Northeast Passage, this time between Svalbard and the islands of Novaya Zemlya, which lie to the east of the Barents Sea. Finding his way again blocked by ice fields, he returned to England.

The third and fourth voyages. Shortly after his return, Hudson was lured to Amsterdam to undertake a third northeast voyage under contract to the Dutch East India Company. While there, he heard reports of two possible channels to the Pacific across North America. One of these, said to be in about latitude 62° N, was described in the logbooks of a voyage made in 1602 by an English explorer, Captain George Weymouth. The other, said to be in the vicinity of about latitude 40° N, was newly reported from Virginia by the English soldier, explorer, and colonist Capt. John Smith. Although his interest in a Northwest Passage had been aroused, Hudson agreed to return directly to Holland if his northeast voyage should prove unsuccessful.

Hudson sailed from Holland in the "Half Moon" on April 6, 1609. When head winds and storms forced him to abandon his northeast voyage, he ignored his agreement and proposed to the crew that they should instead seek the Northwest Passage. Given their choice between returning home or continuing, the crew elected to follow up Smith's suggestion and seek the Northwest Passage around 40° N. While cruising along the Atlantic seaboard, Hudson put into the majestic river that was discovered by the Florentine navigator Giovanni da Verrazano in 1524 but that was thenceforth to be known as the Hudson. After ascending it for about 150 miles (240 kilometres) to the vicinity of what is now Albany, New York, Hudson concluded that the river did not lead to the Pacific.

On his way to Holland, Hudson docked at Dartmouth, England. The English government then ordered him and the English members of his crew to desist from further explorations for other nations. His log and papers were sent to Holland, where his discoveries were soon made known.

Hudson now made ready a voyage to America to follow up Weymouth's suggestion. Weymouth had described an inlet (now Hudson Strait) where a "furious overfall" of water rushed out with every ebb tide. This phenomenon suggested that a great body of water lay beyond the strait. Hudson was confident that it was the Pacific Ocean. The British East India Company contributed £300 toward his voyage, and the Muscovy Company presumably furnished a like amount; Hudson's private sponsors included five noblemen and 13 merchants.

The voyage to Hudson Bay

Sailing from London on April 17, 1610, in the 55-ton vessel "Discovery," Hudson stopped briefly in Iceland, then proceeded to the "furious overfall." Passing through it and entering Hudson Bay, he then followed the east coast southward, rather than striking boldly westward. Finding himself in James Bay at the southernmost extremity of Hudson Bay and with no outlet to the Pacific to be found, Hudson cruised aimlessly until winter overtook him.

In the close confinement of an Arctic winter, quarrels arose. Hudson angered one of his crew, Henry Green, by first giving him a gray gown and then, when Green displeased him, taking it back and giving it to another. Some of his crew suspected that Hudson was secretly hoarding food for his favourites, and tempers flared when Hudson ordered the crew's own sea chests searched for extra victuals. Robert Juet, the mate, had been demoted, and he conspired with Green and others to mutiny. Once the homeward voyage had begun, the mutineers seized Hudson, his son, and seven others, casting them adrift in Hudson Bay in a small open boat on June 22, 1611. Although the "Discovery" sailed home to England, neither of the ringleaders returned with her, having been killed, together with several others, in a fight with Eskimos. No more was ever heard of Hudson and his small party, although in 1631 to 1632 another explorer found the ruins of a shelter, possibly erected by the castaways.

As a commander, Hudson was more headstrong than courageous. He violated his agreement with the Dutch and failed to suppress the 1611 mutiny. He played favourites and let morale suffer. In James Bay he appeared irresolute.

Yet Hudson undertook four dangerous voyages, brought his crew through an Arctic winter, and preserved his vessels amid the dangers of ice and unknown shores. He was a competent navigator who materially extended the explorations of Verrazano, Davis, and Barents. His contribution to geographical knowledge was great, while his discoveries formed the basis for the Dutch colonization of the Hudson River and for English claims to much of Canada.

BIBLIOGRAPHY. G.M. ASHER (ed.), *Henry Hudson the Navigator* (1860, reprinted 1963), provides the surviving journals of Hudson's four voyages, supplementary materials, and an important critical introduction. SAMUEL PURCHAS, *Purchas his Pilgrimes* (1625, reprinted 1906), gives the journals. LLEWELYN POWYS, *Henry Hudson* (1928), is a good popular biography. R.A. SKELTON, *Explorers' Maps* (1958, reprinted 1970), reproduces pertinent charts.

(Jo.E.C.)

Hudson Bay

Roughly oval-shaped, Hudson Bay is one of the major fringing continental seas of the Arctic Ocean, covering 316,000 square miles (819,000 square kilometres)—an area larger than that of France and Italy combined—in the Canadian depression of northeast North America. The northern exit to the bay bifurcates: on either side of the long, narrow Baffin Island, Hudson Strait, curving away to the southeast, provides a narrow link with the Atlantic Ocean; and Foxe Basin and the narrow straits of the Canadian Arctic Archipelago provide a link with the Arctic Basin to the north. The bay is named after Henry Hudson, who, in 1610, on board the aptly named "Discovery," was seeking a Northwest Passage to Asia. The east coast of Hudson Bay proper was mapped two years later; the south coast was traced in 1631; and the explorer Luke Foxe lent his name to Foxe Channel in the same year. The west coast was not mapped until the early 1820s, and the first investigations in depth of the area were made by Canadians during 1929–31, followed by hydrographic researches in 1948. During the 1954–62 period, six research cruises took place; but following observations on ice conditions in 1955 and 1956, air reconnaissance became more frequent. (For related information see NORTH AMERICA; CANADA; ARCTIC ISLANDS; NORTHWEST TERRITORIES. For historical aspects, see CANADA, HISTORY OF; NORTHWEST PASSAGE.)

Exploration and scientific study

Physical characteristics. Hudson Bay has a shallow and quite smooth floor, averaging 330 feet (100 metres) in depth, with a maximum around 900 feet. The coast, situated in a region of permanently frozen earth layers, or permafrost, is a marsh-ridden lowland fed by lake waters and turbulent rivers. In the east and northeast the shores are high and sheer, but elsewhere they are low. Coniferous woods border the southerly James Bay, the shallowest part, but most of the shore is covered with dwarf birch, willow, aspen, and bushes, growing among moss, lichen, and grass.

The eastern coast is bordered, at a distance of some 200 miles, with a set of islands and has cliffs formed of geologically ancient (over 570,000,000 years old) crystalline and sedimentary rocks. The only other islands are a small cluster at the bay exit. Bottom deposits are derived from the earth rather than from the shells of minute organisms, as in the case of the ooze of some of the coastal islands. The bottom deposits are washed in from the nearby coasts or dropped from encrustations below the "dirty ice" of Foxe Basin.

Hudson Bay has a severe continental climate. January temperatures average −20° F (−29° C), while July temperatures are only 47° F (8.3° C). Annual averages are 9.3° F (−12.6° C), but extremes have plunged to −60° F or reached 80° F in the summer. Spring is mild and cloudy, whereas summer is clear, though the bay itself is often coated with fog. Autumn starts cool, with frequent fogs, clearing later; early winter is very cold, clear, and calm, but this pattern is interrupted, after December, by strong winds and snowstorms. The thaw sets in in late April. Annual precipitation totals 12 to 20 inches (300 to 500 millimetres), all but a third of it coming in summer. Only 60 days are without frost.

Hudson Bay is filled by the numerous peripheral rivers and also by currents from Foxe Basin in the north, cre-

ating a counterclockwise general movement. Outflow occurs along the eastern Hudson Strait coast, rounding Chidley Cape (the northernmost tip of the Quebec–Newfoundland border), and passing into the Labrador Current. Flow is highest in July.

Hudson Bay has much ice, mostly local, though there is some influx of pack ice from Foxe Basin. Southern and central areas have solid, floating ice fields only during February and March; but a shore ice belt appears earlier. Freezing begins in October, and floating ice persists here and there into the following summer. Salinity increases with depth: below 80 feet it is 31 parts per thousand (‰); the layer above registers 23‰; and the upper six feet registers only 2‰ when the current is strong and ice is melting. Water temperatures can be as low as 29° F (−1.8° C) at depth in August, although surface temperatures may reach 49° F in September. Tides are important, with currents in the bay responding to the fierce tidal flow off the Labrador coast. The tidal amplitude and the associated current movements cause much mixing of waters in the region.

Biological characteristics. Hudson Bay contains a great quantity of dissolved nutrient salts, since unicellular algae, especially microscopic types, grow fast in the well-heated and illuminated upper layers. Small, shrimplike crustaceans occupy the open waters and form a food source for mollusks, sea urchins, starfish, and worms, together with many other invertebrates living on the bottom.

Variety of sea life

Fish include numerous polar plaice, cod, halibut, salmon migrating to the lakes and rivers to spawn, and also freshwater species. Seals, feeding on fish and invertebrates, inhabit the areas around openings in the ice: they include ringed, bearded, and Greenland varieties. Walrus, dolphins, and killer whales live in the northern sector, and the polar bears come down from the north to hunt seals among the ice. About 200 species of birds gather on the coasts and islands; they include ducks, gulls, eiders, loons, snow geese, swans, sandpipers, snow owls, and crows. There are also such herbivorous mammals as caribou, musk oxen, and rodents, as well as fur-bearing animals.

Human exploitation. Fishing, and hunting for sea mammals or furs, are the main local occupations. Indians inhabit the southern region between Eskimo Point (about halfway along the west coast) to Fort-George (on the eastern coast of James Bay, the embayment thrusting inland from the southeast corner of Hudson Bay); Eskimos live in the north. Population density is very low. For conservation purposes, the Canadian government has designated the whole Hudson Bay Basin a "mare clausum" (closed sea).

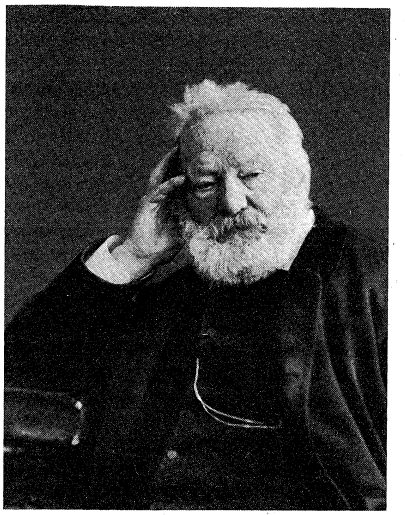
BIBLIOGRAPHY. FARLEY MOWAT (ed.), *Ordeal by Ice* (1960), is a collection of original sketches of British expeditions carried out through the 16th-19th centuries, including a history of the discovery of Hudson Bay. See also F.G. BARBER and C.J. GLENNIE, *On the Oceanography of Hudson Bay: An Atlas Presentation of Data Obtained in 1961* (1964), issued by the Canadian Department of Mines and Technical Surveys.

(M.M.A.)

Hugo, Victor

A poet, novelist, dramatist, critic, and polemicist of vast creative energies, Victor Hugo first gained fame as the champion of the Romantic movement against the classicism and temperate good sense of a long-entrenched public literature in France. Popular during his lifetime as the author of such novels as *Les Misérables*, *Notre-Dame de Paris*, and *Toilers of the Sea*, he is critically important today chiefly for his poetry.

Early years (1802–30). Born at Besançon on February 26, 1802, Victor was the third son of Joseph-Léopold-Sigisbert Hugo, a major and, later, general in Napoleon's army. The paternal family consisted of farmers and artisans; the maternal, seamen and lawyers. Victor claimed that he came of "a stock both Breton and Lorrain," but it was Paris, where he went to live with his mother when he was two, that he called the birthplace of his soul.



Hugo, photograph by Nadar (Gaspard-Félix Tournachon).
Archives Photographiques

Hugo's childhood was coloured by his father's constant travelling with the imperial army and by the disagreements that soon alienated his parents from one another. His mother's royalism, a development perhaps subsequent to marriage, and his father's loyalty to successive governments—the Convention, the Empire, the Restoration—reflected their deeper incompatibility. It was a chaotic time for Victor, continually uprooted from Paris to set out for Elba or Naples or Madrid, yet always returning to Paris, usually with his mother alone. Such an existence at least fed his mind on contemporary history, if it did not provide a settled upbringing. Nevertheless, his intellectual education did achieve a solid basis, especially in Latin literature, and his private reading, though without order, was omnivorous. A few clear images from this period stayed in his mind: of Spain, especially, and of the garden of the Feuillantines in the Latin Quarter—a paradise, long since disappeared, that seemed to hold all the poetry of childhood. The fall of the empire gave him, from 1815 to 1818, a time of uninterrupted study at the Pension Cordier and the Lycée Louis-le-Grand, after which he matriculated at the law faculty at Paris, where his studies seem to have been purposeless and irregular. Memories of his life as a poor student later inspired the figure of Marius in his novel *Les Misérables*.

Indeed, from 1816, at least, he had conceived ambitions other than the law; he was already filling notebooks with verses that he later called "nonsense I perpetrated before my birth": translations—particularly from Virgil—two tragedies, a play, elegies. Encouraged by his mother and excited by the dazzling example of François-René de Chateaubriand, the author, diplomatist, and statesman, he founded a review, the *Conservateur Littéraire* (1819–21), in which his own articles on the poets Alphonse de Lamartine and André de Chénier stand out. His mother died in 1821, and a year later Victor married a childhood friend, Adèle Foucher, to whom he wrote some charming love letters, the *Lettres à la fiancée*, which were published posthumously. They had five children, four of whom survived: Léopold (died 1823), Léopoldine (1824–43), Charles (1826–71), François-Victor (1828–73) and Adèle (1830–1915). In the same year as his marriage, he published his first book of poems, *Odes et poésies diverses*, which earned him a pension from Louis XVIII. Behind Hugo's concern for classical form and his political inspiration, it is possible to recognize a personal voice and his own particular vein of fantasy. Fantasy had become fashionable, and Hugo developed his gift for it in his first published novel, *Han d'Islande* (1823), which in 1825 appeared in an English translation with illustrations by the caricaturist George Cruikshank. Charles Nodier, a journalist and himself author of novels of fantasy, was enthusiastic about the novel and drew Hugo into the group of friends, all devotees of Romanticism,

Early works

who met regularly at the Bibliothèque de L'Arsenal, where he was curator. Hugo continued to frequent this literary circle, called the Cénacle, though he sometimes professed indifference to the theoretical distinctions it debated. He shared also in launching a new review of moderate tendencies, the *Muse Française* (1823–24), to which he contributed essays on the Roman Catholic apologist Félicité-Robert de Lamennais and the poets Alfred de Vigny, Lord Byron, and Walter Scott. In 1824 he published a new collection, *Nouvelles Odes*, and followed it two years later with an exotic romance, *Bug-Jargal*, developed from a short story written in 1819. In 1826 he also published *Odes et ballades*, an enlarged edition of his previously printed verse, the latest of these poems being brilliant variations on the fashionable Romantic modes of mirth and terror. The youthful vigour of these poems was also characteristic of another collection, *Les Orientales*, which appealed to the Romantic taste for Oriental local colour and to sympathy for the Greek War of Independence and was an instant success. Hugo here displayed such technical mastery that some critics, Sainte-Beuve among them, regarded the book as nothing more than "a splendid throne raised to the honour of pure art"—a judgment in which the modern reader can share. Yet it can also be remarked that the poet, skillfully employing a great variety of metres in his verse and using ardent and brilliant imagery, was gradually shedding the legitimist royalism of his youth. It may be noted, too, that "Le Feu du Ciel," a visionary poem, forecast those he was to write 25 years later. The fusion of the contemporary with the apocalyptic is a particular mark of Hugo's genius.

The preface to *Cromwell*

Hugo emerged as a true Romantic, however, with the publication in 1827 of his verse drama *Cromwell* and a once-famous preface. The subject, with its near-contemporary overtones, was that of a national leader risen from the people who sought to be crowned king; but the play's reputation rested largely on the long, elaborate preface, in which Hugo proposed a doctrine of Romanticism that for all its intellectual moderation was extremely provocative. He demanded a verse drama in which the contradictions of human existence—good and evil, beauty and ugliness, tears and laughter—would be resolved by the inclusion of both tragic and comic elements in a single play. Such a type of drama, which abandoned the formal rules of classical tragedy for the freedom and truth to be found in Shakespeare, would provide a vantage point from which to view history and men. *Cromwell* itself, though immensely long and almost impossible to stage, was written in verse of great force and originality.

Success (1830–51). The defense of freedom and the cult of an idealized Napoleon in such poems as the ode "À la Colonne" and "Lui" brought Hugo into touch with the liberal group of writers on the newspaper *Le Globe*, and his move toward liberalism was strengthened by Charles X's restrictions on the liberty of the press as well as by the censor's prohibiting the stage performance of his play *Marion de Lorme* (1829), the story of a courtesan purified by love, in which the character of Louis XIII was portrayed unfavourably. He immediately retorted with *Hernani*, the first performance of which, on February 25, 1830, gained victory for the young Romantics over the traditional Classicists in a literary battle that became known as the "battle of *Hernani*." In this play he extolled the Romantic hero already sketched in the character of Didier in *Marion de Lorme:* the noble outlaw at war with society, dedicated to a passionate love and driven on by inexorable fate. The actual impact of the play owed less to the plot than to the sound and beat of the verse, softened only in the elegiac passages spoken by Hernani and Doña Sol. The triumph of *Hernani* was due to the student audience, for it was an homage from a youth to youth.

Hugo had derived his early renown from his plays; he gained wider fame in 1831 with his historical novel *The Hunchback of Notre-Dame*, an evocation of medieval life under the reign of Louis XI. The novel condemned a society that, in the persons of Frollo the archdeacon and Phoebus the soldier, heaped misery on the hunchback Quasimodo and the gypsy Esmeralda. The theme touched the public consciousness more deeply than had that of his previous novel, *Le Dernier Jour d'un condamné* (1829), the story of a condemned man's last day, in which Hugo launched a humanitarian protest against the death penalty. He later renewed this theme in *Claude Gueux* (1834). While *Notre-Dame* was being written, Louis-Philippe, a constitutional king, had been brought to power by the students and the liberal bourgeois in the three days of the July Revolution. Hugo composed a poem in their honour—*Dicté après juillet 1830;* it was a forerunner of much political verse.

Four books of poems came from Hugo in the period of the July monarchy: *Les Feuilles d'automne* (1831), intimate and personal in inspiration; *Les Chants du crépuscule* (1835), overtly political; *Les Voix intérieures* (1837), at once personal and philosophical in tone; *Les Rayons et les ombres* (1840), in which the poet, renewing these different themes, indulged his gift for colour and picturesque detail. Hugo was not content merely to express personal emotions; he wanted to be the "sonorous echo" of his time and thus to fulfill the poet's function as he saw it. Political and philosophical problems were integrated with the religious and social disquiet of the period; one poem evoked the misery of the workers, another praised the efficacy of prayer. He addressed many poems to the glory of Napoleon, though he shared with his contemporaries the reversion to republican ideals. Independent of party and church, he restated the problems of his century and the great human questions always asked by man, and he spoke with a warmhearted eloquence and reasonableness that moved men's souls. Such was his influence that he even gained an admirer at court: the young Duchess of Orléans, whose protection he enjoyed after 1837.

Dramatic works

So intense was Hugo's creative activity during these years that, having at last achieved a stage production of *Marion de Lorme* in 1831, he continued to pour out plays. There were two motives for this: first, he needed a platform for his political and social ideas, and, second, he wished to write parts for a young and beautiful actress, Juliette Drouet, with whom he had begun a liaison in 1833. Juliette had little talent and, perhaps persuaded by her lover, soon renounced the stage in order to devote herself exclusively to him, becoming the discreet and faithful companion she was to remain until her death in 1883. The first of these plays was another verse drama, *Le Roi s'amuse* (1832), set in Renaissance France and depicting the frivolous love affairs of Francis I while antithetically revealing the noble character of his court jester. This, like *Marion de Lorme*, was at first banned but was used by Giuseppe Verdi as the libretto of his opera *Rigoletto*. Three prose plays followed: *Lucrèce Borgia* and *Marie Tudor* in 1833 and *Angelo, tyran de Padoue* in 1835. *Ruy Blas*, a play in verse, appeared in 1838. The hero, Ruy Blas, becomes prime minister of Spain, wins the love of the queen, and institutes various social reforms. But there is more to this play than its romantic theme of a lowborn poet's passionate love of a queen; the whole piece is filled with the democratic faith and the ideals of reform for which the hero is the author's spokesman. The historical and geographical diversity of the plays is striking and was emphasized by Hugo himself when he came to classify his works on a basis of epochs and countries. Germany, he realized, was absent, and he filled the gap by writing *Le Rhin* in 1842, a book of travel composed of letters written during a series of journeys in the Rhine valley (1839–40), and *Les Burgraves* the following year. An attempt to portray greatness and decadence in German medieval history, *Les Burgraves* emerged as an epic melodrama and a vivid theatrical extravaganza that was hissed off the stage. The failure of *Les Burgraves* interrupted Hugo's career as a dramatist; exile and preoccupation with other tasks discouraged him from writing for the stage, though from time to time he composed the plays, giving rein to his fantasy, that make up the posthumous volume *Théâtre en liberté*, published in 1886.

Hugo's literary achievement was recognized in 1841 by his election, after three unsuccessful attempts, to the Académie Française and by his nomination in 1845 to the Chamber of Peers. From this time he almost ceased to publish, partly because of the demands of society and political life but also as a result of personal loss: his daughter Léopoldine, recently married, was accidentally drowned with her husband in September 1843. He was travelling with Juliette Drouet when it happened and learned the news from a paper. His intense grief found some mitigation in poems that later appeared in *Les Contemplations*, a volume that he divided into "Autrefois" and "Aujourd'hui," the moment of his daughter's death being the mark between yesterday and today. He found relief above all in working on a new novel, which became *Les Misérables*, published in 1862 after work on it had been set aside for a time and then resumed. Other distractions were a liaison with the wife of an artist, Léonie Biard, herself a writer, and, because of his close friendship with Louis-Philippe and his daughter-in-law, the Duchess of Orléans, an assiduous attendance at court festivities. It seemed as if he wished to forget his personal sorrows and the menace of social unrest, but in fact his writings both in verse and prose reflect his troubled mind, and it was in his grief that his genius matured.

Political life

Hugo's political intuition was confirmed when the revolution of February 1848 was seen to be less political than social in its aims, as compared with that of 1830. Hugo's plan to form a regency under the Duchess of Orléans was unsuccessful, and he entered the political arena of the Second Republic, first as a deputy for Paris in the Constituent Assembly and later in the Legislative Assembly, where he upheld his ideals of education for all and the well-being of the people. Such aspirations and his loyalty to the name of Bonaparte led him to support the candidature of Prince Louis-Napoléon for the presidency in 1848. The more the President evolved toward an authoritarianism of the right, however, the more Hugo moved toward the assembly's left. When in December 1851 a coup d'etat took place, which eventually resulted in the Second Empire under Napoleon III, Hugo made one attempt at resistance and then fled to Brussels.

Exile (1851–70). His exile was to last until the return of liberty and the reconstitution of the republic on September 4, 1870. Enforced at the beginning, exile later became a voluntary gesture and, after the amnesty of 1859, an act of pride. He remained in Brussels for a year until, foreseeing expulsion, he took refuge on British territory. He first established himself on Jersey, where he remained from 1852 to 1855. When he was expelled from there he moved to the neighbouring island of Guernsey.

Satire and historical works

Immersed in politics as he was, he devoted the first writings of his exile to satire and recent history: *Napoléon le Petit* (1852), an indictment of the "little Napoleon" (Napoleon III) as opposed to the "great" (Napoleon I), and *Histoire d'un crime*, a day-by-day account of Louis Bonaparte's coup as seen by a dissentient witness. His return to poetry was an explosion of wrath: *Les Châtiments* (1853). This collection of poems unleashed his anger against the new emperor as well as, on a technical level, it freed him from his remaining classical prejudices and enabled him to achieve the full mastery of his poetic powers. During this exile of nearly 20 years he produced the most extensive part of all his writings and the most original. It was accomplished in virtual solitude—his only companion, except for his family and the ever-faithful Juliette, was the restless sea. *Les Châtiments* ranks among the most powerful satirical poems in the French language. The ironic section titles parody the claims of the new leader in his attempt to justify the coup—"Society is saved," "Order is Re-established," etc. On the other hand, Bonaparte is presented as a thief and a killer. All Hugo's future verse profited from this release of his imagination: the tone of this collection of poems is sometimes lyrical as in "Stella," sometimes epic as in "L'Expiation," sometimes moving as in "Souvenir de la nuit du 4," most often virulent as in "Nox" and such poems in which Hugo heaps outrage on the ruler of France in an impassioned rhetoric. The inspiration behind *Les Châtiments*, with its undertone of national and personal frustration, was borne out by a poetic breath that was to be sustained and developed in all his subsequent work, however flawed by repetition and blemishes that inevitably go with it.

Despite the satisfaction he derived from his political poetry, Hugo wearied of its limitations and, turning back to the unpublished poems of 1840–50, set to work on a volume of "pure poetry," *Les Contemplations* (1856). In the meantime, however, he had become obsessed with the preternatural, or the *surréel*, as he called it: table turning, dreams, knockings, and other such phenomena were recorded in the private notebooks. *Les Contemplations* contains the purest of his poetry—the most moving because the memory of his dead daughter is at the centre of the book, the most disquieting, also, because it transmits the haunted world of a thinker. In poems such as "Pleurs dans la nuit" and "La Bouche d'ombre," he reveals a tormented mind that struggles between doubt and faith in its lonely search for meaning and significance.

The apocalyptic approach to reality dominated the poet's inspiration; it was the source of those epic or metaphysical poems for which there has been growing admiration among modern readers: *La Fin de Satan* and *Dieu*—poems of vast scope, the first a tapestry of resplendent visions, the second tense with urgent debate, both of them confrontations of the problem of evil. Written between 1854 and 1860, they were not published until after his death because his publisher preferred the little epics based on history and legend, the "Petites Épopées" of *La Légende des siècles* (1859). These, which for Hugo himself formed a single poem with the metaphysical epics, are the summit of his art; they display all his spiritual power without sacrificing his exuberant capacity to tell a story. Hugo's personal mythology of the human struggle between good and evil lies behind each of the legends: Eve's motherhood is exalted in "Le Sacre de la femme"; mankind liberating itself from all religions in order to attain divine truth is the theme of "Le Satyre"; and "Plein Ciel" proclaims, through utopian prediction of men's conquest of the air, the poet's conviction of indefinite progress toward the final unity of science with moral awareness. Hugo was a prophet of the modern world at a moment when his century was foundering in religious disbelief.

Success of *Les Misérables*

Yet he was prudent enough not to overtax his readers' comprehension. After the publication of three long books of poetry, he turned to prose and took up his abandoned novel, *Les Misérables*. Its extraordinary success with readers of every type when it was published in 1862 brought him instant popularity in his own country, and its speedy translation into many languages won him fame abroad. The plot, that of a detective story, is as well the epic of the people of Paris. Its author claimed it as a "religious" work; and indeed by means of its characters, sometimes a little larger than life yet always vital and engaging, and by its re-creation of the swarming Parisian underworld, the main theme of man's ceaseless combat with evil clearly emerges while the whole gives a faithful picture of the ebb and flow of life. Hugo relives his youth in this vast novel, and the ordinary reader can still be moved by its author's search for justice; even the more critical can admire the novel's complex structure, in spite of lengthy digressions in the narrative.

When the translation of Shakespeare's works by his son François-Victor was nearing its end, Hugo wrote a preface. It became an essay, *William Shakespeare*, published for Shakespeare's tercentenary (1864), in which he elaborated his theory of the poetic imagination as divided between light and darkness, joy and grief. The experience of the great dramatist, he believed, confirmed his own. In this essay, Hugo developed a hierarchy of human genius—Job, St. Paul, Dante, Cervantes, Shakespeare—that went beyond the customary classifications, and there is little doubt that he saw himself in these ranks. Two novels also showed this tendency toward the grandiose: *Les Travailleurs de la mer* (1866), dedicated to the island of Guernsey and its sailors, the "toilers" of

the sea; and *L'Homme qui rit* (1869), a curious baroque novel about the English people's fight against feudalism in the 17th century, which takes its title, "the man who laughs," from the perpetual grin of its disfigured hero. His last novel, *Quatrevingt-treize*, planned at the same period but finally written in Guernsey in 1873, centred around the tumultuous year 1793 in France and portrayed human justice and charity against the background of the French Revolution. These were not the only works of his exile; on his return to France he left behind many manuscripts that were to be published later.

Return to Paris

Last years (1870–85). The defeat of France in the Franco-German War and the proclamation of the Third Republic brought Hugo, faithful to his vow, back to Paris. He agreed to play his part in public affairs as a deputy in the National Assembly (1871) but resigned the following month. Though he still fought for his old ideals, he no longer possessed the same energies. The trials of recent years had aged him, and there were more to come: in 1863 his daughter Adèle had eloped to America, and nine years later she returned to France insane; in 1868 he had lost his wife, a profound sadness to him after their many years together; in 1871 his son Charles died and in 1873, François-Victor. Increasingly detached from life around him, the poet of *L'Année terrible*, in which he recounted the siege of Paris during the "terrible year" of 1870, had become a national hero. He went on with his task, his gaze fixed on the horizons closing in. He had his grandchildren, Georges and Jeanne, for whom he wrote genial verse on the art of being a grandfather, *L'Art d'être grand-père* (1877). In March 1871, outraged by the terms of the peace treaty after the siege, insurrectionists seized power in Paris and established a council called the Commune, which was not suppressed until the end of May. The episode of the commune, which met with Hugo's disapproval, found him once more at Brussels, and he was again expelled for giving shelter to the defeated rebels. After a temporary refuge in Luxembourg he returned to Paris, where he was elected senator. During 1872–73 he was in Guernsey for the writing of *Quatrevingt-treize* and the preparation of his remaining works for publication. In 1878 he was stricken by cerebral congestion, but he lived on for some years in the Avenue d'Eylau, renamed Avenue Victor-Hugo on his 80th birthday. Two years after the death of Juliette, his faithful companion, he died on May 22, 1885. He was given a national funeral, lay in state under the Arc de Triomphe, and was buried in the Panthéon, having been taken there in a pauper's hearse as he had wished.

Reputation. Hugo's enormous output is unique in French literature; it is said that he used to write each morning 100 lines of verse or 20 pages of prose. "The most powerful mind of the romantic movement," as he was described in 1830, laureate and peer of France in 1845, he went on to assume the role of an outlawed sage who, with the easy consciousness of authority, put down his insights and prophetic visions in great prose and verse, becoming at last the genial grandfather of popular literary portraiture and the national poet who gave his name to a street in every town in France.

This instinctive recognition of Hugo as a great poet at the time of his death was followed by a period of critical neglect. A few of his poems were remembered and *Les Misérables* continued to be widely read; his poetic art was prized by the few; only the generosity of his ideas and the warmth of their expression still moved the public mind. What the novelist François Mauriac has called "the heroism of banality" made him a poet of the common man, because he knew how to write with simplicity and power of common joys and sorrows. There was another side to this people's poet—what Paul Claudel called his "panic contemplation" of the universe, the numinous fear that penetrates the sombre poems of the 6th book of *Les Contemplations*, *La Fin de Satan*, and *Dieu*. He knew, as the poet Charles Baudelaire wrote of him, how to evoke the ambience of mystery. His knowledge of the resources of French verse and his technical virtuosity in metre and rhyme, moreover, rescued French poetry from the sterility of the 18th century. The poets who succeeded him—Baudelaire, Rimbaud, Claudel, Apollinaire—are indebted to his vast poetic experience.

André Gide, when asked whom he considered the greatest French poet, replied "Victor Hugo, alas," explaining that if it was a regrettable fact at least it was fact. The time has come to replace this ambiguous verdict by that of another poet, Léon-Paul Fargue: "Victor Hugo, un poète d'avenir," which might be translated as "a poet who has time on his side." Poet of the real and the surreal, Hugo must be accepted whole.

MAJOR WORKS

POETRY: *Odes et poésies diverses* (1822); *Nouvelles Odes* (1824); *Odes et ballades* (1826, enlarged 1828 to include earlier poems and collections); *Les Orientales* (1829); *Les Feuilles d' automne* (1831); *Les Chants du crépuscule* (1835); *Les Voix intérieures* (1837); *Les Rayons et les ombres* (1840); *Les Châtiments* (1853); *Les Contemplations* (1856); *La Légende des siècles* (3 series: 1859, 1877, 1883); *Les Chansons des rues et des bois* (1865); *L'Année terrible* (1872); *L'Art d'être grand-père* (1877); *Les Quatre Vents de l'esprit* (1881). Posthumously published poems and collections include the unfinished *La Fin de Satan* (1886, written 1854–60) and *Dieu* (1891, written 1855); *Toute la lyre* (1888, 2 series; 1893, 1 series); *Les Années funestes, 1852–1870* (1898).

NOVELS: *Han d'Islande* (1823); *Bug-Jargal* (1826); *Le Dernier Jour d'un condamné* (1829); *Notre-Dame de Paris* (1831); *Claude Gueux* (1834); *Les Misérables* (1862); *Les Travailleurs de la mer* (1866); *L'Homme qui rit* (1869); *Quatrevingt-treize* (1874).

DRAMA (POETIC): *Cromwell* (1827); *Marion de Lorme* (written 1829); *Hernani* (1830); *Le Roi s'amuse* (1832); *Ruy Blas* (1838); *Les Burgraves* (1843). (PROSE): *Amy Robsart* (1828, from Sir Walter Scott's novel *Kenilworth* of 1821); *Lucrèce Borgia* (1833); *Marie Tudor* (1833); *Angelo, tyran de Padoue* (1835). *Théâtre en liberté* (posthumously 1886, collects plays written 1854 onward).

PROSE (CRITICAL): The preface to *Cromwell* (1827); *Littérature et philosophie mêlées* (1834); *William Shakespeare* (1864). (POLITICAL): *Napoléon le petit* (1852); Histoire *d'un crime* (1877, written 1852); *Actes et paroles*, 4 series: 1, *Avant l'exil*, 1841–51; 2, *Pendant l'exil*, 1852–70; 3 and 4, *Depuis l'exil*, 1870–85 (collection of articles, essays, etc., from papers and periodicals). (DESCRIPTIVE): *Le Rhin* (1842); *Alpes et Pyrenées* (posthumously 1890); *La France et la Belgique* (posthumously 1894). The posthumous *Choses vues*, 2 vol. (1887–99), contains descriptive pieces, notes, and observations.

TRANSLATIONS: An edition of the *Works of Victor Hugo*, 10 vol., was published in 1907. Many translations from Hugo's poems were made during the 19th century, none satisfactory. Selected *Poems*, ed. by Alfred T. Baker, appeared in 1934; *Selected Prose and Verse*, ed. by W.G. Hartog, in 1927. In 1899 a work entitled *The Novels of Victor Hugo: Complete and Unabridged* was published, and there have since been many adequate translations of individual novels, especially of *Les Misérables* and *Notre-Dame de Paris* (including a 1956 version of the latter by Lowell Blair, under the book's usual English title of *The Hunchback of Notre Dame*). *The Dramas of Victor Hugo*, 4 vol., trans. by I.G. Burnham, was published in 1895–96. In 1964 *Choses vues* was translated as *Things Seen*, by David Kimber.

DRAWINGS: The main collection is in the Musée Victor-Hugo, Paris. It consists of some 350 drawings in many media. They range from the naturalistic representations of buildings and places of his early period, through a more visionary group dating from 1848–50 (among them "La Vue de Paris," "Les Trois Arbres," and "Le Burg à la Croix") to the anticipations of Surrealism of the period *c.* 1853–*c.* 1871. Many drawings of this period are in albums, or on the pages of the manuscripts of Hugo's writings, in the Bibliothèque Nationale, Paris.

BIBLIOGRAPHY. E.M. GRANT, *Victor Hugo: A Select and Critical Bibliography* (1967); and *The Career of Victor Hugo* (1945), a good all-around study in English; J.B. BARRERE, *Hugo, l'homme et l'oeuvre*, 7th ed. (1969), a standard one-volume study for the student and scholar; and *La Fantaisie de V. Hugo*, 3 vol. (1949–60), a thorough exploration of Hugo's imagination in its happy mood; ANDRE MAUROIS, *Olympio ou la vie de V. Hugo* (1954; Eng. trans., *Victor Hugo*, 1956), the best biography for the layman, better in its first half; PIERRE ALBOUY, *La Création mythologique chez Victor Hugo* (1964), a thorough study of Hugo's epic and myths; JEAN GAUDON, *Le Temps de la contemplation . . ., 1845–1856* (1969), a brilliant study of Hugo's imagination in its darker shades.

(J.-B.Ba.)

Human Behaviour, Development of

The essential phenomena of human development are enlargement and change. Bone, muscle, and fat become larger and heavier but retain the same basic structure throughout life. In contrast, psychological processes (including observable behaviours) undergo major changes during the first dozen or so years of life. Delineation of these changes deserves close attention for a number of reasons. First, knowledge of how a structure develops aids in understanding how that structure functions in its mature state. For example, the sociological principles that describe how contemporary cities function throughout the world are different from the principles describing how those cities grew to their current size and form.

Similarly, an understanding of the psychological functions of an adult is facilitated if one knows how the adult developed. Second, the organization of motives, beliefs, anxieties, skills, defenses, and cognitive functions varies with age. Attempts to teach a child a new skill should acknowledge the organization of his abilities at that time. Finally, appreciation of developmental change can serve the practical need to predict the quality of a child's adult behaviour and to anticipate possible personality disturbance from knowledge of a child's life experiences. Many developmental theorists assume that what is learned in childhood directs future behaviour and provides a thread of continuity in human development. It is also assumed that there are especially sensitive periods in development and that experiences during these periods are most likely to predispose the individual to develop particular ways of acting as an adult.

Sensitive periods

The notion of sensitive periods implies that specific psychological processes are changing rapidly during particular phases of development and that environmental changes that intrude during a sensitive period are likely to alter the course of the behaviour or structure undergoing development at that time. The term sensitive period refers to those times when an event has its most dramatic influence on a specific psychological process. Examples of sensitive-period phenomena in human development are not as frequent as they are in the development of other animals. One example comes from the child's physical growth. The first three years of life roughly constitute the period of most rapid growth in stature; this is also the time when quality of diet exerts a maximal effect on eventual adult height. There is a coincidence of the phase during which stature is developing most rapidly and the time when stature is most vulnerable to environmental deficiency. Many psychologists regard the first few years of life as a sensitive period for the establishment of a trusting attachment to adults.

Contrasting conceptions of development

AREAS OF CONTROVERSY

The child as active or passive

Basic philosophical differences regarding the fundamental nature of children and their growth are found among psychologists. One dimension of controversy involves the degree to which a child is viewed as active or passive in response to his setting. In the English empiricist tradition (of which 17th-century philosopher John Locke is called the father) the child is seen as a passive product of environmental forces, his behaviour being molded by empirical, sensory experience. According to the continental tradition (of which the 18th-century moralist Jean-Jacques Rousseau was an exemplar) a child is regarded as actively manipulating and engaging his environment, trying to mold it to his skills, and attempting to understand it in terms of his current information. The psychological doctrines called behaviourism and psychoanalysis include views that regard the child as passive; while in the theories of Swiss psychologist Jean Piaget children are considered as active.

A second controversy pivots on the degree of continuity to be expected in psychological development. Does a child's behaviour progress continually, by gradual accretions, without sudden dramatic shifts? Or is the course of developmental progress steplike and segmented into stages? The term stage in psychological theory has no explanatory value but implies that the psychological function or behaviour is different during different stages. Most stage theories assume that each child must pass through successive stages in a fixed order. Some psychoanalytic theories, for example, indicate a fixed sequence of focal points for psychosexual developments, beginning with what is called the oral stage and progressing through anal, phallic, and genital stages. Piaget's theory states that each child passes from a sensorimotor stage (of relatively primitive coordination of perception and action) through a concrete operational stage in thinking to a final stage of formal operations (characterized by abstract thinking).

Most stage theories in psychology are descriptive rather than explanatory, for they lack an explicit set of rules that indicate the mechanisms by which a child progresses from one stage to the next. Behaviourism, one approach that does not make use of such stage concepts, is concerned with processes that produce changes in behaviour.

Idealized goals versus open-ended development

Another controversy turns on whether development proceeds toward an idealized goal, in contrast to being an open-ended program, much like biological evolution, with no special terminal state defining the goal of growth. Sigmund Freud and Piaget may be understood in such ideal terms, Freud suggesting that, at psychological maturity, the adult would have invested his libidinal energy (the energy of the sex drive) in a love object so that his anxiety over sexuality would be minimal. Piaget wrote that a capacity for logical reasoning is the goal of development; this suggests that when an adolescent can test hypotheses and solve problems in a systematic manner, he has reached the final stage of formal operations. Behaviourism and what is called modern cognitive theory, by contrast, do not conceptualize development as proceeding in a specific direction. A child is understood to acquire beliefs and responses, but no specific end point is defined as being fundamentally more mature or desirable than another; and no developmental goals are perceived as being independent of the culture in which the child is growing.

Most psychologists at least implicitly take a theoretical position on each of these dimensions. The discussion to follow is dominated by the view that each child is active rather than passive, that there are stages in development, but that there is no idealized goal independent of environmental needs and pressures.

In retrospect, theories of human development seem to have been neither logically rigorous nor to have attempted to account for both intellectual and motivational phenomena. Historically, major personality theories approaching the motivational development of the child have included psychoanalytic theory and social learning theory. Influential theorists of intellectual development have been Piaget and Heinz Werner.

PERSONALITY THEORIES

Psychoanalytic stages

There are critical differences between the conceptions of psychoanalytic and social learning theories with respect to the emergence of individual differences in children. Psychoanalytic theory, as first offered by Freud, assumes that the child passes through a series of stages hopefully culminating in an idealized adult state characterized by low anxiety, capacity for sexual satisfaction, effective defenses against stress, and an ability to love and work. It is a major tenet of psychoanalytic theory that the source of energy (that underlies such basic drives as sex) changes as development proceeds.

Thus, during the first postnatal year, excitation is held to focus on the mouth; infants indeed apparently obtain gratification through a reduction of excitation in the oral regions. Expectably, in psychoanalytic theory, this period is called the oral stage of development. During the second year the source of excitation is said to shift to the anal area, and new objects and behaviours are believed to bring gratification as the child enters this so-called anal stage. During the pre-school years, the source of excitation is expected to shift to the genitals, but not yet

in a fully heterosexual sense. Freud called this the phallic period of development. During the final and so-called genital stage of development, gratification is said to be sought in a love relationship with another.

A second important Freudian principle states that although id impulses are constantly directed toward gratification of the major drives of the organism, ego and superego functions act to set limits on this process. In Freud's language, as the child grows the reality principle gradually begins to control the pleasure principle; the child learns that the environment does not always permit immediate gratification. Development is primarily concerned with the emergence of so-called ego functions, those that are held to be responsible for the acquisition of behaviours, beliefs, and controls that channel the discharge of fundamental drives through thinking and via defense mechanisms.

Social learning theory

Social learning theory takes a more relativistic and arbitrary stance with respect to idealized notions of adult maturity. Social learning theory is concerned with identifying those mechanisms that can be offered to explain differences in behaviour, motives, and values among children. Its major principles stress the effects of external reward and punishment (administered by parents, teachers, and peers) on the child's tendency to adopt the behaviours and values of others. Also, this theory differs from psychoanalysis with respect to primary focus. Psychoanalytic theory is concerned with inferred internal processes surrounding motives, fears, and defenses, and only secondarily with behaviour. Social learning theory primarily is directed to the overt actions of the child, and less to inner mechanisms. As a result, while the principles and concepts of the theories differ, they often may be complementary rather than contradictory.

THEORIES OF INTELLECTUAL DEVELOPMENT

Piaget and Werner both postulated ideal states of intellectual development toward which the child is supposed to be progressing. Piaget's theory is the more detailed and rests on the fundamental notion that the child develops through stages until he arrives at a stage of thinking that resembles that of an adult from one of the western cultures. The five stages given by Piaget (and characterized below) are: (1) the sensorimotor stage from birth to 18 months; (2) the intuitive stage of primitive reasoning from 18 months to 3 years; (3) the preoperational stage from 3 to 7 years; (4) the concrete-operational stage from 7 to 12 years; and (5) the stage of formal operations that characterizes the adolescent and adult.

Piaget's stages

Most research and theoretical concerns have been focused on the last three of Piaget's stages. Prior to the concrete-operational stage the child seems unable to reason about several important classes of problems, even though he may have adequate language function. These problems include those that require the ability to represent a series of actions conceptually. Only when he becomes a concrete-operational child can he think representationally of a route he might take to a friend's house. The concrete-operational child shows what is called the ability to conserve; he states that liquids and solids do not change their amount or quantity merely because their external shape changes. In addition, concrete-operational children understand relational words such as darker, bigger, and larger, and cease to think about them in absolute terms. Such a child is able to reason simultaneously about the whole and about part of the whole. If a five-year-old child is shown eight yellow candies and four brown candies and is asked, "Are there more yellow candies or more candies?," he is likely to say, "More yellow candies," evidence that he cannot reason about a part and a whole simultaneously. But a child at Piaget's stage of concrete operations answers this question correctly. He also can organize objects along a quantified dimension; this is called serialization.

Piaget characterized the stage of formal operations, beginning at about age 12, as revealing the adolescent's ability to think about hypothetical problems that are not necessarily in accord with his experience. He shows willingness to think about possibilities. The adolescent is said to generate a number of possible solution hypotheses to a particular problem. Consider the following: A man was found in the back seat of a car that had hit a telephone pole. What happened? A child who conformed to Piaget's stage of concrete operations might think up one reasonable answer that satisfied him and state it. The older child typically would generate more than one of the possible ways this event might have happened, and then select the one that seemed the most logical. The adolescent's thinking tends to be self-consciously deductive and resembles that of a scientist. Piaget's stage of formal operations (typical of adolescents and adults) is basically an orientation toward problem solving, reflecting a tendency to isolate the elements of the problem and systematically to explore possible solution hypotheses, regardless of how they match what is called reality. Formal thinking is rational and systematic. Finally, the adolescent often broods about his own thinking and is self-conscious about his introspections.

Werner's theory

Werner's theory, which has had considerably less influence than that of Piaget, is that development proceeds from a state of relatively global, undifferentiated ability to a state of increasing articulation, differentiation, and integration. Although this idea makes intuitive sense, recent research indicates that the young infant is more differentiated than Werner would have led us to expect. Differentiation refers to a differentiation of functions. Initially all parts of the organism, according to Werner, operate in a similar way and only gradually does each assume different functions. Werner wrote that the child's thinking becomes more discrete, more analytic, and more flexible as he matures. Werner suggested that the same objective solution to a problem can be brought about by different functions: this is in contrast to Piaget, who posited only one principle of solution to certain problems. Werner emphasized how the child generates hypotheses and remains flexible; Piaget stressed the gradual attainment of a single hypothesis. Finally, Werner tended to be more concerned with language than was Piaget, and assigned language a more vital part in the child's subjective life.

Development before and at birth

INFLUENCE OF THE PRENATAL ENVIRONMENT

Periods of the ovum, embryo, and fetus

Development during the roughly 40 weeks prior to normal human birth is divisible into three phases. The first, which may be called the period of the ovum, lasts from fertilization until the time the fertilized egg is implanted in the wall of the uterus, a process that typically takes 10 to 14 days. A second phase, lasting from the second to the eighth week after conception, is called the period of the embryo, and is characterized by differentiation of the major organs that normally are present in the newborn. The last phase, from the eighth week to delivery, called the period of the fetus, is characterized by dramatic growth in the size of the organism.

Prenatal development is extremely rapid; by the 18th day the embryo already has taken some shape and has established a longitudinal axis. By the ninth week the embryo is about an inch long; face, mouth, eyes, and ears have begun to take on a well-defined form; arms, legs, hands, feet, and even fingers and toes have appeared. The sex organs, along with muscle and cartilage, have begun to form. The internal organs have a definite shape and assume some primitive function. The period of the fetus (from about the second month until birth) is characterized by increased growth of the organism and by gradual assumption of physical functions. At 32 weeks the normal fetus is capable of breathing, sucking, and swallowing, and by 36 weeks can show a response to light and sound waves.

Since prenatal environment is not the same for all fetuses, such variations can lead to differences in development. Factors related to these variations include the age of the mother, maternal diet, drugs, radiation, and disease. Even maternal emotional states may influence the infant: for example, extreme maternal anx-

iety during pregnancy seems to increase the chances of difficult labour and delivery.

THE BIRTH PROCESS

Hypoxia and prematurity

Major problems surrounding birth include hypoxia (partial lack of oxygen) and prematurity. Hypoxia in a newborn is most likely to injure nerve cells of the brain stem. Damage to these cells can produce such motor defects as those shown in paralysis of limbs, in tremor of face or fingers, or in an inability to use vocal muscles. The term cerebral palsy refers to a variety of motor defects associated with damage to the brain cells, usually as a result of oxygen deficiency during the birth process. Hypoxic infants commonly are unusually irritable during the first weeks of life and show more muscular tension and rigidity than do normal infants.

Prematurity also influences the child's development during the early years of postnatal life. According to one criterion, an infant born prior to 37 weeks of gestation is classified as premature. Since it is difficult to obtain information on the exact date of conception, physicians use the child's birth weight as an index. An infant with a birth weight under five and a half pounds is likely to be regarded as premature; with a birth weight under four pounds he is apt to be classified as severely premature. Worldwide data on the incidence of prematurity vary from 6 percent to 10 percent from country to country. About 7.6 percent of hospital births in the United States are categorized as premature. This figure is higher in underdeveloped countries. Premature infants remain relatively small in height and light in weight until they are about five years of age, and generally earn comparatively low scores on tests of cognitive and motor development during the first five years. However, after the child enters school, these differences tend to disappear. In the United States, most health problems associated with pregnancy and delivery are roughly four times more frequent among families with low incomes than among middle-class families.

The infant

Infancy may be defined as covering the first 18 months of life; alternatively, a child may be called an infant until he begins to speak meaningful language and is able to understand complex communications from adults. A universally accepted definition is not available. Statistical data to follow are based on body growth data provided by the World Health Organization. Average values, of course, vary from country to country and from one ethnic group to another.

BODY GROWTH

Full-term male babies tend to be slightly larger than females in all body dimensions, averaging about 20 inches tall and about 7½ pounds in weight. Body length increases by more than a third, and weight almost triples in the first year, so that by the age of 12 months the typical baby is about 28 inches high and weighs about 20 pounds. The infant's body proportions change rapidly, particularly during the second half of the first year. The total length of the head of a three-month fetus is about half the length of his body. At birth the head length is about one-quarter the body length and at adulthood only one-tenth. Bones of the body originate from soft tissue, which gradually hardens (becomes calcified or ossified). Most infant bones are incompletely ossified and, being relatively soft and pliable, are particularly susceptible to deformity. Tiny beginnings of teeth are present in the fetus at ten weeks and calcification has begun by the fifth prenatal month. The first tooth, generally the lower front, commonly erupts when the infant is about seven months of age; by one year the child has about six teeth. The newborn has all the muscles he will ever have, but they are small in relation to his size.

ENDOWMENT AT BIRTH

A newborn human being is a remarkably capable organism. From the moment he begins to breathe, he normally behaves as if he can see, hear, smell, and is sensitive to touch, painful stimulation, and changes in position. Perhaps taste is not yet functioning during the first day after birth. The newborn has a variety of reflexes, some of which serve his very survival and many of which are complex. An infant only two hours old typically will follow a moving light with his eyes, will show dilation and constriction of his pupils as darkness changes to light, will suck almost anything (*e.g.*, a finger) inserted into his mouth, and will turn his head in the same direction as the side on which his cheek or the corner of his mouth is touched. The newborn baby can cry, turn away, vomit, lift his chin from a front position, and grasp an object placed in his palm. The outer layers of the brain (cerebral cortex) tend to be incompletely functional in the newborn and only gradually assume a fuller role in the infant's behaviour during the first weeks of postnatal life. As the cortex gains function, it begins to inhibit the lower brain centres that mediate early reflexes (see Table).

Sensorimotor functions

MILESTONES DURING THE FIRST YEAR

The period of the newborn covers the first five to seven days during which the infant normally is recovering from the stresses of delivery. During the succeeding eight weeks, the child typically becomes markedly more attentive to his surroundings, and his vocalizations increase dramatically. Between eight and ten weeks after birth he apparently begins to appreciate visual depth, for he will look longer and smile more frequently at a three-dimensional face than at one shown in two dimensions. He also will seem to begin to lose interest in repetitions of the same event and to show frequent smiles to human faces. At about seven months of age the child may display signs of anxiety in response to unexpected events, especially when he is confronted by an unfamiliar adult. Between 12 and 18 months of age he commonly will show signs of walking and increased mobility, together with some understanding of language.

Visual development

The normal infant seems able to see light and colour at birth and shows signs of respectably good visual acuity. His pupils dilate and constrict to changes in brightness, he will efficiently follow a moving light with his eyes, and he shows some degree of visual accommodation (the ability to focus on near and far objects). By the age of six months his visual acuity appears comparable to that of normal adults. Evidence of duration of sustained attention in the young infant seems at first to depend on the degree of movement and contrast in the visual stimulus. Thus, stimuli with relatively sharp black and white contours and those that move will tend to elicit the most prolonged signs of attention. At about three months of age the child seems to attend most readily to stimuli that are moderately discrepant from those to which he has been accustomed. Attention appears to be attracted least when there is either a relatively small or a comparatively large discrepancy between an event and what one might infer the infant's expectation to be. During the latter part of the first year the child commonly behaves as if he has begun to think of ways in which different experiences may be classified. An unusual or discrepant event seems to provoke the child to activate hypotheses about its classification. The richer the set of hypotheses the child appears to have available to him, the longer and more sustained his attention seems to be. To summarize: the duration of the infant's visual attention is related to contrast and movement, discrepancy, and richness of his store of hypotheses for classifying sensory experiences.

Auditory perception

Normally capable of hearing at birth, a newborn is sensitive to the location of the sound source as well as to differences in the frequency of the sound wave. Low-frequency waves tend to increase motor behaviour if the child is normally alert, and to inhibit distress if the child is crying. High-frequency auditory stimuli tend to freeze motor behaviour or to elicit what seems to be a dramatic alerting reaction. Sound stimuli of a few seconds' duration have a minimal effect, while those that last 5 to 15 seconds are most apt to maximally affect the infant's

Reflexes of the Newborn

effective stimulus	reflex
Tap upper lips sharply	lips protrude
Tap bridge of nose	eyes close tightly
Bright light suddenly shown to eyes	closure of eyelids
Clap hands about 18 inches from infant's head	closure of eyelids
Touch cornea with light piece of cotton	eyes close
With baby held on back turn face slowly to right side	jaw and right arm on side of face extend out; the left arm flexes
Extend forearms at elbow	arms flex briskly
Put fingers into infant's hand and press his palms	infant's fingers flex and enclose finger
Press thumbs against the ball of infant's feet	toes flex
Scratch sole of foot starting from toes towards the heels	big toe bends upward and small toes spread
Prick soles of feet with pin	infant's knee and foot flex
Tickle area at corner of mouth	head turns toward side of stimulation
Put index finger into mouth	sucks
Hold infant in air, stomach down	infant attempts to lift head and extends legs

activity level. Newborns also are likely to become quieter when auditory stimulation is rhythmic, and intermittent patterns have a more quieting effect than does a steady sound. Contrast, discrepancy, and richness of hypotheses also seem likely to modify attention in the auditory mode.

The newborn infant can make some discrimination among odours. The typical two-week-old shows positive sucking responses to sweet substances and makes negative responses to bitter stimuli. Highly sensitive to changes in spatial position of the body, the newborn baby who is jarred or who falls from a sitting position will make energetic postural adjustments.

Many motor responses are maturational and develop with no special training. Given the opportunity to use limbs and body, every normal child will begin to creep, stand, walk, sit up, and grab objects. Most babies are able to sit for a minute or so with support at three months of age, and by eight months can sit without support. Most begin creeping at about 40 weeks and are crawling adequately by 50 weeks of age. The typical infant is standing and walking by 64 weeks of age. Maturational changes in the neuromuscular systems largely seem to determine the time at which the child will sit, stand, and walk.

Vocalization and speech

Babbling and vocal sounds are almost universal during the infancy period in normal children, and no relationship has been established between the amount of babbling during the first six months and the amount or quality of speech when the child is two years old. Vocalization in the young infant often accompanies motor activity and usually occurs when the child appears to be excited by something he sees or hears. During the second half of the first year the child often will grow quiet while apparently listening with interest and may begin to show signs of excitement by babbling when the sound ends.

Use of meaningful words differs from simple babbling in that speech primarily helps obtain goals, rather than simply reflecting excitement. The first of two basic sounds includes all those related to crying; this is present at birth. A second category, described as cooing, emerges at about eight weeks, and includes sounds that eventually become part of meaningful speech. Environmental experiences ordinarily do not begin to influence vocalization seriously before two months of age, and differences in vocalization related to social class do not emerge until about 18 to 24 months of age. During the first two months of postnatal life the vocalizations of deaf children born to deaf-mute parents are indistinguishable from those of infants born to normal parents. Environmental effects on the variety and frequency of the infant's sounds become more evident after roughly eight weeks of age. The typical child utters his first meaningful word at about one year and his first short sentences between 18 and 27 months.

Piaget's views of infant development

Piaget stated that an infant has a set of so-called internal schemes that control his external coordinations. He identified the first phase of mental development as the sensorimotor stage; this, in turn, was held to differentiate into six subphases covering the first year and a half of life. The first four of these are said to be achieved during the first year; during the first (reported to last one month) innate reflexes become more efficient. Piaget labeled the second subphase as primary circular reactions, characterized by simple acts that are repeated for their own sake; *e.g.*, sucking, opening and closing the fists, and fingering a blanket. He attributed no intent on conscious purpose to this activity. The last four of these six initial subphases are interpreted as being more intentional. During the third (called secondary circular reactions), said to last from the fourth to sixth months, the child is described as repeating responses that produce interesting effects; for example, he may kick his legs to produce a swinging motion in a toy. In the fourth subphase (lasting from the seventh to the tenth months and called coordination of secondary reactions) the child is expected to begin to solve simple problems, building on actions he has mastered previously. Thus, he may knock down a pillow to obtain a toy hidden behind it. During the fifth (tertiary circular reactions, 11th to 18th months) he is described as carrying on trial-and-error experimentation. At this time he may change his responses toward the same object or try out new responses to obtain the same goal.

The final stage of infancy, held to be achieved by the age of 18 months, is characterized as involving the invention of new ways to solve problems. By now the child is said to have developed a primitive form of representation, defined as a kind of imagery that is used to solve problems. According to Piaget, the child gradually learns during the first year that an object exists permanently, even though it is not always in view. Prior to six months of age he is not expected to behave as if objects that are out of sight continue to exist. For example, he may grab for objects he can see, but probably will not reach for objects outside his immediate visual field. However, from about nine months on he will be expected to reach for objects hidden from view if he has watched them being hidden. During the first half of the second year he even may search for objects that he has not watched being hidden, indicating an inference about their location. In the final stage he is described as searching for objects that he has not actually seen being hidden. Show such a child a toy placed in a box, put both under a cover, then remove the box and its contents and the child will search under the cover as though he inferred the location of the toy.

Social learning in the first year

The infant is born showing very few specific emotional or motivational responses to other people. But his experiences with adults during the first year presumably lay the foundation for his relationship with and later behaviour toward them. From the moment the baby is born he exhibits responses. Some are spontaneous, others are reactions to what seem to be his needs for nourishment, alleviation of pain, and warmth. The infant spontaneously scans the environment, vocalizes, sucks, smiles, cries, and thrashes. During the third and fourth month he is apt to begin to cling to objects, manipulating his own fingers, toys, and parts of his mother or nurse. Human infants, like those of other species, have many unlearned, built-in reactions that they direct at whatever interacts with them. As a result, the infant under typical conditions becomes attached to the biological mother. Each vertebrate species is provided with a special set of responses that offspring can exhibit at birth (or hatching) or very soon afterwards. Each makes these responses to the first appropriate object that the environment supplies. The objects that elicit these responses are likely to become objects of attachment for the young mammal or bird. The mother is usually the most effective stimulus for these reactions. In addition, the mother commonly provides food, alleviates discomfort, and is a source of pleasant stimulation.

In humans, important consequences of these early ex-

periences of attachment to and contentment with the mother or nurse include a generalization of responses from her to other people, and an articulated imaginary representation of her face, form, and voice. It is inferred from the infant's smile in response to a human face that the child has developed a subjective representation (or schema) of the person who takes care of him. Most four-month-old infants smile at a motionless representation of a human face. Neither movement nor voice is necessary, although facial movement and voice increase the probability that the smile will occur.

Relation between attachment and anxiety

Increasing familiarity with a specific adult seems to increase the chance that the infant will become afraid when he encounters a less familiar face. This is called stranger anxiety and even may be elicited by the father's face at the appropriate point in the infant's development. A common source of concern among fathers who have grown accustomed to smiles from the infant, this abrupt change should be welcomed and understood. It indicates that the child has developed sufficiently to distinguish one face from another.

The emergence of stranger anxiety at about seven months indicates that by this age most infants have developed a good representation of the mother's face. The infant is alerted by the relatively strange person but is too immature to do much more than cry. Five months later he usually is able to ask who the stranger is, run to his mother, or run away. Or he will have been exposed to strangers so often that he will have learned not to cry.

Another kind of infant distress called separation anxiety is elicited when the mother leaves the infant. The more closely attached the infant is to her the more likely he is to cry upon her departure. When the mother leaves such a child in an unusual or discrepant location, he is alerted and cries. As he develops he experiences more frequent separations from his mother in unusual locations and gradually seems to acquire the ability to interpret her actions, or to reassure himself of her return, and his signs of anxiety diminish. Children who do not have a particular adult devoted to their care often do not become attached to one adult, are less likely to show separation and stranger anxiety, and are less socially responsive—less likely to smile, vocalize, laugh, or approach adults. Similar behaviour has been observed in monkeys reared in isolation and in children raised in relatively impersonal institutional surroundings.

Intellectual development during the preschool and early school years

Intelligence is a murky word that resists easy definition. Many psychologists define the term as the quality of the subjective processes activated when one tries to solve a problem, to attain a desired goal, or to make sense of his experience. The development of adaptive thinking or behaviour is said to require a set of so-called intellectual units and a set of routines by which they are organized. Units or elements of thinking include what commonly are called schemata, images, symbols, concepts, and rules; these may be called cognitive units. The routines include perceiving, remembering, generating hypotheses, evaluating, and deducing; these often are called cognitive processes. To the extent that cognitive units are processed appropriately, one may be said to be thinking intelligently.

Among the many philosophical attitudes toward thinking, one view assumes that there is within each person a kind of psychological executive that continually monitors the cognitive processes, much as an architect supervises the construction of a house. An alternative view is that monitoring is a function of the units and processes themselves, analogous to the behaviour of chemicals in a beaker. The chemical units called hydrochloric acid and sodium hydroxide interact, facilitated by the process of ionization, yielding the products salt and water. No external executive or separate force watches over the beaker to guarantee the proper reaction. The reaction is understood to be inherent in the nature of the chemicals. North American psychologists generally have been friendly to the view that regards thinking as a more or less mechanistic activity. European theorists such as Piaget hold that higher-order "mental structures" organize and keep thinking adaptive, coordinated, and efficient. This view has strong intuitive appeal; although a ten-year-old child retains considerable information, he quickly selects the correct facts and applies the precise routine when asked how a bird and a fly are alike.

Units in thinking

Those who posit such cognitive units as schemata, images, symbols, concepts, and rules hasten to add that such units are not to be viewed as physical things with size or shape. Instead, they are regarded as properties or activities of the thinker, in much the same way as temperature is considered to be a property of molecules in motion. Each cognitive unit provides the child with the ability to make sense of experience, to remember past events, and to generate solutions to problems.

Schemata, probably the earliest of these to develop, are defined as abstract representations of specific events. These representations are not photographic copies or visual images but are more like schematic blueprints that emphasize the arrangement of a set of salient elements. The salient elements supply the schema with distinctiveness and differentiate it from those of similar events.

Images are reconstructions from schemata that preserve the spatial and temporal detail of the event.

Symbols represent the next level of abstraction from experience and are arbitrary names for things and qualities. Common examples of symbols are the names for letters, numbers, or objects. While a schema or image represents a specific experience, such as a sight or a sound, and preserves the relations in that experience, a symbol is an arbitrary representation of an event. The child who can name as the letter *A* a specific arrangement of three lines, and can point to *A* when asked, is said to possess the symbol for *A*. Children beyond six years of age commonly have schemata, images, and symbols for letters of the alphabet. Symbols are used in the development of other cognitive units called concepts.

A concept may be thought of as a special kind of symbol that represents a set of attributes common to a group of symbols or images. The concept defines a common attribute or meaning from a diverse array of experiences, while a symbol stands for a particular class of events. Consider a drawing of a cross. An eight-month-old is apt to represent this object as a schema or image. The three-year-old names it a cross and thereby represents it as a symbol. An adolescent may regard it as a cross of Christianity, thinks of it in relation to religion and the church, and therefore represents it as a concept. A concept is a person's way of extracting common traits from a variety of experiences. For additional discussion and other definitions see CONCEPT FORMATION.

A rule may be understood to state a relation among or between the dimensions of two or more concepts. A rule implies that the presence of one dimension is correlated with the presence of another. Consider the rule that candy is sweet. This states a relation between two concepts (candy and sweet) and describes the dimensions they share. A second meaning may be assigned to the notion of rule. A rule may be understood to state a routine or function imposed on a set of concepts to produce a new concept. Multiplication is a rule imposed on concepts called numbers to yield a new concept.

Informal rules refer to imperfect relations involving two or more dimensions. The statement that candy is sweet is an informal rule, for occasionally one finds sour candy. Most of a child's beliefs are informal rules; *e.g.*, snakes are dangerous, water is cool, men are tall.

What are designated as formal rules state relations that are asserted always to be true and specifiable. The rules of mathematics and physics are familiar examples. The formal rule $6 \times 11 = 66$ states a fixed relation between the concepts 6 and 11.

COGNITIVE PROCESSES

Cognitive processes include two very general types: undirected and directed. Undirected thinking refers to free

association, dreaming, or reverie, and includes the apparently free flow of thinking of which one is aware as he walks home or stares out the window. Directed thinking refers to activities that proceed when one tries to solve a problem that has been given to him or that he has set for himself. A normal preschool child seems to know that there may be a solution to a problem and apparently realizes when he has reached an answer. Problem-solving processes typically involve the sequence: perceiving, remembering, generating concepts, evaluating, deducing, and (under some circumstances) reporting the answer to someone.

Consider the general changes that normally occur from 3 through 12 years of age. The richness of the child's supply of symbols, concepts, and rules increases; and his tendency to rely on images in problem solving decreases. The child shows growing concern with whether his concepts and rules agree with those of other children; he may appear more apprehensive about making mistakes and his memory function may show dramatic improvement.

Perceiving

Perceiving refers to the process by which an individual selects, organizes, and initially interprets sensory impressions. To perceive is to extract information from the environment. These perceptual interpretations change with experience. As a child develops there is evidence of greater precision in recognizing similarities and differences; the child makes more differentiations among stimuli in his environment. Acquisition of language helps the child to learn by distinguishing specific features of new events. When the child gives objects distinctive names, it is easier for him to perceive them as separate and different from each other. Once he learns the words red and pink he is more likely to notice differences among objects of these two colours than if he had learned neither word or only one of them. Each event is defined by its salient elements: eyes are salient for a face; legs help define an animal; a break in the continuity of a line is a salient element for defining a letter of the alphabet.

In contrast to older children, a young child is apt to exhibit difficulty in attending to more than one event at a time. If he seems to try to listen or watch many things at once, he often appears to become confused. An adult is likely to divide his attention among competing stimuli with better efficiency. For example, adults can attend alternately to two conversations with greater efficiency than can children, who are less familiar with word sequences and who have a harder time making sensible constructions from partial sentences (see also ATTENTION).

Language

A remarkable development during the second year is the typical child's enormous progress in acquiring language. During this period he begins to relate many symbols with objects and to use words meaningfully. He understands verbal signals better and his speech becomes more complex. Available data indicate that a typical one-year-old understands only three words. By 15 months the average is 19 words and by 21 months, 118 words. At two years of age, this comprehension vocabulary contains an average of 272 words. Children in the 1970s tend to have larger vocabularies than did their counterparts of the 1940s, and they use longer clauses and sentences. In the U.S., three-year-olds in 1930 used about 3.4 words per sentence; about 25 years later children of the same age averaged 4.0 words. It is thought possible that the widespread use of radio and television, a drop in migration from other countries and in the proportion of bilingual children, and the availability of nursery schools contributed to the change.

Early forms of grammar

Ordinary verbal language has a vocabulary and grammatical rules for arranging words into meaningful sentences. Even a very young child may have a kind of grammar, although these rules for arranging words are very basic or elementary. The first words spoken at the end of the first year generally are simple labels for persons, objects, or simple actions; *e.g.*, milk, go, mommy. A child also begins to make requests or to describe the environment: "Want out," "Daddy away." Single words may stand for entire sentences. Thus, the word shoe may mean, "Take off my shoe." Eat may signify, "Can I eat now?"

At about 20 months most children begin putting a few words together to make simple statements that are abbreviated versions of adult sentences. The sentence, "That's the ball" becomes "that ball"; the sentence, "Where is the ball?" becomes "where ball?" So-called function words (as, and, at, the, his) are almost completely absent at this stage. The young child seems to have his own special, simplified set of grammatical rules for forming sentences. It consists of two classes of words: pivot words and variable words. A series of sentences might be comprised of the verb see followed by a variety of words, to produce "see hat," "see sock," "see horsey," "see boy." The pivot word is see and the noun objects are the variable words. The variable class is large and open, and consists of almost all of the child's vocabulary, except for some of the pivots. After using these pivot-variable constructions for about a half-year most children begin to utter two-word sentences in which the variable word occupies both utterance positions. Examples would be "man car," for "The man is in the car"; "car bridge," for "The car is under the bridge." These are not pivotal constructions but illustrate primitive sentence forms in which both components can be expanded.

It is not yet clear how the child acquires grammar so early and how he comes to be able to generate so many new sentences he has never heard before. His grammar does not seem to be acquired by direct training, for parents do not consistently reward grammatically correct sentences nor do they punish incorrect usage each time. His speech does not appear to be the result of direct imitation alone for he creates new, grammatically correct sentences that he does not seem to have encountered before.

Memory functions

Memory involves both the storage of traces of past experience and the retrieval of that stored information at a later time.

It is useful to distinguish between what are called short-term and long-term memory processes. Short-term memory may be defined as referring to traces available for a maximum of 30 seconds immediately after stimulation, but typically for a much shorter period. Unless such factors as heightened attention or repetition serve to transfer information from short-term traces to long-term storage, the ability to remember a specific experience may be lost forever.

One usually judges how much a child can remember by asking him to recall what he saw or heard, or by asking him to recognize selected stimuli. When he is asked to recall an event he is required to retrieve the necessary information and is given no hints. When he is asked to recognize an event he is presented with an array of information from which he is to select only that which seems familiar. Children and adults tend to perform much better when they have to recognize than when they have to recall. This difference is most dramatic in young children. Thus, a young child may only recall 3 of 50 pictures he has seen, but typically can recognize almost all of them. The young child's ability to recall is less efficient than that of the adolescent, perhaps because he has a less adequate set of cognitive units to label incoming information.

Images, symbols, and concepts seem to function metaphorically as the glue of thinking; simply naming an event tends to increase the length of time over which it is remembered. The very act of naming requires increased attention to the event and tends to facilitate rehearsal. A young child is less likely to have learned to use the trick of rehearsal. He is not likely to repeat events to himself spontaneously in an effort to aid in their storage. Remembering is a most elusive yet a most central process in thinking, and is easily disturbed when one is anxious, distracted, or fatigued.

The process of explaining

Perceiving and remembering typically are the first two processes in a problem-solving sequence. Next, one com-

monly generates possible solutions to the problem. Important steps in the effort to explain include: searching the available set of concepts and rules for precursors of an event that is not immediately understood; and checking each precursor for consistency with presently established rules or causal beliefs about the object or event. If the explanation that results contradicts an old rule, it is likely to be rejected. Finally, if the explanation is judged to match experience and does not seem to contradict old rules, it is apt to be accepted.

Critical factors in generating acceptable solution hypotheses include a set of available cognitive units appropriate to the event and the absence of a firmly held rule that contradicts the new hypothesis. Some children fail in this effort by simply being unable to think of new explanations; some encounter strongly held rules that contradict the new explanation; others seem never to feel sure their hypotheses are correct. An important change, usually observed to occur between 7 and 11 years of age, is an increasing ability to generate possible explanations of hypothetical events that violate conclusions drawn from the child's experience.

An obstacle to the adaptive use of concepts and rules by young children is their tendency to regard a concept as absolute rather than relative. When a four-year-old first learns the concept dark, he is likely to regard it as descriptive of an absolute class of colours (namely, black and other dark hues). The phrase dark yellow may make no sense to him if dark does not yet mean relative darkness. It often is difficult to teach a five-year-old to grasp both the absolute and relative aspects of a number; this involves persuading him that a concept can have more than one meaning.

Children seem to vary in the degree to which they pause to evaluate the quality of their thinking and the accuracy of their hypotheses and conclusions. Some children behave as if they accept and report the first hypothesis they produce, acting upon it with only the barest consideration of its accuracy. Others seemingly devote a long period of time to considering their ideas, and evidently censor many of their hypotheses. This dimension of difference can be seen as early as two years of age. As children mature they are likely to devote longer and longer periods of time to considering the validity of their thinking. It may be inferred that they have become increasingly afraid of making mistakes.

Deductive thinking

The complementary processes of generating hypotheses and of deductive thinking have been called the essence of problem solving. To deduce is to apply a rule to solve a problem, and a most important limiting factor in the quality of deductive activity is the child's store of rules, which increases with age. Of major theoretical interest is the question that centres on whether there are basic changes in the use of rules over the first dozen years. According to one view, the child merely learns more rules each day, storing them for future use; but it also is asserted that there is no rule too difficult for a child of any age to acquire. An alternative assumption is that some rules are too complex for young children to understand and apply. This viewpoint (vocally advocated by Piaget) assumes that there are stages in the development of reasoning.

MOTIVATIONAL DEVELOPMENT

Most children evidently think about goals they would like to attain. These ideational activities are called motivational, and the evidence is that they undergo regular changes with development.

The concept of a motive may be distinguished from that of a biological drive. A drive is definable as a state of deprivation or discomfort associated with an imbalance in the physiology of an organism. Hunger, thirst, cold, and pain are common biological drives. Each drive involves the activation of certain structures in the central nervous system, often with internal sensory experiences that are unpleasant.

As distinguished from drives, motivational activities include wishing, striving, hoping; they represent what the individual desires to experience or to possess. Such activities are cognitive and have no necessary relation to overt action. A child may have wished for many years to be able to fly but may not have taken observable steps toward attaining that goal. Motivational states can be potential or active, much as the energy of movement has a potential and a kinetic state. Most of the time a child is not aware of being motivated toward a given goal; at these times the motivational state is potential. The motivational state becomes active when some arousing event, called an incentive, stimulates the child.

Children differ in the ease with which an incentive can evoke such striving or wishing. Some are easily provoked to hope that their mothers will hug them; others rarely respond this way. Some react to being slightly frustrated by becoming motivated in a hostile manner; others require more substantial disappointment before they show hostility. The easier it is for a particular incentive to elicit a motivational response, the stronger that response tendency is said to be. Such tendencies seem to form a hierarchy; that is, at any given stage of development some incentives are more potent than others and tend to dominate the child's ideational activities. At two years of age the child's interest in a close relationship with the parent tends to dominate most other motivational preoccupations, while at 15 years of age this normally is much weaker as the adolescent strives for a satisfying heterosexual relationship (see also MOTIVATION).

Efforts toward gratification

Since subjective activities have no necessary link to overt action, factors that prompt a child overtly to attempt to gratify himself should be understood. The child first must have learned responses that are effectively gratifying. A six-year-old who strongly desires to control his older brother may not have learned how to effect this control and, therefore, will not show signs of this motivational tendency in his everyday behaviour. The child also should expect that his overt action will be successful in obtaining the goal. A five-year-old who has lived in one foster home after another, none of which responded to his requests for affection and help, is likely to stop displaying such signals and become sullen and withdrawn. The less the child expects to be gratified the less likely he will be to direct his behaviour toward attaining the goal involved.

Feelings of anxiety over displaying the goal-related behaviour also modify the link between motivational function and direct action. Many inner strivings in a two-year-old are likely to lead to behaviour aimed at gratification, ostensibly because such a young child has not learned to delay or inhibit these attempts. A normal seven-year-old, by contrast, has learned that he must delay satisfying many of his desires and may inhibit certain actions because he feels guilty or fears parental rejection.

A final factor in goal-directed behaviour is the immediate situation in which the child becomes aroused. A child who feels strong hostility toward his mother is less likely to reveal these feelings when he is sitting in a classroom than he would be if he were at home. In general, any member of the social environment reacts to and classifies each child primarily on the basis of his observable actions, rather than on his inferred motives; it is through his overt behaviour that others initially come to know him. People generally care more about whether a child hits other children than they do about his hostile thinking. More concern is shown about his observable sexual behaviour than about the quality of his sexual fantasies.

Efforts to reduce uncertainty

A basic motivational characteristic of all but the earliest phases of human development normally is the desire to keep feelings of uncertainty at a minimum. Particular classes of events have the common effect of making one seriously uncertain about the future. The feeling of uncertainty may begin when one encounters an unfamiliar situation or person, when he doubts the consequences of an action, or is unsure about being loved or rejected, about passing or failing a test, or about being fired or promoted. Thus, a child may become anxious when he is in an unfamiliar situation or when he is uncertain about the likelihood of various events and is not sure what response

to make or as yet has no response appropriate to the unfamiliar situation.

Feelings of anxiety are most likely to occur when a person does not know how to behave in an uncertain situation. Since the sources of uncertainty change as the child develops, the dominant aspects in the motivational hierarchy will change. During the first postnatal year, uncertainty about the mother's presence and a need for signs of her approval are dominant; thus, the child is said to have a strong motive for parental presence and affection. During the second year uncertainty about parental punishment becomes salient and the child is motivated to avoid parental disapproval. During the preschool years the child is apt to be uncertain about his competence and sex role. When he enters school, the major uncertainties centre on mastering school subjects and on gaining the acceptance of the other children (his peers). A primary source of uncertainty in early adolescence is the tendency to doubt one's ability to maintain a satisfactory heterosexual relationship; hence, sexual motives become dominant.

Important changes in salient motives during the first 15 years are related to changes in major sources of uncertainty. Uncertainty may prompt a person to wish for another to intervene and to reduce the unpleasant feeling of anxiety by offering affection, reassurance, advice, or money; such a person is said to have the motive for affiliation. Anxiety does not always lead to the motive for affiliation, and each can occur separately. Thus, some children may wish to be alone when anxious rather than to seek out others.

Anger and hostility

Some classes of incentive have the effect of thwarting or of threatening the child's attempts to attain desired goals. In a general sense, these conditions frustrate the child and he normally becomes emotionally (affectively) aroused; in this case the affective reaction is called anger. Other events may threaten the beliefs or values of the child and indicate that his standards are incorrect or morally reprehensible. This experience also may lead to the emotional arousal called anger. Anger has to be differentiated from the motivational state of hostility. To be hostile is to wish to inflict pain, distress, or anxiety on another person or on a substitute for (surrogate of) that person. The person to whom the hostility is directed is inferred to be the one the child identifies as the thwarting agent or the one who threatens the child's standards. Although anger and hostility often seem to occur together, they can occur separately. A five-year-old who cannot open a screen door to enter a house may become angry and may stamp his feet, but may not experience hostile wishes toward anyone at that moment.

The situations that are likely to arouse anger change as the child develops. During the first two years of life, restriction of the child's tendency to explore the environment is a major factor in eliciting signs of anger. The child is apt to become angry at the parents' efforts to force him to go to bed, to stay within fenced areas, and to keep away from attractive but fragile objects. During the preschool years, when the child is apt to be given more freedom, deprivation and postponement of specific goals typically elicit signs of anger. During the school years, when even more freedom may be granted to the child, he usually seems angered by events that threaten his standards regarding sex role, passivity, autonomy, or his acceptance by others. These values may be threatened when someone implies that the child has an undesirable set of characteristics, or when he encounters a salient person who holds values different from those he holds.

Although it usually is assumed that all aggressive behaviour is derived from hostility, this does not seem to be the case. Some behaviours that manifestly hurt people are not in the service of anger or hostility. Aggression, therefore, is defined as an action that inflicts pain, anxiety, or distress on another, and is in the service of a hostile motive or of the emotion of anger. The advantage of this definition is that it allows us to reject actions that unintentionally hurt others, and to include acts that, on the surface, appear to be culturally valued but that intentionally cause distress to another person.

Sexual motives

Most children are motivated to see and explore their own genitals and to manipulate them for pleasant sensations. The child who wants to view and explore the genitals of others may do so to satisfy a general curiosity about discrepant events. Discrepancy in physical appearance between the child's own anatomy and that of the opposite sex seems to elicit a strong motive to understand the difference. Thus, the young child's sexual motives appear, in part, to be tied to pleasant sensations derived from genital manipulation and, in part, from an effort to understand differences in sexual anatomy. They are not yet associated with the romantic interactions that adults label sexual love. As the child approaches adolescence the nature of the motive apparently changes, as does the nature of his behaviour.

Effectance motivation

Effectance motivation may be defined as a desire to master a new skill or to perfect an old one. A major basis for the effectance motive during the first three postnatal years seems to be the child's desire to match his actions and their products to an internal standard (or schema). As early as four months of age the infant is likely to smile when he studies a face that almost matches the schema he has acquired for a face. The dynamic process of matching to a schema is a pleasant experience. Similarly, the two-year-old who builds a tower may smile as the tower gradually becomes taller. The child appears to have an internal representation for the tower he wishes to build and, as his product comes closer to matching that idea, he behaves as if he experiences pleasure. The effectance motive also stems from what seems to be a desire to predict and control events. When a one-year-old discovers that hitting the keys of a toy xylophone produces varied sounds, he is likely to repeat this action many times. When he appears to be able to predict the outcome of his behaviour every time, he is apt to become bored.

The involvement generated by a challenging task seems to entail uncertainty of prediction. The challenge of many school-age games tends to be maximized when the child seems unsure of what will happen. The more varied and unpredictable the outcome, the more likely the child will be to perceive a particular game as interesting. Another possible basis for effectance motivation involves a desire for self-definition. According to this interpretation, the child wishes to know his attributes and needs to believe that he is competent in at least one skill. A particular skill is understood to have salience for the child if it is distinctive and is not possessed by everyone.

In sum, the preschool and early school years are witness to the growth of several motivational and behavioural systems. By the time the child is ready for school he normally is motivated toward affiliation, genital stimulation, instrumental help, affection, effectance, and hostility. Several factors determine whether any of these motivational tendencies will lead to observable attempts at gratification. The child's possession of behaviours that can gratify a motive, the degree of his anxiety associated with gratification, and his expectancy of gratification all influence his goal-directed behaviour. Finally, a behaviour that to the observer seems intended to gratify a specific motive may be aimed at gratifying a quite different motive. It is difficult to make absolute statements about the meaning of particular goal-oriented behaviours or to predict what behaviour will result from an inferred motive. Each child apparently possesses his private code, and more must be learned about the way he thinks if the aims of his actions are to be interpreted correctly.

MORAL DEVELOPMENT

One's morality embraces his beliefs about the appropriateness or goodness of what he does, thinks, or feels. These beliefs reflect what are called standards. The middle childhood years, from ages 5 to 11, represent a sensitive period during which moral standards typically develop at a rapid rate. According to Piaget, from the ages of 5 to 12 the child's concepts of right and wrong pass

from rigid and inflexible notions learned from his parents to a sense of equity and moral judgment that takes into account the specific situation in which a presumed moral violation has occurred. Five-year-olds commonly view lying as bad, regardless of the situation or the circumstances; with age, the child tends to become more flexible and to consider exceptions to this strict rule. His judgments become less absolute or authoritarian, and depend more on the needs and desires of members of the group within which he functions. He becomes a moral relativist in contrast to being a moral realist.

The U.S. psychologist Lawrence Kohlberg hypothesized that the child's development of moral standards passes through developmental stages grouped into three moral levels. At the early level the child is characterized as being guided by an orientation toward punishment and obedience; standards are held to be based strictly on what will avoid punishment or bring pleasure. At the intermediate level, the child is said to view moral standards as a way of maintaining the approval of others; the child is portrayed as accepting the precepts of authority. Moral standards at this level are held to rest on a positive evaluation of authority, rather than on fear of punishment. At the third level Kohlberg describes the child as viewing moral principles as obligations to others and to his individual conscience. Thus, the bases for justifying moral standards are understood to pass from avoidance of punishment, to avoidance of adult disapproval, to avoidance of self-recrimination. This progression is described as being accompanied by a progression from fear of being hurt, to anxiety over rejection, to guilt and self-derogation.

Defenses against feelings of guilt

Violations of moral standards often can produce symptoms of psychological distress. In Western culture, violation of standards surrounding hostility, aggression, dependence, sexuality, or incompetence are most likely to produce signs of anxiety and guilt. As a result, a child often erects defenses against the sources of anxiety or guilt.

Repression, which Freud called the most basic defense, is said to keep the person from being aware of ideas that make him anxious. In denial (another defense mechanism), the child insists that a situation that makes him feel anxious is not true, to the point that he believes in the accuracy of his denial. In the mechanism called projection, unacceptable feelings or impulses are acknowledged but the child attributes them to other sources. In projecting, he ascribes his own undesirable impulses or actions to other persons. The child may place the blame for injury suffered by a child he was chasing by saying, "He made me chase him. If he didn't make me chase him, I wouldn't have bumped into him." In what is called displacement, the child directs his actions toward a substitute target. The child who wishes to hit his father may displace this response to a younger sibling (brother or sister) or to a pet cat. In exhibiting signs of the Freudian mechanism of rationalization, the child offers a socially acceptable reason for his unacceptable behaviour, attitudes, or thinking. In behavioural withdrawal, a defense mechanism frequently employed by preschool children, the child simply avoids threatening situations or people.

Evidence of anxiety and fear

The importance of feelings of anxiety in the lives of school-age children is attested by the observation that almost a third of children's drawings (in response to the request that they draw the most important event in their lives), illustrate fear- or anxiety-provoking experiences. Although many childhood fears are related to direct experience with dangerous events (*e.g.*, a speeding car), or are a result of parental warnings (*e.g.*, to stay away from fire), many symbolic fears are also noted. Young children are only moderately fearful of immediate and possible dangers such as being hit by a car, but seem intensely afraid of remote or highly improbable events such as runaway lions or ghosts. When children are very anxious, they are prone to develop phobias, obsessions, compulsions, or such psychophysiological symptoms as asthma or ulcerative colitis (an intestinal disorder). Some children with initially intractable asthma experience rapid relief as soon as they are taken from their mothers to live in residential treatment centres (see PSYCHIATRIC TREATMENT, CONCEPTS OF; PSYCHONEUROSES).

THE ROLE OF THE FAMILY

Effect of parents

A child is influenced by a vast array of stimuli, including his parents, siblings, peers, teachers, and the public media. Parents and siblings (brothers and sisters) tend to have their major effect during the child's first six years; peers (fellow children) and teachers usually are most influential during the child's preadolescent period; the public media (*e.g.*, television) tend to prevail during the person's early adolescence. Parents can mold the behaviour of a child through reward, punishment, instruction, and by serving as models for the child. If he values parental acceptance and praise, these adult responses can be used to establish and maintain selected behaviour in the child. Punishment (whether it be spanking, threat of rejection, or deprivation) also produces changes in behaviour. The child's feelings of anxiety generated by anticipated punishment lead him to inhibit undesired responses and to substitute more acceptable acts. Parents also influence behaviour through instruction and by serving as models, facilitating imitation and identification.

Important dimensions of parental behaviour include what are called affection-hostility and control-permissiveness. The affectionate parent is described as accepting, approving, and understanding. The hostile parent is characterized as excessively punitive, critical, and as holding a negative view of the child. The dimension of control-permissiveness is defined by the degree to which the parent restricts the child and enforces demands on the child's sexuality, hostility, manners, neatness, orderliness, care of household goods, obedience, and so on. A child may be living with parents who are: affectionate and permissive; affectionate and controlling; hostile and permissive; or hostile and controlling. Some psychologists state that each of these classes of parent tends to produce a different kind of personality in the child. The affectionate-permissive parent is said to make the child friendly, dominant, and possessed of high self-esteem. The affectionate-controlling parent is held to favour the development of a more dependent, less friendly, less creative, more hostile child. The hostile-controlling parent is said to produce a child with strong tendencies toward feelings of guilt and one who is disposed to shyness and social withdrawal. It is stated that the hostile-permissive parent is apt to have an aggressive child who has poor control over his behaviour.

Effect of siblings

The psychological influence of siblings is likely to be felt most keenly by a child who is between three and nine years of age. Rivalry with a younger brother or sister represents a major threat to the firstborn's relationship to the mother. The secondborn child, on the other hand, is apt to perceive the older sibling as omnipotent, competent, and entitled to special privileges. Critical factors associated with being a firstborn child include the child's tendency to orient to his parents for values and to identify with parental models; he commonly exhibits anxiety over the loss of parental affection when a sibling arrives. He also tends to think of himself in a privileged role, because parents are apt to accord firstborn children special advantages. Firstborn children generally hold more standards that conform most closely to those values that parents promote than do later-born children. The firstborn child is more likely to identify with adult authority figures outside the home and is more apt to trust adults.

The later-born child tends to have a greater sense of inadequacy when he compares himself with his older sibling, especially if their age difference is less than four years and when both are of the same sex. However, the later-born child does not tend to excuse his relative incompetence as the result of being younger than his sibling. He commonly seems to assume that he should be as talented as his older brother or sister. A disadvantage in the role of the firstborn child is that he tends to experience more anxiety over loss of parental affection; he also may become more dependent in time of stress. The first-

born child is especially vulnerable to guilt over his hostility toward the second-born sibling, probably because he is not readily able to justify his resentment to the younger child.

Adolescence

PHYSICAL CHANGES IN ADOLESCENCE

Puberty refers to the first stage of adolescence during which the reproductive system matures. The bodily changes at this time result, in part, from increased amounts of gonadotropic hormone released by the pituitary gland. This hormone stimulates the sex glands (gonads: testes or ovaries) to produce sex hormones and induces the growth of mature sperm or ova. This typically is accompanied by a growth spurt, an accelerated rate of increase in height and weight. In boys, the growth spurt may begin as early as 10 years of age or as late as 13½, but usually begins at 13 years of age, with a peak between 13½ and 14 years. In girls, the growth spurt may begin as early as 8 years of age or as late as 11½, but usually begins at about age 11, two years before the boy, and reaches a peak at 11½ to 12 years of age. The impression that girls mature earlier stems primarily from the observation that girls tend to attain adult height and weight about two years earlier than do boys. Almost every part of the body undergoes some change during adolescence and many tissues increase in size. The heart beats faster and the reproductive organs increase in size. Bones grow in size, changing proportion and shape. At the same time cartilage begins to calcify, and the bones become harder, denser, and more brittle (see BONE).

The average age at which girls in temperate climates begin to menstruate (reach menarche) is about 13 years; about 97 percent of these girls reach menarche between 11 and 15 years of age. In warmer lands menstruation tends to begin earlier, and in colder countries, later. Sexual maturity in boys tends to be reached between 13 and 14 years of age, if the male's development of pigmented pubic hair is considered equivalent to the menarche in girls (see MENSTRUATION).

Major developmental tasks facing the adolescent include the establishment of personal independence, sexual adjustment, peer relations, and preparation for a vocation. Establishing independence from parents is not a simple matter. In the face of cultural pressures for independence from peers, parents, and teachers, the child tends to strive toward an adult state of independence. However, the child commonly is not assigned a clear functional role in many Western societies. Since parents sometimes put pressure on him to remain passive and obedient, the child often is placed in conflict.

Many societies have seen a significant increase during recent years in the expression of sexual interests and related behaviour during adolescence, and these activities continue to be considerably greater among boys than among girls. Recent changes in British, European, and North American cultures indicate more openness and candor about sexual motives and behaviour than were observable in the 1950s. Thus, 52 percent of one sample of males attending school said that sexual intercourse was acceptable for a male before marriage if he was engaged; a lower proportion of the females in the sample gave a similar answer for their own sex. Within a group composed of 20 percent of the males, each said that sexual intercourse was acceptable to him before marriage, even if he did not have any particular affectionate feelings toward his partner, while only 10 percent of the females seemed to accept such values for themselves (see SEXUAL BEHAVIOUR).

THE INFLUENCE OF PEERS

Relationships with peers serve many important functions, providing an opportunity to learn how to interact with others, how to control social behaviour, and promoting the development of relevant skills and interests. The peer group also may act as a therapeutic agent; *e.g.*, if it helps the adolescent understand and allay his feelings of anxiety over conflicts. The peer group can promote discussion of sexuality, may give license for aggressive play, and permit the adolescent to vent hostility. The child is likely to feel safe with his peers in discussing and expressing officially prohibited actions and ways of thinking. Friends sometimes function as mutual psychotherapists and may help to buffer feelings of conflict.

The peer group also can be influential when it communicates its evaluation of the child, positive or negative, by giving or withholding group acceptance. The child tends to trust his own group's evaluation more than he does that of adults; he is likely to feel that adults either are overly accepting of his undesirable qualities or at other times are overcritical. The adolescent is apt to require evaluation by peers since he ordinarily lacks absolute standards to help him judge how bright, handsome, wise, or likable he is. The adolescent is prone to use his immediate peer group as the reference for the qualities he wishes to assess in himself. His evaluation is partially dependent on the size of the reference group, and on how he perceives its quality relative to the larger community in which he lives. The larger the peer group, the less likely the adolescent is to decide that he is exceptionally talented or capable of leadership.

The peer group provides role-playing models with which he may identify. He is likely to respect those peers who command positions of power. The young adolescent tends to adopt the leader's values, imitating his behaviour in the hope of eventually commanding the same powers and competencies. Moreover, peers teach the child a role to play in the group. Unlike the family, the adolescent peer group composed of members of the same sex has no obvious differentiation by age and sex. Instead, differentiation in the group rests on assigning different roles to various members. Essential roles include those of the leader, the leader's advisor, and the scapegoat. Other roles may be those of the intellectual, the clown, the rebel, the athlete, or the ladies' man. Adolescents cast in these roles commonly take them seriously and each tends to behave in accordance with how he perceives the role given to him. Apparently the child adopts the behaviour appropriate to his role in an effort to gain acceptance and distinction for playing the role properly.

Antisocial behaviour

An adolescent who drops out of the school enterprise is most likely to come from an emotionally troubled, socially isolated, lower-class home that typically is located in a segregated ethnic or economic urban ghetto. He is likely to have friends of whom his parents do not approve, who share his aversion to academic activities, and who already have dropped out of school or who soon will drop out themselves. He usually has had academic difficulty (especially in reading) for many years, even though he may be of average or higher intelligence. He is likely to have been held back to repeat one or more grades and to have begun to feel frustrated. Troubled by feelings of inadequacy, anger toward authority, and lacking a clear sense of his own identity, he tends to be a creature of the moment; impulsive and low in frustration tolerance.

Often these children become juvenile delinquents; in typical Western societies boys are about four times more likely to commit offenses than are girls. The most frequent delinquent acts recorded for boys involve such aggressive behaviours as burglary, malicious mischief, larceny, and automobile theft. For girls, the commonest offenses include running away from home and promiscuous sexual behaviour. The incidence of cases of juvenile delinquency in the U.S. has increased substantially since 1948; estimates in the 1970s suggest that at least 12 percent of all children (this includes 22 percent of all boys) turn up in juvenile-court records before the end of their adolescence. While delinquents often are found in lower-class neighbourhoods, specific relationships also have been reported that link personality, intelligence, and delinquency among children from more privileged homes.

SUMMARY

The major theme of this discussion has been that the components of human psychological development (in-

cluding motives, moral standards, expectancies, emotions, sources of anxiety, cognitive structures, and actions) are reorganized as development proceeds. This discussion has been largely descriptive since developmental psychology lacks a tight net of interlocking theoretical propositions that reliably permit satisfying explanations. Developmental psychologists still are unable to achieve agreement in explaining how a child passes from one presumed developmental stage to another or how environmental experiences facilitate or obstruct this passage. However, a few related conclusions seem appropriate at this time: psychological organization changes as the child grows; also, the exact relationship of a child's overt behaviour to his subjective state is always to be regarded as uncertain. The major task of developmental psychology is to map the child's observable behaviour on the internal, dynamic processes that are inferred as adapting him to his environment and as helping him make sense of his experiences.

BIBLIOGRAPHY. P.H. MUSSEN, J.J. CONGER, and J. KAGAN, *Child Development and Personality,* 3rd ed. (1969), a basic textbook in child development covering both cognitive and personality growth; HEINZ WERNER, *Einführung in die entwicklungspsychologie* (1926; Eng. trans., *Comparative Psychology of Mental Development,* 1940, rev. ed. 1948), the theoretical point of view of one of psychology's most original theorists; JEAN PIAGET, *La naissance de l'intelligence chez l'enfant,* 2nd ed. (1948; Eng. trans., *Origins of Intelligence in Children,* 1952), the fundamentals of Piaget's theory of intelligence, and with B. INHELDER, *La genèse des structures logiques élémentaires* (1959; Eng. trans., *The Early Growth of Logic in the Child,* 1969), descriptions of experiments on how a child passes through various stages of operational thought; M.L. and L.N.W. HOFFMAN (eds.), *Review of Child Development Research,* vol. 1 (1964), chapters on contemporary issues in child development, ranging from sexual identity through cognitive development; A.L. BALDWIN, *Theories of Child Development* (1967), a summary of the basic position of major theories of human development, including Piaget, social learning theory, Werner, and Freud; J. KAGAN and H.A. MOSS, *Birth to Maturity* (1962), research monograph summarizing the results from a longitudinal study of adults who were followed from infancy through early adulthood, with comments upon stability and change in personality growth; P.H. MUSSEN (ed.), *Manual of Child Psychology,* 3rd ed. (1969), basic handbook of child psychology with fundamental chapters on infancy, cognition, and psychopathology; E.H. WATSON and G.H. LOWREY, *Growth and Development of Children,* 5th ed. (1967), a textbook emphasizing aspects of physical, including physiological, growth in children.

(J.K.)

Human Behaviour, Innate Factors in

Many people believe that, although most bodily characteristics are inherited, psychological or behavioural traits are not. Nevertheless, one's hereditary endowment (genotype) indeed predisposes him to become bright or dull, or to be prone to show specific personality traits fully as much as to be tall or short, or to have dark or light hair. The biochemical and physiological processes that are involved in linking the potentials of genotype to the actualization called the phenotype are complex in all such cases.

Properties of behaviour

Behaviour is somewhat more difficult to study genetically than are such traits as eye colour. Since most psychological characteristics are continuous (*i.e.*, each such trait varies by degrees), more complex extensions of the simple methods of classical genetics must be used in their investigation. Further, behaviour is fluid or environment sensitive in that its phenotypic values may fluctuate widely under environmental influence. Thus one's exposure to a learning task tends to change his performance. Likewise the environment in which a child is reared may markedly affect his measured intelligence and personality.

Since behaviour is complex, even its special definitions tend to be imprecise. Most psychologists, for example, consider intelligence to be a loose term that covers many distinctive kinds of ability. These difficulties are not unique to the study of behaviour. Many bodily traits are also continuous, environment sensitive, and complex; height and weight are examples.

NATURE AND NURTURE

Phenotypic characteristics result from the interaction of nature (genotype) and nurture (environment). A main goal of geneticists is to establish the degree to which each genetic or environmental component contributes to phenotypic variation.

The term genetic commonly is used interchangeably with such words as innate, inherited, or heritable. Likewise, environmental corresponds to acquired. The designation congenital refers merely to a trait or syndrome that is present at birth, for whatever reason: genetic, prenatal, paranatal, or whatever. However, these equivalences are loose; thus, to say that a trait is genetically based is to indicate that its expression in the individual depends on genes. The heritability of a trait, on the other hand, technically refers to the degree to which the expression of that character in a sample of individuals is controlled by genotype. Heritability concerns populations rather than individuals, and heritability estimates for a trait may vary widely from one population to another.

The proposition that characteristics acquired through exposure to environment can be transmitted to succeeding generations was proposed by Jean Lamarck (1744–1829). Widely debated in the 19th century, championed even by Charles Darwin, forms of Lamarckism continued (despite contradictory evidence) to be espoused, notably in the U.S.S.R. where so-called Michurin-Lysenko genetics once had official government approval. Reports offered in support of this position strained the credulity of most non-Soviet geneticists. Thus, a report of the transformation (by environmental treatment) of one species of wheat (*Triticum durum*) into another (*Triticum vulgare*) was compared by U.S. geneticist Theodosius Dobzhansky to having a domestic cat give birth to a lion.

To the extent that behaviour is genetically based, it should follow the same kind of rules that apply to bodily traits. Consequently, such changes as those induced in temperament or intelligence through learning are not expected to be incorporated in the chromosomes to be passed on to offspring. Indeed, most studies to test the inheritance of acquired behavioural traits in laboratory animals have given negative results. The few that seemed to support the Lamarckian view either had obvious flaws or could not be verified by other investigators.

This is not to say that environmental factors cannot change the genotype. Thus, actual chromosomal aberrations may arise during embryological development; mongolism (Down's syndrome) is such a case. The origin of this form of retardation in intelligence remained a mystery until 1959, when it was reported that a sample of mongoloids showed a complement of 47 chromosomes rather than the usual 46, the extra one belonging to a chromosome group labelled number 21. These groups normally contain two chromosomes, and in this anomaly (designated as trisomy 21) there are three. The probability is that half of the children of female mongoloids will also show the syndrome. Assuming that trisomy arises environmentally, then this acquired character is genetically transmitted; similarly, heritable changes may be induced by environmental radiation (*e.g.*, X-rays) through chromosomal mutation. (See also BIRTH DEFECTS AND CONGENITAL DISORDERS.)

Chromosomes and mental retardation

Likewise, when fruit-fly (*Drosophila*) pupae are exposed to severe heat, about half of them develop defective wing structure (absence of cross-veins). Further treatment and breeding of the abnormal insects eventually produces descendants that show the defect even without the heat treatment. A character that originally appears only under environmental manipulation now emerges spontaneously. This does not, however, give clear support to the simple notion that all acquired characters can be inherited. Presumably a line of *Drosophila* has been bred from ancestors whose threshold for the expression of a genetically based character is extremely low, the tendency to express the trait clearly existing in the original genotypes.

GENETICALLY MEDIATED INDIVIDUAL DIFFERENCES

Bodily characteristics. The pattern of human inheritance for some bodily characteristics is not as simple as was once supposed. Eye colour, for example, long has been thought to depend on a dominant gene B (for brown), and a recessive gene b (for blue), so that BB and Bb produce brown, and only bb yields blue. (Genes that have such alternative forms as B and b are called alleles.) While this model often applies, sometimes two blue-eyed parents may provide genetic material that produces a brown-eyed child; apparently small, ordinarily undetectable brown spots in the iris of the eye were present in the parents. Blond hair seems to be recessive to brown and black, and red hair apparently depends on a different pair of recessive genes. There are so many intermediate variants, that the presence of other modifying (epistatic) genes must be considered.

Genes and eye colour

The inheritance of skin pigmentation probably also depends on a number of gene pairs. This continuous (quantitative) trait can be measured with special methods (*e.g.*, spectrophotometry) or by matching skin hue against a standard scale. Work using these techniques indicates at least five allelic gene pairs to be involved, largely producing intermediate degrees of pigmentation.

Human bodily form showed significant heritable components: stature; height-weight proportionality; some fat, bone, and muscle indexes; physique (somatotype); and bodily signs of bisexuality (masculinity-femininity).

Studies in biochemical genetics have generated information about a number of specific diseases; for example, so-called hemoglobinopathies and inborn errors of metabolism. A notable hemoglobin disorder is sickle cell anemia, called the prototype of the molecular diseases. Involving a tendency of red blood cells to assume an abnormal sickle shape, its clinical manifestations include anemia, jaundice, and obstruction of blood vessels. These symptoms have been found to reflect the substitution of an amino acid (valine) for a glutamic acid in a complex amino-acid chain found in the hemoglobin molecule. Sickle cell anemia is an example of what is called a balanced polymorphism. Transmitted as a dominant by a single pair of alleles (SS), this double-dominant, killing disorder should tend to die out, being subject to strong negative selection; but the recessive condition (ss) is especially susceptible to malaria. Hence ss is subject to negative selection, and the genotype Ss has an advantage in resisting malaria. (Indeed, the sickling trait is largely confined to parts of Africa where malaria is a major disease.) Thus, both genes are maintained in the population, an example of balanced polymorphism.

Inborn errors of metabolism

It has been suggested that a continuing high frequency of psychiatric disturbance, in spite of strong negative selection, may also represent a case of balanced polymorphism.

More than a hundred inborn errors of metabolism have been identified. Common examples include phenylketonuria, tyrosinosis, galactosemia, albinism, and alkaptonuria. Besides the physiological abnormalities produced by the enzyme defects involved, such behavioural changes as retardation in intelligence may also result. Phenylketonuria (PKU), for example, in which an amino acid (phenylalanine) found in some protein foods is improperly metabolized and accumulates in the body, may result in IQ (Intelligence Quotient) levels of 20 or less. In some (but not all) PKU sufferers, more intelligent behaviour appears when a diet low in phenylalanine is instituted. (Perhaps the degree of responsiveness to such treatment is itself genetically controlled.) PKU generally is thought to depend on a pair of recessive genes, and heterozygotes (people who carry just one of the pair) are detectable through a phenylalanine tolerance test. When they ingest enough phenylalanine, they show about twice the normal level of this amino acid in their blood.

Hemophilia, a disorder in which blood fails to clot normally, is carried by a recessive gene that is sex-linked (located on a sex chromosome). Diabetes mellitus, a deficiency of insulin, does not seem to fit such a simple genetic model; it shows different degrees of severity and exhibits variable age of onset. There may, however, be a complex of syndromes involved; perhaps a juvenile variant and an adult form of diabetes.

Genetic transmission of epilepsy also is incompletely understood, although there is clear evidence for a strong heritable component. For example, abnormal brain-wave patterns frequently are observed among close relatives of epileptics, being especially pronounced in mothers of those who show abnormal electrical activity associated with specific parts of the brain (focal psychomotor discharge). Evidence based on one family pedigree suggests that at least one form of catalepsy (in which bodily motion is arrested during seizures) may be inherited, possibly through the action of a single dominant gene.

Psychological characteristics. *Sensory functions.* From a genetic point of view, one of the better understood sensory functions is the tendency of different people to experience a chemical called phenylthiocarbamide (PTC) as bitter or as tasteless. The substance is one of a number of compounds containing the carbamide group. In most populations, taste thresholds (*e.g.*, for hydrochloric acid) are distributed according to chance (*i.e.*, normally). This is not the case, however, for PTC, the proportion of non-tasters ranging from about 20 percent among Africans to about 35 percent in Europeans. Pedigree data suggest that PTC "taste-blindness" depends on one pair of recessive alleles, although other studies have not always given support to this simple notion. The ability to taste PTC as bitter appears to be highly specific, in that the substance can be recognized by a "taster" only when it is dissolved in his own saliva. The trait is neither related to overall taste acuity nor to the general ability to taste non-carbamides (*e.g.*, quinine sulfate) as bitter. Since the carbamides exist only as man-made synthetics, it is difficult to guess what adaptive function the taster gene might serve in human evolution. Suspicion of its association with diabetes has not been confirmed. Taste thresholds for PTC and for such compounds as quinine, however, seem to be related to food preferences, smoking, performance on intelligence tests, and proneness to peptic ulcers; age and sex were found not to be directly implicated.

Taste and smell

Little work has been done on the genetics of differences in the sense of smell (olfactory function). Scattered pedigree reports may implicate hereditary factors in selective inability to detect particular odours (specific anosmia).

Deafness can have both hereditary and environmental bases; the origins of congenital deafness (present at birth) or deafness appearing early in childhood may be assigned as shown in Table 1. These figures are based on data from large populations in several countries.

Normal auditory function also depends heavily on genotype. There is evidence of heritability for loudness discrimination, rhythm, and auditory acuity; tone deafness in human populations probably depends on a single chromosomal locus. This is not to say that these abilities may not be improved with practice, however.

Genetic factors are important in many forms of blindness and in defective ability to converge the eyes or to focus (accommodate) for near and far objects. Some common forms of colour blindness are clearly hereditary. Most varieties of red-green blindness usually are regarded as classic examples of sex-linked transmission, but the genetic situation probably is more complex. Thus there appear to be at least four distinct phenotypes (protanopia,

Vision

Table 1: The Causes of Early Childhood and Congenital Deafness

cause	percent
Assignable disease of trauma	25
Multiple autosomal recessive loci at complete penetrance	40
Autosomal dominant loci with an average of 85% penetrance	15
Nonsegregating factors (either unrecognized diseases or effects of polygenes)	19
Sex-linked or sex-influenced loci	<2

Source: K.S. Brown and C.S. Chung.

protanomaly, deuteranopia, and deuteranomaly) distinguishable by noting how a subject matches a yellow with mixtures of red and green. Pedigree data suggest that protanopia and protanomaly are mediated by sets of multiple allelic genes (independent of those producing deuteranopia and deuteranomaly), each set producing a specific chemical defect as in the inborn errors of metabolism.

Personality and intelligence often are said to be reflected by performance on selected visual tasks. Whatever the validity of such statements, many sets of data (from twins and other family members) suggest that heritable components mediate susceptibility to selected visual illusions, eidetic imagery ("photographic" memory), spatial orientation and afterimage characteristics, and differences in ability to detect flicker in a source of light (flicker-fusion frequency). Since most studies have involved small samples of people and have done little to control for prior experience, they call for cautious interpretation.

Sensorimotor functions and special abilities

There is a paucity of data bearing on genetic factors in motor skill. Twin and family studies suggest some heritability for such abilities as tapping speed, athletic performance, motor-reaction time, card sorting, hand steadiness, mirror-image drawing, and manual dexterity. In some cases, however, results have varied considerably between initial and final performance. Many other kinds of performance tend to change with training, and it should not be assumed that heritability estimates will show a constant value over practice trials. Such estimates also tend to be higher when based on the performance of the right hand (preferred by most people) and could simply reflect superior, more reliable performance.

Based on studies of the behaviour of twins, it has been reported that outstanding athletic ability (as in Olympic competition) strongly depends on heredity.

Musical ability shows a strong tendency to run in families. The pedigree of composer Johann Sebastian Bach shows many relatives with substantial musical talent and represents a notable, but not unique illustration.

Data gathered at the Juilliard School of Music in New York City show a clear degree of musical talent in 50–75 percent of the parents of opera singers, instrumental virtuosos, and music students. It would appear that significant musical ability emerges very early in life; in general, such aptitudes improve steadily among the gifted, even without practice, and in a variety of environmental settings. Such data point to maturation in the gradual unfolding of genotype as a major factor.

Many other studies implicate heredity in musical ability, whether measured as a global score or by means of different subtests. For example, musical quotients based on tests of musical intelligence have shown identical (monozygotic: MZ) twins to be much more alike than are fraternal (dizygotic: DZ) twins.

Genetic models put forward to account for the transmission of musical talent include: single-gene action with incomplete dominance; a single or a few recessive genes; several major genes with multiple alleles. None of these theories seemed to account completely for the data available in the 1970s.

Intelligence. In the 19th century, Sir Francis Galton analyzed a large number of pedigrees of distinguished statesmen, scientists, military commanders, literary men, artists, divines, and other groups. He concluded that high intelligence (as manifested in such activities) strongly depends on heredity; his later work with twins pointed to the same conclusion. Although noting that his method failed to control for them, Galton accorded environmental influences little importance. Subsequent studies strongly tend to endorse Galton's conclusion that intelligence runs in families. Except in such extreme cases as infantile malnutrition, environment indeed seems to play a minor role; for example, adopted children resemble their biological parents in measured intelligence more than they do their foster parents. This tendency increases with the age of the child and applies even for widely disparate socioeconomic environments. Also, children living in such homogeneous environments as orphanages show as great a range in intelligence as that found among comparable children living in their own homes, each home representing a different environment.

Heritability of intelligence

Up to about puberty, raw scores on intelligence tests tend to increase regularly with age (as does any biologically rooted character), and are largely independent of all but the most pronounced deviations in environment. For example, monozygotic twins reared apart (though less alike than those reared together) still tend to be more similar in intelligence than are dizygotic twins raised in the same home. This has been found to hold no matter how early in life such twins are separated. Within wide limits, long exposure to a common environment does not seem to make people more similar in intelligence. Scores of studies clearly emphasize the role of heredity in intelligence among human beings and other animals.

Concerning the extent of genetic determination in human intelligence, most investigations have yielded heritability estimates between 70–80 percent. Since such values are relative to the population studied and to the method of estimation, some disagreement should be expected. It seems most unlikely, however, that genotype contributes less than 50 percent of the variability and it is conceivable that the figure is closer to 80 percent.

Major, independent studies concur in showing that specific aspects of intelligent behaviour (verbal ability, spatial intelligence, and word fluency) have strong genetic dependence. In one especially sophisticated study, numerical ability also was found to be significantly heritable.

Thus both general intelligence (*e.g.*, IQ) and its specific component abilities seem to be heritable, some of the latter being more dependent on genotype than are others. The precise genetic mode for the transmission of intelligence was incompletely understood in the 1970s. Some theorists favoured the notion of multiple genes; one suggested a major gene pair (Nn), dominant for normal intelligence (IQ range 90–110), with five additional minor pairs of genes acting as modifiers that could push IQ up or down in the absence of the dominant gene N. A mathematically advanced attempt to fit a genetic model to sets of twin and family IQ data suggested that about a hundred chromosomal loci were involved, and indicated dominance for higher intelligence; it yielded no evidence for interaction between genotype and environment.

Retardation

Cases of severely low intelligence tend to involve gross neurological, biochemical, or chromosomal impairment; single dominant or recessive genes generally are implicated. When there is only mild retardation with few or no physical defects, the individual may be regarded simply as representing the low end of the IQ range.

At least a quarter of the known inborn errors of metabolism have been found to produce severe defects in intelligence. Some examples are Tay-Sachs disease (idiocy associated with blindness), Gaucher's disease (anemia of the spleen), Wilson's disease (a defect in copper metabolism), galactosemia (a disturbance in sugar metabolism), and phenylketonuria (PKU). The majority of phenylketonurics shows IQs below 50, although their group average is about 50. Adaptive and language functions are most adversely affected, while motor and personal-social skills show the least retardation. A marked deficit in attention span has been demonstrated among PKU children. All or most of these aberrations can be alleviated in at least some PKU groups that are maintained on a low-phenylalanine diet.

A major breakthrough in understanding the etiology of many gross forms of retardation was the identification of chromosomal anomalies. Mongolism (Down's syndrome) has been noted above as being associated with autosomal trisomy. Risk increases from less than 1:1,500 for mothers under 30 years of age to 1:60 for mothers age 45 and over.

In another type of chromosome abnormality part of one of the short arms in the 4–5 chromosome group is missing. The syndrome thus produced has been called *cri du chat* because an affected infant's vocalization may be compared to the mewing of a cat. Well-known aber-

Cri du chat, Turner's and Klinefelter's syndromes

rations involving the sex chromosomes include Turner's syndrome (gonadal dysgenesis), associated with absence of one of the X chromosomes among females; and Klinefelter's syndrome (seminal duct dysgenesis) in which males have two or more X chromosomes. Most of these conditions produce subnormal intelligence and other psychological changes. In this connection, there may be a relation between the presence of surplus chromosomes and antisocial behaviour. Men who show extra Y chromosomes have been reported to be unusually tall and aggressive; but it is likely that any connection between this chromosome anomaly and criminal behaviour is very indirect.

Thus it has been suggested that any chromosomal aberration produces a variety of physiological symptoms, including cerebral changes akin to minimal brain damage. This, in turn, may result in changes in personality that dispose to many forms of abnormal behaviour. A link has been suggested between chromosomal anomalies and some of the major psychoses (*e.g.*, schizophrenia), but available evidence has failed to support the hypothesis.

Attitudes and beliefs. Although family tendencies to hold specific beliefs and attitudes have been demonstrated, it seems most unlikely that these could be under genetic control. Inherited personality dispositions (*e.g.*, aggressiveness) might, however, predispose an individual to, for example, authoritarian attitudes. Indeed, a study of MZ and DZ twins demonstrated significant hereditary contributions to scores on several scales of vocational interest (*e.g.*, those for physicist, mathematician, osteopath, dentist). Any connection between genetic makeup and such vocational interests is, however, apt to be rather remote. It often is argued that MZ twin pairs share more similar environments than do DZ twins (a general criticism of the twin method).

Personality. Studies on the heritability of basic personality characteristics (largely among twins) have yielded variable results. Inconsistencies appear among reports regarding tests of dominance, psychoneurotic tendency, and need for achievement. A number of studies suggest sex differences in heritability estimates for personality traits, however. And there is some consensus that the dimension of introversion-extroversion is strongly heritable; perhaps to the same extent as intelligence. One analysis of twin data showed a heritability of 67 percent for this trait. It was concluded that the introvert genotype is more modifiable by environmental factors than is that for the extrovert. This seems to converge with the notion that introverts are more readily trainable through conditioning than are extroverts. Other dimensions that show evidence of heritability (perhaps of evolutionary significance) are aggression, anxiety, attention to detail, social attachment, activity, and emotionality. Individual differences in some of these appear early in life; infants, for example, tend to be reliably indentifiable as "cuddlers" and "noncuddlers." It has been suggested that cuddlers may become extroverts as adults, and that noncuddlers are prone to become introverts.

Psychiatric disorders

While many gross deviations such as metabolic errors or chromosomal abnormalities are inherited, the evidence is less clear for so-called psychiatric disorders. The psychotic disturbance called schizophrenia has received most attention and Table 2 summarizes genetic studies from several European countries and the U.S. While there is general support for schizophrenic inheritance, the average values in the table show little of the variation typically found. Different reports of rates at which pairs of MZ twins both are affected (concordance rates) range from as low as zero to higher than 80 percent. In one study, schizophrenic symptoms were found in five of 47 foster-home children of schizophrenic mothers; no symptoms were observed among a matched group of similarly reared children of normal mothers. Few authorities discount environmental influences in schizophrenia.

Likely factors contributing to such variability include sex differences (female twins being more concordant), severity and chronicity of the disturbance, age, criteria for identifying DZ and MZ twins and for diagnosing schizophrenia itself, and gene frequency and incidence (or penetrance) in particular populations. Some theorize the action of one or a few pairs of major recessive genes; others favour a dominant mode of transmission; a few hold that a model based on the cumulative effects of a number of genes (a so-called polygenic, threshold model) most suitably accommodates the data. If the diagnostic label schizophrenia covers more than one disorder, all of these mechanisms may be operating.

Table 2: Studies on the Incidence of Schizophrenics in Relatives of Varying Degree of Kinship

relationship to schizophrenic	expectation of schizophrenia (percent)	number of investigators	number of countries
Unrelated (gen. pop.)	0.86	19	6
Parents	5.07	14	8
Siblings	8.53	12	7
Children	12.31	6	2
Uncles and aunts	2.01	4	4
Nephews and nieces	2.24	5	4
First cousins	2.91	4	4
Twin pairs	**Both twins affected**		
DZ OS	5.6	6	5
DZ SS	12.0	9	7
MZ	57.7	10	8

Population levels of schizophrenia hold at about one percent in spite of apparently strong selection against it. Schizophrenics are relatively unlikely to marry, have fewer children than the average, and these have a comparatively high rate of early mortality. It has been suggested that negative selection of any schizophrenic genotype may be compensated by a positive selection for carriers (as in sickle cell anemia). Indeed, unaffected sisters of schizophrenic females tend to have more children than average; and unaffected offspring of schizophrenic mothers included more conspicuously successful adults than were observed among a control group.

Among other forms of psychological disorder, significant contributions of genotype have been claimed for manic-depressive psychosis, involutional depression, senile psychosis, homosexuality, alcoholism, childhood schizophrenia and autism, hysteria, obsessional and anxiety neuroses, psychopathy, cyclothymia, schizoid personality, bedwetting (enuresis), and reading difficulties (dyslexia). Few of the studies adequately controlled for environment by using separated twins, and the data often do not permit specific genetic hypotheses. In the case of manic-depressive psychosis, several investigators agreed that a dominant gene (or genes) is involved but differed about whether it is sex-linked or not. Dominant inheritance also has been suggested for enuresis and for anxiety neurosis; a polygenic model has been offered for hysteria. It has been hypothesized that genotype does not so much predispose to neurotic breakdown as it does to determine the form the disorder will take if it does occur.

While considerable evidence points to genetic factors in normal and abnormal personality, exact modes of transmission are poorly understood; environmental influences should be considered in themselves and as interacting with particular genotypes.

GENETICALLY MEDIATED GROUP DIFFERENCES

Passionate opinions are held on whether human groups in relative cultural, economic, or geographic isolation differ genetically. The explosive political implications have led some to argue that the problem should not even be examined by scientists. Most authorities suggest, however, that emotional attitudes and conflict are generated more by ignorance than by empirical evidence.

Early attempts at classifying human groups stressed such variations as skin colour, facial features, or bodily measurements. Later classifications emphasized geographic isolation and criteria of gene frequencies for traits; for example, local genetic groups (microraces) in Wales as defined by blood groups.

Efforts to identify so-called racial groups

People with some common feature (*e.g.*, dark skin) may not differ from other groups in gene frequencies for other traits and may not be geographically isolated, and the same geographic region may have genetically isolated populations. Combinations of these variables can identify groups for specific research purposes. Thus, groups distinguished on the basis of skin colour and geographic location (*e.g.*, U.S. white, Charleston Negro, U.S. non-Charleston Negro, and West African Negro) also can be compared genetically. Estimates of gene flow (exchange of genes through interbreeding) can be depicted in terms of a combination of bodily traits, blood-group factors, and so-called biological distances. Such a multivariate scheme tends to avoid the oversimplification of cruder systems of classification. Since multivariate classification had not been used widely up to the 1970s, the bulk of earlier information about alleged racial differences in behaviour seemed primitive at best. Nevertheless, for genetically distinct groups between which there is little or no gene flow, inherited differences should be anticipated both for adaptive and nonadaptive forms of behaviour.

Most available data concern broad comparisons among rather ill-defined groups, particularly of European and African origin. Intergroup differences have been claimed for such psychological functions as reaction time, visual and auditory acuity, colour-blindness, pain sensitivity, and discrimination of tone, weight, touch, taste, and smell; however, problems of design and control make these findings somewhat ambiguous. Differences in complex traits also have appeared (*e.g.*, susceptibility to visual illusions, and performance on tests of intelligence and personality) and for these the problems of interpretation are even more formidable. Average intelligence test scores clearly have been higher among so-called whites than among populations identified as black, tending to divert attention from millions of superior blacks and inferior whites. Since test performance is related strongly to expenditure for schools, it is controversial whether genotypes or environmental factors contribute most to such differences. Until environmental variation can be controlled more fully even such simplistic abstractions as group averages will be difficult to interpret.

Even greater caution needs to be applied to score differences on personality tests for such crudely defined so-called racial groups. Their significance is badly clouded by such variables as language skill, motivation, test-taking attitudes, and rapport with the tester.

Traits related to social class

The possibility of genetic bases for similarly reliable differences among socioeconomic groups has been considered. Social constraints indeed may tend to isolate such groups biologically, each potentially constituting a kind of microrace. If environmental factors alone were operating, those reared in impoverished settings invariably would tend to develop low ability, and, if a rich environment or special education were provided, could rise to a high socioeconomic level. Yet genetic factors seem to modify this view; for example, the IQ's of children within any social class tend to show regression (to approach the average IQ of all classes combined). Thus, many individuals are born into environments geared to occupations above or below their abilities. A British psychologist, Sir Cyril Burt, has estimated from empirical data that about 45 percent of adults are "incorrectly" placed in society. He suggested that a substantial proportion of the population would need to move from one class to another if any distribution of ability across occupational strata were to be preserved. On the other hand, selective mating on the basis of class-related traits (*e.g.*, IQ, educational background) should tend to reduce variability and to diminish "incorrect" placement within social classes. In any case, the evidence fails to show that people who hold socioeconomic power constitute a genetically elite class.

STATISTICAL CONSIDERATIONS

The main goals in studying the genetics of human behavioural traits are (1) to assess the relative influence of hereditary as compared with environmental contributions to the variance of the traits; (2) to assess the importance both of correlations and interactions among these components; (3) to identify the particular genetic mechanisms involved in the transmission of the traits. Basically, all methods used assume that the phenotypic differences among individuals (that is, the observed trait variance) may be partitioned into genetic, environmental, and interactional components as follows:

(1) $V_P = V_G + V_E + V_I$

in which

V_P = phenotypic or observed variance of a trait;
V_G = genetic contribution to that variance;
V_E = environmental contribution to that variance;
V_I = contribution of interaction between genetic and environmental factors.

Contributions to phenotypic variance

V_I represents a component of variance that will occur when different genotypes respond differently to the same environmental treatment. For example, a genetically "bright" child may gain more from enriched rearing conditions than a genetically "dull" child. It is important to note that this source of variance is difficult to estimate in ordinary human populations, since, more often than not, there is already present a correlation between genotype and environment. This reflects evidence that "brighter" children tend to be more favoured than dull children in respect to environmental treatment. The terms V_G and V_E can, under special circumstances, be broken down further. Thus V_G may be partitioned into additive and nonadditive or dominance components. Likewise, V_E may be divided into as many components as can be observationally identified. One study on heritability of human birth weight can serve as an example. The data are shown as Table 3.

Table 3: Genetic and Environmental Components of Variation in Human Birth Weight

component of variation	percent of total
Genetic	
Additive	15
Nonadditive	1
Sex	2
Total genotype	18
Environmental	
Maternal genotype	20
Maternal environment, general	18
Maternal environment, immediate	6
Age of mother	1
Parity	7
Intangible	30
Total environmental	82

Generally, heritability is estimated as the ratio of additive genetic to total phenotypic variance. But, since in much work with human beings this component cannot be separated readily, the ratio of total genetic to total phenotypic variance is used. This is sometimes called degree of genetic determination rather than heritability.

Basically, all methods used to study the genetics of human data are based on a comparison of similarity among relatives of varying degrees of kinship as well as similarity among relatives as against nonrelatives. Methods commonly used to make these comparisons include the classical approach through correlations involving relatives, particularly monozygotic (MZ, or identical) and dizygotic (DZ, or fraternal) twin pairs. Heritability estimates may be calculated from correlations between parents, half-brothers, half-sisters, and full siblings. Some information regarding mode of inheritance is yielded by comparing correlations among siblings with those between parents and their children. When dominant genes are present sibling correlations are higher than are those between parents and children. Likewise, twin data provide a number of indexes of heritability (H or HR) in the broad sense. Some examples are:

(2) $$H = \frac{r_{MZ} - r_{DZ}}{1 - r_{DZ}}$$

and

(3) $$HR = \frac{2(r_{MZ} - r_{DZ})}{r_{MZ}}$$

in which r_{MZ} and r_{DZ} are correlations for identical and fraternal twin pairs, respectively.
A variant of these is given by:

(4) $$E = \frac{r_{MZT} - r_{MZA}}{1 - r_{MZA}}$$

where E is environmental influence, and r_{MZT}, r_{MZA} are correlations for identical twin pairs reared together and apart, respectively.

These formulas and alternative forms of them also can be expressed in terms of variances rather than with correlations. Further, if the investigator is interested merely in establishing that hereditary factors have a statistically significant effect, (rather than attempting an estimate of their magnitude) MZ and DZ variances can be compared by means through which a probability value may be ascertained (*e.g.*, the so-called F ratio).

Multiple abstract variance analysis

Another approach, known as multiple abstract variance analysis (MAVA), leads to the estimation of so-called nature-nurture ratios both within and between families and to estimations of correlations between genetic and environmental influences as found within families and within whole cultures. Basically, MAVA starts with sets of observed variances that are then expressed in terms of abstract or hypothetical variances and covariances. For example, the variance between identical twins reared apart represent the variance due to environment, much as does E in equation (4). Likewise, the variance between siblings reared together is expressed as:

(5) $$\sigma^2_{ST} = \sigma^2_{wg} + \sigma^2_{we} + 2r_{wg \cdot we}\sigma_{wg}\sigma_{we}$$

in which

σ^2_{ST} = variance of siblings reared together
σ^2_{wg} = variance component due to genetic differences within the family
σ^2_{we} = variance component due to environmental influences within the family
$2r_{wg \cdot we}\sigma_{wg}\sigma_{we}$ = twice the covariance between genetic and environmental components within the family.

A great many other empirical variance estimates can be obtained that can likewise be equated to a number of abstract variances, some of which will be common to several equations. In addition, some of the unknowns (particularly, covariance terms) can be dropped if they seem not to correspond to empirical situations. The result is a possible solution for the abstract variances separately in terms of the empirical variances to which they have been equated. Applications of this method to data on human intelligence and personality have indicated strong hereditary determination for intelligence and for 2 out of 11 personality factors (schizothymia and impulsivity). This promising MAVA approach, however, takes no explicit account of interactions between genetic and environmental influences. Also the number of abstract variances that can be put into an equation are so numerous as to make analysis unwieldy. For example, it may be elected to specify that the kinds of treatment accorded to identical twins, fraternal twins, and siblings are a priori distinct. Again, a special term representing degree of assortative mating in the population could be introduced. The question of which of these terms can be logically dropped or included is thorny. Nor does MAVA generate conclusions regarding the mode of genetic transmission of a trait, nor does it indicate hypotheses concerning its evolutionary history.

Another approach used to study human behavioural data is the biometrical genetic analysis initiated by British statistician Sir Ronald A. Fisher (1890–1962). Application of the method to twin and family data has permitted the remarkably sophisticated conclusions that IQ shows high heritability, about 70 to 80 percent being attributable to genetic factors; that it is transmitted polygenically (about 100 genes); that superior IQ has a high level of dominance; and that the gene action involved suggests that IQ has undergone much natural selection in human evolution.

All of these methods have been developed for use with continuous traits only. With discrete traits (*e.g.*, metabolic defects), applications of the so-called Hardy-Weinberg rule may be made; that is, gene frequencies are estimated for the trait and then predictions are generated concerning incidence of the trait in offspring from parents of various genotypes. This method has been successfully used with phenylketonuria, Huntington's chorea (a hereditary disorder of the nervous system), and many other syndromes. An attempted application of the method to twin and family schizophrenia data, however, yielded indefinite results.

BIBLIOGRAPHY. J.L. FULLER and W.R. THOMPSON, *Behavior Genetics* (1960), a comprehensive text that covers basic methods and surveys human and laboratory animal studies of the genetics of sensory and response systems, intelligence, temperament, and psychiatric disturbance; D.C. GLASS (ed.), *Genetics* (1967), contains a variety of approaches to behaviour genetics, and particularly interesting chapters relating genetics and the social sciences; J. HIRSCH (ed.), *Behavior-genetic Analysis* (1967), a relatively technical treatment of the subject; M. MANOSEVITZ, G. LINDZEY, and D.D. THIESSEN (eds.), *Behavioral Genetics: Method and Research* (1969), includes reviews on most aspects of behaviour genetics in humans and other animals; D. ROSENTHAL and S.S. KETY, *The Transmission of Schizophrenia* (1968), an account of data on the heritability of the disorder; S.G. VANDENBERG (ed.), *Methods and Goals in Human Behaviour Genetics* (1965), a survey that takes up a number of the major statistical problems as well as the methodology in dealing with the genetic analysis of human intelligence and personality.

(W.R.T.)

Human Culture

Culture may be defined as behaviour peculiar to *Homo sapiens*, together with material objects used as an integral part of this behaviour; specifically, culture consists of language, ideas, beliefs, customs, codes, institutions, tools, techniques, works of art, rituals, ceremonies, and so on. The existence and use of culture depends upon an ability possessed by man alone. This ability has been called variously the capacity for rational or abstract thought, but a good case has been made for rational behaviour among subhuman animals, and the meaning of abstract is not sufficiently explicit or precise. Thus the term symbolling has been proposed as a more suitable name for man's unique mental ability—symbolling consisting of assigning to things and events certain meanings that cannot be grasped with the senses alone. Articulate speech—language—provides a good example. The meaning of the word dog is not inherent in the sounds themselves; it is assigned, freely and arbitrarily, to the sounds by human beings. Holy water, "biting one's thumb" at someone (*Romeo and Juliet*, Act I, scene 1), or fetishes are other examples. Symbolling is a kind of behaviour objectively definable (symbolling should not be confused with symbolizing, which has an entirely different meaning).

Culture and "symbolling"

This article is divided into the following sections:

The concept of culture
  Various definitions of culture
    Universalist approaches to culture and the human mind
    Relativist approaches to sociocultural systems
  Culture and personality
  Cultural comparisons
    Ethnocentrism
    Cultural relativism
    Evaluative grading
  Cultural adaptation and change
    Ecological or environmental change
    Diffusion
    Acculturation
    Evolution
Approaches to the study of culture
  Viewing culture in terms of patterns and configurations

Viewing culture in terms of institutional structure and functions
Social organization
Economic systems
Education
Religion and belief
Custom and law

The concept of culture

VARIOUS DEFINITIONS OF CULTURE

What has been termed the classic definition of culture was provided by the 19th-century English anthropologist Edward Burnett Tylor in the first paragraph of his *Primitive Culture* (1871):

> Culture . . . is that complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society.

In *Anthropology* (1881) Tylor made it clear that culture, so defined, is possessed by man alone. This conception of culture served anthropologists well for some 50 years. With the increasing maturity of anthropological science, further reflections upon the nature of their subject matter and concepts led to a multiplication and diversification of definitions of culture. In *Culture: A Critical Review of Concepts and Definitions* (1952), U.S. anthropologists A.L. Kroeber and Clyde Kluckhohn cited 164 definitions of culture—ranging from "learned behaviour" to "ideas in the mind," "a logical construct," "a statistical fiction," "a psychic defense mechanism," and so on. The definition—or the conception—of culture that is preferred by Kroeber and Kluckhohn and also by a great many other anthropologists in recent years is that culture is an abstraction or, more specifically, "an abstraction from behaviour."

These conceptions have defects or shortcomings. The existence of behavioral traditions—that is, patterns of behaviour transmitted by social rather than by biological hereditary means—has definitely been established for nonhuman animals. "Ideas in the mind" become significant in society only as expressed in language, acts, and objects. "A logical construct" or "a statistical fiction" is not specific enough to be useful. The conception of culture as an abstraction led, first, to a questioning of the reality of culture (inasmuch as abstractions were regarded as imperceptible) and, secondly, to a denial of its existence: thus, the subject matter of nonbiological anthropology, "culture," was defined out of existence; and without real, objective things and events in the external world there can be no science.

Kroeber and Kluckhohn were led to their conclusion that culture is an abstraction by reasoning that if culture is behaviour it, ipso facto, becomes the subject matter of psychology; therefore, they concluded that culture "is an abstraction from concrete behavior but is not itself behavior." But what, one might ask, is an abstraction of a marriage ceremony or a pottery bowl, to use Kroeber and Kluckhohn's examples? This question poses difficulties that were not adequately met by these authors. A solution was perhaps provided by Leslie A. White in an essay, "The Concept of Culture" (1959). The issue is not really whether culture is real or an abstraction, he reasoned; the issue is the context of the scientific interpretation.

Importance of the context

When things and events are considered in the context of their relation to the human organism, they constitute behaviour; when they are considered not in terms of their relation to the human organism but in their relationship to one another, they become culture by definition. The mother-in-law taboo is a complex of concepts, attitudes, and acts. When one considers them in their relationship to the human organism—that is, as things that the organism does—they become behaviour by definition. When, however, one considers the mother-in-law taboo in its relationship to the place of residence of a newly married couple, to the customary division of labour between the sexes, to their respective roles in the society's mode of subsistence and offense and defense, and these in turn to the technology of the society, the mother-in-law taboo becomes, again by definition, culture. This distinction is precisely the one that students of words have made for many years. When words are considered in their relationship to the human organism—that is, as acts—they become behaviour. But when they are considered in terms of their relationship to one another—producing lexicon, grammar, syntax, and so forth—they become language, the subject matter not of psychology but of the science of linguistics. Culture, therefore, is the name given to a class of things and events dependent upon symbolling that are considered in a kind of extrahuman context.

Universalist approaches to culture and the human mind. Culture, as noted above, is due to an ability possessed by man alone. The question of whether the difference between the mind of man and that of the lower animals is one of kind or of degree has been debated for many years, and even today reputable scientists can be found on both sides of this issue. But no one who holds the view that the difference is one of degree has adduced any evidence to show that nonhuman animals are capable, to any degree whatever, of a kind of behaviour that all human beings exhibit. This kind of behaviour may be illustrated by the following examples: remembering the sabbath to keep it holy, classifying one's relatives and distinguishing one class from another (such as uncles from cousins), defining and prohibiting incest, and so on. There is no reason or evidence that will lead one to believe that any animal other than man can have or be brought to any appreciation or comprehension whatever of such meanings and acts. There is, as Tylor argued long ago, a "mental gulf that divides the lowest savage from the highest ape" (*Anthropology*).

In line with the foregoing distinction, human behaviour is to be defined as behaviour consisting of, or dependent upon, symbolling rather than upon anything else that *Homo sapiens* does; coughing, yawning, stretching, and the like are not human.

Next to nothing is yet known about the neuroanatomy of symbolling. Man is characterized by a very large brain, considered both absolutely and relatively, and it is reasonable—and even obligatory—to believe that the central nervous system, especially the forebrain, is the locus of the ability to symbol. But how it does this and with what specific mechanisms remain to be discovered. One is thus led to the conclusion that at some point in the evolution of primates, a threshold was reached in some line, or lines, when the ability to symbol was realized and made explicit in overt behaviour. There is no intermediate stage, logical or neurological, between symbolling and nonsymbolling; an individual or a species is capable of symbolling or he or it is not. The life of Helen Keller makes this clear: when, through the aid of her teacher, Anne Sullivan, Helen was enabled to escape from the isolation to which her blindness and deafness had consigned her and to effect contact with the world of human meanings and values, the transformation was instantaneous.

Evolution of "minding." But even if almost nothing is known about the neuroanatomy of symbolling, a great deal is known about the evolution of mind (or "minding," if mind is considered as a process rather than a thing), in which one finds symbolling as the characteristic of a particular stage of development. The evolution of minding can be traced in the following sequence of stages. First is the simple reflexive stage, in which behaviour is determined by the intrinsic properties of both the organism and the thing reacted to—for example, the contraction of the pupil of the eye under increased stimulation by light. Second is the conditioned reflex stage, in which the response is elicited not by properties intrinsic in the stimulus but by meanings that the stimulus has acquired for the responding organism through experience—for example, Pavlov's dog's salivary glands responding to the sound of a bell. Third is the instrumental stage, as exemplified by a chimpanzee knocking down a banana with a stick. Here the response is determined by the intrinsic properties of the things involved (banana, stick, chimpanzee's neurosensory–muscular system); but a new element has been introduced into behaviour, namely,

the exercise of control by the reacting organism over things in the external world. And, finally, there is the symbol stage, in which the configuration of behaviour involves nonintrinsic meanings, as has already been suggested.

Evolution toward increased security and expanse of life

These four stages exhibit a characteristic of the evolution of all living things: a movement in the direction of making life more secure and enduring. In the first stage the organism distinguishes between the beneficial, the injurious, and the neutral, but it must come into direct contact with the object or event in question to do so. In the second stage the organism may react at a distance, as it were—that is, through an intermediate stimulus. The conditioned reflex brings signs into the life process; one thing or event may serve as an indication of something else—food, danger, and so forth. And, since anything can serve as a sign of anything else (a green triangle can mean food, sex, or an electric shock to the laboratory rat), the reactions of the organism are emancipated from the limitations that stage one imposes upon living things, namely, the intrinsic properties of things. The possibility of obtaining life-sustaining things and of avoiding life-destroying things is thus much enhanced, and the security and continuity of life are correspondingly increased. But in stage two the organism still plays a subordinate role to the external world; it does not and cannot determine the significance of the intermediary stimulus: the bark of a distant dog to the rabbit or the sound of the bell to Pavlov's dog. This meaning is determined by things and events in the external world (or in the laboratory by the experimenter). In stages one and two, therefore, the organism is at the mercy of the external world in this respect.

In the third stage the element of control over environment is introduced. The ape who obtains food by means of a stick (tool) is not subordinate to his situation. He does not merely undergo a situation; he dominates it. His behaviour is not determined by the juxtaposition of things and events; on the contrary, the juxtaposition is determined by the ape. He is confronted with alternatives, and he makes choices. The configuration of behaviour in stage three is constructed within the dynamic organism of the ape and then imposed upon the external world.

The evolution of minding is a cumulative process; the achievements of each stage are carried on into the succeeding one or ones. The fourth stage reintroduces the factor of nonintrinsic meanings to the advances made in stages two and three. Stage four is the stage of symbolling, of articulate speech. Thus one observes two aspects of the evolution of minding, both of which contribute to the security and survivability of life: the emancipation of behaviour from limitations imposed upon it by the external world, on the one hand, and increased control over the environment, on the other. To be sure, neither emancipation nor control becomes complete, but quantitative increase is significant.

Evolution of culture. The direction of biological evolution toward greater expansion and security of life can be seen from another point of view: the advance from instinctive behaviour (*i.e.*, responses determined by intrinsic properties of the organism) to learned and freely variable behaviour, patterns of which may be acquired and transmitted from one individual and generation to another, and finally to a system of things and events, the essence of which is meanings that cannot be comprehended by the senses alone. This system is, of course, culture, and the species is the human species. Culture is a man-made environment, brought into existence by the ability to symbol.

Once established, culture has a life of its own, so to speak; that is, it is a continuum of things and events in a cause and effect relationship; it flows down through time from one generation to another. Since its inception 1,000,000 or more years ago, this culture—with its language, beliefs, tools, codes, and so on—has had an existence external to each individual born into it. The function of this external, man-made environment is to make life secure and enduring for the society of human beings living within the cultural system. Thus, culture may be seen as the most recent, the most highly developed means of promoting the security and continuity of life, in a series that began with the simple reflex.

Absence of culture in primate societies

Society preceded culture; society, conceived as the interaction of living beings, is coextensive with life itself. Man's immediate prehuman ancestors had societies, but they did not have culture. Recent studies of monkeys and apes have greatly enlarged scientific knowledge of their social life—and, by inference, the scientific conception of the earliest human societies. Data derived from paleontological sources and from accumulating studies of living, nonhuman primates are now fairly abundant, and hypotheses derived from these are numerous and varied in detail. A fair summary of them may be made as follows: The growth of the primate brain was stimulated by life in the trees, specifically, by eye–hand coordinations involved in swinging from limb to limb and by manipulating food with the hands (as among the insectivorous lemurs). Descent to the ground, as a consequence of deforestation or increase in body size (which would tend to restrict arboreal locomotion and increase the difficulty of obtaining enough food to supply increased need), and the assumption of erect posture were other significant steps in biological evolution and the eventual emergence of culture. Some theories reject the arboreal stage in man's evolutionary past, but this does not seriously affect the overall conception of his development.

The Australopithecines of Africa, extinct manlike higher primates about which reliable knowledge is very considerably today, exemplify the stage of erect posture in primate evolution. Erect posture freed the arms and hands from their earlier function of locomotion and made possible an extensive and versatile use of tools. Again, the eye–hand–object coordinations involved in tool using stimulated the growth of the brain, especially the forebrain. It is not possible to determine on the basis of paleontological evidence the precise point at which the ability to symbol (specifically, articulate speech) was realized, as expressed in overt behaviour. It is believed by some that man's prehuman ancestors used tools habitually and that habit became custom through the transmission of tool using from one generation to another long before articulate speech came into being. In fact, some theorists hold, the customary use of tools became a powerful stimulus in the development of a brain that was capable of symbolling or articulate speech.

Immediate effects of symbolling

The introjection of symbolling into primate social life was revolutionary. Everything was transformed, everything acquired new meaning; the symbol added a new dimension to primate—now human—existence. An axe was no longer merely a tool with which to chop; it could become a symbol of authority. Mating became marriage, and all social relationships between parents and children, between brothers and sisters became moral obligations, duties, rights, and privileges. The world of nature, from the stones beside the path to the stars in their courses, became alive and conscious spirits. "And all that I beheld respired with inward meaning" (Wordsworth). The anthropoid had at last become a man.

Relativist approaches to sociocultural systems. Thus far in this article, culture has been considered in general, as the possession of all mankind. Now it is appropriate to turn to particular cultures, or sociocultural systems. Human beings, like other animal species, live in societies, and each society possesses culture. It has long been customary for ethnologists to speak of Seneca culture, Eskimo culture, North American Plains culture, and so on—that is, the culture of a particular society (Seneca) or an indefinite number of societies (Eskimo) or the cultures found in or characteristic of a topographic area (the North American Plains). There is no objection to this usage as a convenient means of reference: "Seneca culture" is the culture that the Seneca tribe possesses at a particular time. Similarly, Eskimo culture refers to a class of cultures, and Plains culture refers to a type of culture. What is needed is a term that defines culture precisely in its particular manifestations for the purpose of scientific study, and for this the term sociocultural

system has been proposed. It is defined as the culture possessed by a distinguishable and autonomous group (society) of human beings, such as a tribe or a modern nation. Cultural elements may pass freely from one system to another (cultural diffusion), but the boundary provided by the distinction between one system and another (Seneca, Cayuga; United States, Japan) makes it possible to study the system at any given time or over a period of time.

Every human society, therefore, has its own sociocultural system: a particular and unique expression of human culture as a whole. Every sociocultural system possesses the components of human culture as a whole—namely, technological, sociological, and ideological. But sociocultural systems vary widely in their structure and organization. These variations are attributable to differences among physical habitats and the resources that they offer or withhold for human use; to the range of possibilities inherent in various areas of activity, such as language or the manufacture and use of tools; and to degree of development. The biological factor of man may, for purposes of analysis and comparison of sociocultural systems, be considered as a constant. Although the equality or inequality of races, or physical types, of mankind has not been established by science, all evidence and reason lead to the conclusion that whatever differences of native endowment may exist, they are insignificant as compared with the overriding influence of the external tradition that is culture.

CULTURE AND PERSONALITY

Since the infant of the human species enters the world cultureless, his behaviour—his attitudes, values, ideals, and beliefs, as well as his overt motor activity—is powerfully influenced by the culture that surrounds him on all sides. It is almost impossible to exaggerate the power and influence of culture upon the human animal. It is powerful enough to hold the sex urge in check and achieve premarital chastity and even voluntary vows of celibacy for life. It can cause a person to die of hunger, though nourishment is available, because some foods are branded unclean by the culture. And it can cause a person to disembowel or shoot himself to wipe out a stain of dishonour. Culture is stronger than life and stronger than death. Among subhuman animals, death is merely the cessation of the vital processes of metabolism, respiration, and so on. In the human species, however, death is also a concept; only man knows death. But culture triumphs over death and offers man eternal life. Thus, culture may deny satisfactions on the one hand while it fulfills desires on the other.

The powerful influence of culture

The predominant emphasis, perhaps, in studies of culture and personality has been the inquiry into the process by which the individual personality is formed as it develops under the influence of its cultural milieu. But the individual biological organism is, itself, a significant determinant in the development of personality. The mature personality is, therefore, a function of both biological and cultural factors, and it is virtually impossible to distinguish these factors from each other and to evaluate the magnitude of each in particular cases. If the cultural factor were a constant, personality would vary with the variations of the neurosensory–glandular–muscular structure of the individual. But there are no tests that can indicate, for example, precisely how much of the taxicab driver's ability to make change is due to innate endowment and how much to cultural experience. Therefore, the student of culture and personality is driven to work with "modal personalities," that is, the personality of the typical Crow Indian or the typical Frenchman insofar as this can be determined. But it is of interest, theoretically at least, to note that even if both factors, the biological and the cultural, were constant—which they never are in actuality—variations of personality would still be possible. Within the confines of these two constants, individuals might undergo a number of profound experiences in different chronological permutations. For example, two young women might have the same experiences of (1) having a baby, (2) graduating from college, and (3) getting married. But the effect of sequence (1), (2), (3) upon personality development would be quite different than that of sequence (2), (3), (1).

CULTURAL COMPARISONS

Ethnocentrism. Ethnocentrism is the name given to a tendency to interpret or evaluate other cultures in terms of one's own. This tendency has been, perhaps, more prevalent in modern nations than among preliterate tribes. The citizens of a large nation, especially in the past, have been less likely to observe people in another nation or culture than were members of small tribes who were well acquainted with the ways of their culturally diverse neighbours. Thus, the American tourist could report that Londoners drive "on the wrong side of the street" or an Englishman might find some customs on the continent "queer" or "boorish," merely because they were different. Members of a Pueblo tribe in the American Southwest, on the other hand, might be well acquainted with cultural differences not only among other Pueblos but also in non-Pueblo tribes such as the Navajo and Apache.

Ethnocentrism in primitive and modern cultures

Ethnocentrism became prominent among many Europeans after the discovery of the Americas, the islands of the Pacific, and the Far East. Even anthropologists might characterize all preliterate peoples as being without religion (as did Sir John Lubbock) or as having a "prelogical mentality" (as did Lucien Lévy-Bruhl), merely because their ways of thinking did not correspond with those of the culture of western Europe. Thus, inhabitants of non-Western cultures, particularly those lacking the art of writing, were widely described as being immoral, illogical, queer, or just perverse ("Ye Beastly Devices of ye Heathen").

Cultural relativism. Increased knowledge led to or facilitated a deeper understanding and, with it, a finer appreciation of cultures quite different from one's own. When it was understood that universal needs could be served with culturally diverse means, that worship might assume a variety of forms, that morality consists in conforming to ethical rules of conduct rather than inhering in the rules themselves, a new view emerged that each culture should be understood and appreciated in terms of itself. What is moral in one culture might be immoral or ethically neutral in another. For example, it was not immoral to kill a baby girl at birth or an aged grandparent who was nonproductive when it was impossible to obtain enough food for all; or wife lending among the Eskimo might be practiced as a gesture of hospitality, a way of cementing a friendship and promoting mutual aid in a harsh and dangerous environment, and thus acquire the status of a high moral value.

The view that elements of a culture are to be understood and judged in terms of their relationship to the culture as a whole—a doctrine known as cultural relativism—led to the conclusion that the cultures themselves could not be evaluated or graded as higher and lower, superior or inferior. If it was unwarranted to say that patriliny (descent through the male line) was superior or inferior to matriliny (descent through the female line), if it was unjustified or meaningless to say that monogamy was better or worse than polygamy, then it was equally unsound or meaningless to say that one culture was higher or superior to another. A large number of anthropologists subscribed to this view; they argued that such judgments were subjective and therefore unscientific.

The question of subjective judgments

It is, of course, true that some values are imponderable and some criteria are subjective. Are people in modern Western culture happier than the Aborigines of Australia? Is it better to be a child than an adult, alive than dead? These certainly are not questions for science. But to say that the culture of the ancient Mayas was not superior to or more highly developed than the crude and simple culture of the Tasmanians or to say that the culture of England in 1966 was not higher than England's culture in 1066 is to fly in the face of science as well as of common sense.

Evaluative grading. Cultures have ponderable values as well as imponderable, and they can be measured with

objective, meaningful yardsticks. A culture is a means to an end: the security and continuity of life. Some kinds of culture are better means of making life secure than others. Agriculture is a better means of providing food than hunting and gathering. The productivity of human labour has been increased by machinery and by the utilization of the energy of nonhuman animals, water and wind power, and fossil fuels. Some cultures have more effective means of coping with disease than others, and this superiority is expressed mathematically in death rates. And there are many other ways in which meaningful differences can be measured and evaluations made. Thus, the proposition that cultures have ponderable values that can be measured meaningfully by objective yardsticks and arranged in a series of stages, higher and lower, is substantiated. But, it should be noted, this is not equivalent to saying that man is happier or that the dignity of the individual (an imponderable) is greater in an industrialized or agriculturized sociocultural system than in one supported by human labour alone and sustained wholly by wild foods.

Actually, however, there is no necessary conflict between the doctrine of cultural relativism and the thesis that cultures can be objectively graded in a scientific manner. It is one thing to reject the statement that monogamy is better than polygamy and quite another to deny that one kind of sociocultural system contains a better means of providing food or combatting disease than another.

CULTURAL ADAPTATION AND CHANGE

Ecological or environmental change. Every sociocultural system exists in a natural habitat, and, of course, this environment exerts an influence upon the cultural system. The cultures of some Eskimo groups present remarkable instances of adaptation to environmental conditions: tailored fur clothing, snow goggles, boats and harpoons for hunting sea mammals, and, in some instances, hemispherical snow houses, or igloos. Some sedentary, horticultural tribes of the upper Missouri River went out into the Great Plains and became nomadic hunters after the introduction of the horse. The culture of the Navajos underwent profound change after they acquired herds of sheep and a market for their rugs was developed. The older theories of simple environmentalism, some of which maintained that even styles of myths and tales were determined by topography, climate, flora, and other factors, are no longer in vogue. The present view is that the environment permits, at times encourages, and also prohibits the acquisition or use of certain cultural traits but otherwise does not determine culture change. The Fuegians living at the southern tip of South America, as viewed by Charles Darwin on his voyage on the "Beagle," lived in a very cold, harsh environment but were virtually without both clothing and dwellings.

Diffusion. "Culture is contagious," as a prominent anthropologist once remarked, meaning that customs, beliefs, tools, techniques, folktales, ornaments, and so on, may diffuse from one people or region to another. To be sure, a culture trait must offer some advantage, some utility or pleasure, to be sought and accepted by a people. (Some anthropologists have assumed that basic features of social structure, such as clan organization, may diffuse, but a sounder view holds that these features involving the organic structure of the society must be developed within societies themselves.) The degree of isolation of a sociocultural system—brought about by physical barriers such as deserts, mountain ranges, and bodies of water—has, of course, an important bearing upon the ease or difficulty of diffusion. Within the limits of desirability on the one hand and the possibility of communication on the other, diffusion of culture has taken place everywhere and in all times. Archaeological evidence shows that amber from the Baltic region diffused to the Mediterranean coast; and, conversely, early coins from the Near East found their way to northern Europe. In aboriginal North America, copper objects from northern Michigan have been found in mounds in Georgia; macaw feathers from Central America turn up in archaeological sites in northern Arizona. Some Indian tribes in northwestern regions of the United States had possessed horses, originally brought into the Southwest by Spanish explorers, years before they had ever even seen white men. The wide dispersion of tobacco, maize, coffee, the sweet potato, and many other traits are conspicuous examples of cultural diffusion.

Evidence of diffusion in the ancient past

Acculturation. Diffusion may take place between tribes or nations that are approximately equal in political and military power and of equivalent stages of cultural development, such as the spread of the sun dance among the Plains tribes of North America. But in other instances, it takes place between sociocultural systems differing widely in this respect. Conspicuous examples of this have been instances of conquest and colonization of various regions by the nations of modern Europe. In these cases it is often said that the culture of the more highly developed nation is "imposed" upon the lesser developed peoples and cultures, and there is, of course, much truth in this; the acquisition of foreign culture by the subject people is called acculturation and is manifested by the indigenous populations of Latin America and other regions. But even in cases of conquest, traits from the conquered peoples may diffuse to those of the more advanced cultures; examples might include, in addition to the cultivated plants cited above, words ("coyote"), musical themes, games, and art motifs.

One of the major problems of ethnology during the latter half of the 19th and the early decades of the 20th centuries was the question "How are cultural similarities in noncontiguous regions to be explained?" Did the concepts of pyramid building, mummification, and sun worship originate independently in ancient Egypt and in the Andean highlands and in Yucatán or did these traits originate in Egypt and diffuse from there to the Americas, as some anthropologists have believed? Some schools of ethnological theory have held to one view; some, to another. The 19th-century classical evolutionists (Edward Burnett Taylor, Lewis H. Morgan, and others) held that the mind of man is so constituted or endowed that he will develop cultures everywhere along the same lines (see also CIVILIZATION AND CULTURAL EVOLUTION). "Diffusionists"—those, such as Fritz Graebner and Elliot Smith, who offered grand theories about the diffusion of traits all over the world—maintained that man was inherently uninventive and that culture, once created, tended to spread everywhere. Each school tended to insist that its view was the correct one and to hold it unless definite proof of the contrary could be adduced. The tendency nowadays is not to side categorically with one school as against another but to decide each case on its own merits. The consensus with regard to pyramids is that they were developed independently in Egypt and the Americas because they differ markedly in structure and function: the Egyptian pyramids were built of stone blocks and contained tombs within their interiors. The American pyramids were constructed of earth, then faced with stone, and they served as the bases of temples. The verdict with regard to the bow and arrow is that it was invented only once and subsequently diffused to all regions where it has been found. The probable antiquity of the origin of fire making, however, and the various ways of generating it—by percussion, friction, compression (fire pistons)—indicate multiple origins.

Evolution. Evolution of culture—that is, the development of forms through time—has taken place. No amount of diffusion of picture writing could of itself, for instance, produce the alphabetic system of writing; as Tylor demonstrated so well, the art of writing has developed through a series of stages: picture writing, hieroglyphic writing, and alphabetic writing. In the realm of social organization there was a development from territorial groups composed of families to segmented societies (clans and larger groupings). Sociocultural evolution, like biological evolution, exhibits a progressive differentiation of structure and specialization of function.

A misunderstanding has arisen with regard to the relationship between evolution and diffusion. It has been argued, for example, that the theory of cultural evolution

was unsound because some peoples skipped a stage in a supposedly determined sequence; for example, some African tribes, as a consequence of diffusion, went from the Stone Age to the Iron Age without an intermediate age of copper and bronze. But the classical evolutionists did not maintain that peoples, or societies, had to pass through a fixed series of stages in the course of development, but that tools, techniques, institutions—in short, *culture*—had to pass through the stages. The sequence of stages of writing did not mean that a society could not acquire the alphabet without working its way through hieroglyphic writing; it was obvious that many peoples did skip directly to the alphabet.

Relationship between diffusion and evolution

Approaches to the study of culture

VIEWING CULTURE IN TERMS OF PATTERNS AND CONFIGURATIONS

Cultural traits. The concept of culture embraces the culture of mankind as a whole. An understanding of human culture is facilitated, however, by analyzing "the complex whole" into component parts or categories. In somewhat the same sense that the atom has been regarded as the unit of matter, the cell as the unit of life, so the culture trait is generally regarded as the unit of culture. A trait may be an object (knife), a way of doing something (weaving), a belief (in spirits), or an attitude (the so-called horror of incest). But, within the category of culture, each trait is related to other traits. A distinguishable and relatively self-contained cluster of traits is conventionally called a culture complex. The association of traits in a complex may be of a functional and mechanical nature, such as horse, saddle, bridle, quirt, and the like, or it may lie in conceptional or emotional associations, such as the acts and attitudes involved in seclusion in a menstrual hut or retrieving a heart that has been stolen by witches.

Cultural areas. The relationship between an actual culture and its habitat is always an intimate one, and therefore one finds a more or less close correlation between kind of habitat and type of culture. This results in the concept of culture area. This conception goes back at least as far as the early 19th century, but it was first brought into prominence by the U.S. anthropologist Clark Wissler in *The American Indian* (1917) and *Man and Culture* (1923). He divided the Indian cultures (as they were in the latter half of the 19th century) into geographical cultural regions: the Caribou area of northern Canada; the Northwest coast, characterized by use of salmon and cedar; the Great Plains, where tribes hunted the bison with the horse; the Pueblo area of the Southwest; and so on. Others later distinguished culture areas in other continents.

Cultural types. Appreciation of the relationship between culture and topographic area suggests the concept of culture type, such as hunting and gathering or a special way of hunting, such as use of the horse in bison hunting in the Plains or the hunting of sea mammals among the Eskimo; pastoral cultures centred upon sheep, cattle, reindeer, etc.; horticulture (with digging stick and hoe) and agriculture (with ox-drawn plow). Less common are trading cultures such as are found in Melanesia or specialized production of some object for trade, such as pottery, bronze axes, or salt, as was the case in Luzon.

Culture configurations and patterns

Configuration and pattern, especially the latter, are concepts closely related to culture area and culture type. All of them have one thing in common; they view culture not in terms of its individual components, or traits, but as meaningful organizations of traits: areas, occupations, configurations (art, mathematics, physics), or patterns (in which psychological factors are the bases of organization). Clark Wissler's "universal culture pattern" was a recognition of the fact that all particular and actual cultures possess the same general categories: language, art, social organization, religion, technology, and so on.

VIEWING CULTURE IN TERMS OF INSTITUTIONAL STRUCTURE AND FUNCTIONS

Social organization. A sociocultural system presents itself under two aspects: structure and function. As culture evolves, sociocultural systems (like biological systems) become more differentiated structurally and more specialized functionally, proceeding from the simple to the complex. Systems on the lowest stage of development have only two significant kinds of parts: the local territorial group and the family. There is a corresponding minimum of specialization, limited, with but few exceptions, to division of function, or labour, along sex lines and to division between children and adults. The exceptions are headmen and shamans; they are special organs, so to speak, in the body politic. The headman is a mechanism of social integration, direction, and control, expressing, however, the consensus of the band. The shaman, though a self-appointed priest or magician, is also an instrument of society; he may be regarded as the first specialist in the history of human society.

"Classes" and "segments"

All human societies are divided into classes and segments. Class is defined as one of an indefinite number of groupings each of which differs in composition from the other or others, such as men and women; married, widowed, and divorced; children and adults. Segment is defined as one of an indefinite number of groupings all of which are alike in structure and function: families, lineages, clans, and so on. On more advanced levels of development there are occupational classes, such as farmers, pastoralists, artisans, metalworkers, and scribes, and territorial segments, such as wards, barrios, counties, and states.

Segmentation is a cultural process essential to the evolution of culture; it is a means of increasing the size of a society or a grouping within a sociocultural system (such as an army) and therefore of increasing its power to make life secure, without suffering a corresponding loss of effectiveness through diminished solidarity; segmentation is a means of maintaining solidarity at the same time that it enlarges the social grouping. A tribe could not increase in size beyond a certain point without resorting to segmentation: the formation of lineages, clans, and the like. The word clannish points to one of the functions of segments in general: the fostering of solidarity. Tribes become segments in confederacies; and above the tribal level, the evolution of civil society employs barrios, demes, counties, and states in its process of segmentation. In present-day society, the army and the church offer illuminating examples of increased size and sustained solidarity proceeding hand in hand.

Economic systems. Division of labour along occupational lines is rare, although not wholly lacking, in preliterate societies—despite a widespread notion that one member of a tribe specializes in making arrows, which he exchanges for moccasins made by another specialist. Occupational groupings were virtually lacking in all cultural systems of aboriginal North America, for example. Guilds of metalworkers are found in some African tribes and specialists in canoe making and tattooing existed in Polynesia. But it is not until the transition from preliterate society, based upon ties of kinship, to civil society, based upon property relations and territorial distinctions (the state), that division of labour along occupational lines becomes extensive. On this level there are found many kinds of specialists: metalworkers, scribes, astrologers, soldiers, dancers, musicians, alchemists, prostitutes, eunuchs, and so forth.

Distribution and exchange of goods

Production of goods is everywhere followed by distribution and exchange. Among the Kurnai of Australia, for example, game was divided and distributed as follows: the hunter who killed a wallaby, for example, kept the head; his father received the ribs on the right side, his mother the ribs on the left side, plus the backbone, and so on; the various parts of the animal went to various classes of relatives in accordance with fixed, traditional rules.

Distribution along kinship lines constitutes a system of circulation and exchange within the tribe as a whole, for everyone is a relative of everyone else. It takes the form of bestowing gifts to relatives on all sorts of occasions—such as birth, initiation, marriage, death. In some cases there is an exchange of goods on the spot, but more often A gives something to B who gives A a gift at a later date.

All this takes place in the network of rights and obligations among kindred; one has both an obligation to give and a right to receive on certain occasions and in certain contexts. The whole process is one of mutual aid and cooperation.

The consequence of this form of distribution and exchange is that the recipient receives kinds of things that he already has; each household has the same kinds of foods, utensils, ornaments, and other things that every other household has. Why, then, it might be asked, does this form of exchange take place? Two reasons may be distinguished. First, this kind of exchange fortifies ties of kinship and mutual aid—as neighbourhood exchange among households in modern American culture initiates friendships that in times of need constitute mutual aid. Second, this system of circulation of goods is in effect a system of social security: a household in need, due to illness or accident, receives help from the community ("No household can starve as long as others have corn," as the Iroquois put it). Here we have an economic system subordinated to the welfare of the society as a whole.

Exchange or circulation of goods and services (a basket is the material form of "a service," that is, human labour) must, of course, take place in sociocultural systems where division of labour finds expression in specialization: the ironworker must obtain food; the horticulturalist needs an iron hoe.

Exchange of goods between sociocultural systems is universal and takes place on the lowest levels of cultural development. In some instances it is the only form of nonhostile communication: in the so-called silent trade the actual exchange takes place in a neutral zone without the presence of the participating parties. Archeological evidence shows that intergroup exchange occurred in remote times and over great distances, as already noted above in the discussion of diffusion.

An interesting form of the circulation of goods—usually referred to as redistribution—occurs among more highly developed tribes. The head of the sociopolitical system, that is, the chief or priest-chief, imposes levies upon all households, thus acquiring a large amount of goods—food, utensils, art objects, and so on—which he then redistributes to the households of the tribe. This may take the form and occasion of ceremonies and feasts or distribution may be made in cases of need. This widespread and interesting form of redistribution serves the same ends as those served by distribution as a function of the kinship system, namely, fostering solidarity and social security—an equitable distribution that tends to iron out inequalities among households.

Ownership and use

Some economic concepts in modern Western culture do not correspond closely with conceptions and customs in many preliterate societies. Ownership is a case in point. Complete possession of and exclusive right to use something in an economic context, such as land, a dwelling, or a boat, is rare, if not wholly lacking, in preliterate societies (although one might have exclusive rights to a dream, spell, or charm). In general, one has merely the right to use or occupy a tract of land or a house; when its use has terminated, anyone can take it over. In some societies it might be said that a boat "belonged" to the men who made it or even to the individual who initiated its construction. But anyone else in the community would have the right to use it when the "owners" (the men who made it) were not using it. It is right to use, rather than exclusive and absolute possession, that is significant; there is no such thing as absentee ownership in primitive society.

A band or tribe "holds" the land it occupies; here again, it is tenure rather than ownership that is significant; the land "belongs" to Nature, or Mother Earth; people merely hold and use it. There is usually an intimate relationship between the people and "their" land. Navajo Indians fell on their knees and kissed the earth when they were returned to their former territory after forcible detention in an alien land. Land is defended against outsiders, except when they are accepted as guests, but the significant thing is not that the outsiders do not own the land but that they pose a threat to those who occupy it.

In some tribes there is a distinct conception that the land held "belongs" to the tribe, the chief of which allots plots or tracts to individuals or households for their use. But when use terminates, the land reverts to the tribal domain.

During the latter part of the 19th century there was considerable discussion of "primitive communism." This doctrine came to be interpreted as meaning that private property, private right to hold or use, was nonexistent in primitive society. It was extended also to communism in wives and children in some tribes; this was interpreted to be a vestige of a former stage of "primordial promiscuity." Many ethnologists, however, launched a vigorous attack upon "the doctrine of primitive communism." Some of the conceptions of earlier anthropologists—such as group marriage—were shown to be unwarranted in the light of later research.

Today, with these polemics well in the past, the situation with regard to property rights in tribal societies may be summarized as follows. Tenure and use, rather than ownership in fee simple, were the significant concepts and practices. Private, or personal, possession of goods and use of land were recognized. But possession and right were qualified by the rights and obligations of kinship: one had an obligation both to give and to receive within the body of kindred, according to specific rules. In a de facto sense, things belonged to the body of kindred; rights of possession and use were regulated by customs of kinship. In some cultures a borrower was not obliged to return an object borrowed, on the theory that if one could afford to lend something he relinquished his right to its possession. The mode of life in preliterate society, based upon kinship and functioning in accordance with the principles of cooperation and mutual aid, did indeed justify the adjective communal; it was the noun communism that was resented—if not feared—because of its Marxist connotation.

One of the most important, as well as characteristic, features of the economic life of preliterate societies, as contrasted with modern civilizations, is this: no individual and no class or group in tribal society was denied access to the resources of nature; all were free to exploit them. This is, of course, in sharp contrast to civil society in which private ownership by some, or a class, is the means of excluding others—slaves, serfs, a proletariat—from the exploitation and enjoyment of the resources of nature. It is this freedom of access, the freedom to exploit and to enjoy the resources of nature, that has given primitive society its characteristics of freedom and equality. And, being based upon kinship ties, it had fraternity, as well (see also ECONOMIC SYSTEMS, PRIMITIVE).

Education. In the human species, individuals are equipped with fewer instincts than is the case in many nonhuman species. And, as already noted, they are born cultureless. Therefore an infant *Homo sapiens* must learn a very great deal, acquire a vast number of conditioned reflexes and habit patterns in order to live effectively, not only in society but in a particular kind of sociocultural system, be it Tibetan, Eskimo, or French. This process, taken as a whole, is called socialization (occasionally, enculturation)—the making of a social being out of one that was at birth wholly individualistic, egoistic.

Socialization or enculturation

Education in its broadest sense may properly be regarded as the process by which the culture of a sociocultural system is impressed or imposed upon the plastic, receptive infant. It is this process that makes continuity of culture possible. Education, formal and informal, is the specific means of socialization. By informal education is meant the way a child learns to adapt his behaviour to that of others, to be like others, to become a member of a group. By formal education is meant the intentional and more or less systematic effort to affect the behaviour of others by transmitting elements of culture to them, be it knowledge or belief, patterns of behaviour, or ideals and values. These attempts may be overt or covert. The teacher may make his purpose apparent, even emphatic, to the learner. But much education is effected in an unobstrusive way, without teacher or learner being

aware that culture is being transmitted. Thus, in myths and tales, certain characters are presented as heroes or villains; certain traits are extolled, others are deplored or denounced. The impressionable child acquires ideals and values, an image of the good or the bad.

The growing child is immersed in the fountain of informal education constantly; the formal education tends to be periodic. Many sociocultural systems distinguish rather sharply a series of stages in the education and development of full-fledged men and women. First there is infancy, during which perhaps the most profound and enduring influences of a person's life are brought to bear. Weaning ushers in a new stage, that of childhood, during which boys and girls become distinguished from each other. Puberty rites translate children into men and women. These vary enormously in emphasis and content. Sometimes they include whipping, isolation, scarification, or circumcision. Very often the ritual is accompanied by explicit instruction in the mythology and lore of the tribe and in ethical codes. Such rituals as confirmation and Bar Mitzwa in modern Western culture belong to the category of puberty rites.

With marriage come instruction and admonition, appropriate to the occasion, from elder relatives and, in more advanced cultures, from priests.

In some sociocultural systems men may become members of associations or sodalities: men's clubs, warrior societies, secret societies of magic or medicine. In some cases it is said that in passing through initiation rites a person is "born again." Women may belong to sodalities, also, and in some instances they may become members of secret, magical societies along with men.

Religion and belief. Man's oldest philosophy is animism, the doctrine that everything is alive and possesses mental faculties like those possessed by man: desire, will, purpose, anger, love, and the like. This philosophy results from man's projection of his own self, his psyche, into other things and beings, inanimate and living, without being aware of this projection. "To the Omaha," wrote anthropologist Alice Fletcher,

> nothing is without life: . . . He projects his own consciousness upon all things and ascribes to them experiences and characteristics with which he is familiar; . . . akin to his own conscious being ("Wakonda," in F.W. Hodge [ed.], *Handbook of American Indians North of Mexico*).

Development of religion

"A belief in spirits is," according to Edward Burnett Tylor, the great pioneer of English anthropology, "the minimum definition of religion." Preceding this stage, according to some students, is a belief in impersonal, supernatural power, or mana (manitou, orenda, etc.). In any case, these two elements of religion are virtually worldwide and undoubtedly represent a very early stage in the development of religion. In some cultures spirits are virtually innumerable, but, in the course of time, the more important spirits become gods. Thus, there has been a tendency toward monotheism in the evolution of religion. The German Roman Catholic priest and anthropologist Father Wilhelm Schmidt argued not only that some primitive peoples believe in a Supreme Being but that monotheism was characteristic of the earliest and simplest cultures. Schmidt's thesis, however, has been severely criticized by other ethnologists. Also, as Tylor pointed out many years ago, the Supreme Being of some very primitive peoples is not, in fact and effect, a god but rather a philosophic explanatory device that accounted for the existence and structure of the world; after his work was completed he had no other significance; he was not worshipped and played no part in the daily lives of the people.

In the past there was much discussion—and debate—about the difference between magic and religion. Both were deemed expressions of a belief in the supernatural. Some argued that religion was social (moral) whereas magic was antisocial (immoral). Another distinction was that magic was the use of supernatural power divorced from a spiritual being. The distinction between religion and magic was so beset with exceptions as to render most definitions of these terms logically imperfect. Another difficulty was the tacit assumption that different entities, religion and magic, exist per se, and therefore that "correct" definitions of them must exist also (Adam called the animal a horse because it was a horse). Much confusion and debate would have been obviated if it had been recognized (as it generally is now) that there is no such thing as a "correct" definition—all definitions are man-made and arbitrary—and that the problem is not what is religion or magic but what beliefs, events, and experiences does one wish to designate with the words religion and magic (see also MAGIC; RELIGION, SOCIAL ASPECTS OF; RELIGION, STUDY OF).

Custom and law. Sociocultural systems, like other kinds of systems, must have means of self-regulation and control in order to persist and function. In human society these means are numerous and varied. The kinship organization specifies reciprocal and correlative rights, duties, and obligations of one class of relatives to another. Codes of ethics govern the relationship of the individual to the well-being of society as a whole. Codes of etiquette regulate class structure by requiring individuals to conform to their respective classes. Custom is a general term that embraces all these mechanisms of regulation and control and even more. Custom is the name given to uniformities in sociocultural systems. Uniformities are important because they make anticipation and prediction possible; without them, orderly conduct of social life would not be possible. Custom, therefore, is a means of social regulation and control, of effecting compliance with itself in order to render effective conduct of social life possible.

Ethics, etiquette, law, and other norms

As in the case of religion and magic, much effort and debate have been spent in attempts to achieve a clean-cut distinction between custom and law. There is little or no difficulty when one is concerned with the extremes of the spectrum of social control. The way that a Hopi Indian holds his corn-husk cigarette in his hand is a matter of custom rather than law, as most ethnologists would probably agree. At the other extreme, a state edict prohibiting the manufacture and sale of alcoholic beverages is a law, not a custom. But in other situations, the distinction is far from clear and disagreement with regard to definitions arises. For example, in marriage the obligation to wed someone within a specified group or class (endogamy) or outside a group or class (exogamy) has been called both law and custom. Probably the most useful distinction between custom and law is the following. If an infraction of a social rule or deviation from a norm is punished merely by expressions of social disapproval, gossip, ridicule, or ostracism, the rule is called custom. If, however, infractions or violations are punished by an agency, designated by society and empowered to act on its behalf, then the rule is called a law. But even here there is difficulty. The same kind of offense may be punished by custom in one society, by law in another: for example, adultery, incest, miscegenation, black magic.

It is the ethnologist, rather than the historian, who is disturbed by instances of ambiguity with regard to custom and law; in preliterate societies the distinction between the two is not always clear. But in civil societies—that is, states brought into being by the agricultural revolution and their more recent successors—the distinction is usually sharper and more apparent, though instances of sumptuary laws that prohibit the wearing of silk or that limit the length of a garment merge law and custom or reinforce the latter by the former.

One need not be unduly disturbed by the difficulty of making sharp distinctions among sociocultural phenomena and of formulating definitions. The phenomena of culture, like those of the external world in general, are what they are, and if man-made concepts and words do not correspond closely with them, one may regret the lack of fit. But it is better to do this than to distort real phenomena by trying to force them into artificial concepts and definitions.

BIBLIOGRAPHY. For a general account of man and culture, see RALPH L. BEALS and HARRY HOIJER, *An Introduction to Anthropology*, 3rd ed. (1965); ELMAN R. SERVICE, *Primitive Social Organization: An Evolutionary Perspective* (1962); and SOL TAX (ed.), *Horizons of Anthropology* (1964). EDWARD

BURNETT TYLOR, *Anthropology* (1881; abr. ed., 1960), is still valuable. The question of whether man is unique psychologically or "just a more highly developed primate" is discussed by LESLIE A. WHITE, "The Symbol: The Origin and Basis of Human Behavior," in *The Science of Culture*, 2nd ed., pp. 22–39 (1969); and by ERNST CASSIRER, *An Essay on Man: An Introduction to a Philosophy of Human Culture* (1944 and 1956). A compilation and critical review of the numerous conceptions of culture are provided by A.L. KROEBER and CLYDE KLUCKHOHN, *Culture: A Critical Review of Concepts and Definitions* (1952). A conception of culture based on the material reviewed by Kroeber and Kluckhohn is found in LESLIE A. WHITE, "The Concept of Culture," *Am. Anthrop.*, 61:227–251 (1959).

(L.A.W.)

Human Cultures, Primitive and Non-Urban

So great are the variations in ways of life, past and present, that comparisons among them are difficult. Any simple classification of human societies and cultures can only be viewed as arbitrary. From a modern urban point of view, nevertheless, there is the obvious distinction between the primitive and the civilized: between *simple* and *complex* societies; between *tiny* and *huge* social agglomerations; between *scattered* and *dense* populations; and, above all, between *prestate* societies and societies that have developed *states*. In general, civilization involves the rise of legal institutions and the acquisition of a legal monopoly of force by a government. Those developments made possible the cities and empires of classical times and the growth of dense populations. Thus "civilized" is nearly synonymous with "urban."

The varieties of non-urban, or primitive, societies may be further classified. One way is by the methods they use to get food. Those who hunt and gather behave quite differently, as societies, from herdsmen and mounted predator-warriors, the *pastoralists,* who in turn live quite differently from the various kinds of *agriculturalists.* These distinctions are not sharp, for of course there are societies that combine foraging with some agriculture; others, some agriculture and some herding; and, in a few cases, a class of herders may live in the same society with a class or caste of agriculturalists. A whole continuum of societies may be constructed, ranging from tiny, simple bands of hunter-gatherers in poor environments to large, dense populations of irrigation agriculturalists —that is, from the entirely nomadic to the fully sedentary. The degree to which societies approach the sedentary deserves prominence in any classification, since sedentary ways are accompanied by so many other cultural traits and institutions.

Primitive societies

NOMADIC SOCIETIES

Throughout 99 percent of the time *Homo sapiens* has been on earth, or until about 8,000 years ago, all peoples were foragers of wild food. There were great differences among them; some specialized in hunting big game, fishing, and shellfish gathering, while others were almost completely dependent on the gathering of wild plants. Broadly speaking, however, they probably shared many features of social and political organization, as well as of religions and other ideologies (in form though not in specific content). The hunting-gathering societies declined with the growth of agricultural societies, which either drove them from their territories or assimilated or converted them.

The later rise of the nation-states, especially after the Industrial Revolution in Europe, resulted in the near extermination of hunting-gathering societies. Today, the remaining ones are confined to desert, mountain, jungle, or Arctic wastelands. Some have been studied and described by anthropologists; the central and northern Australians, the Bushmen of the Kalahari in southern Africa, the Pygmies of the central African forests, the Pygmies of the Andaman Islands in the Indian Ocean, the Ona and Yahgan Indians of southern South America, the "Digger" Indians of Nevada, the Indians of the northern Canadian forests, and the Canadian, Alaskan, and Greenland Eskimos.

All of these peoples inhabit areas representing almost every extreme in climate and environment, but they have one thing in common: their marginality to, or relative isolation from, modern economic systems. Their techniques and forms of food getting vary greatly. The Eskimos, for example, are entirely dependent on hunting and fishing; the African Bushmen, the Australian aborigines, and the Nevada Indians are chiefly dependent on the gathering of seeds, nuts, and tubers.

The significance of nomadism to the student of primitive cultures may be suggested by a comparison of the Ona and Yahgan of Tierra del Fuego. The Ona inhabit the interior forests and depend heavily on hunting guanaco (a small New World camel). The Yahgan are canoe-using fishermen and shellfish gatherers. Yet, despite their utterly different ecological adaptation, the two Indian societies share a common culture that is so similar that anthropologists conventionally lump them together with the neighbouring Chono and Alakaluf of Chile as a "Fuegian culture area." They are all nomadic, though the Ona are "foot Indians" and the others are "canoe Indians"; they are all relatively sparsely scattered over the landscape, poor in material culture, and with similar social, political, ceremonial, and ideological customs and institutions.

Common characteristics of nomadic societies

All of the nomads so far mentioned share important general characteristics. The first and most obvious is that their nomadism severely restricts the amount of their "baggage," or material culture. Bows and arrows (except in Australia, where the unique boomerang is used instead) and perhaps a simple spear javelin, or in some areas throwing sticks or clubs, are the usual hunting and fighting weapons. In warmer zones shelter is a simple lean-to or small beehive hut of sticks, twigs, and leaves. In Arctic zones there are the caribou-skin tent and the famous Eskimo igloo—or, in more permanent or revisited places, the stone hut.

Camps are small and impermanent. The nuclear family likes to camp near related families when possible, and usually this forms the patrilineally extended family—a group of brothers with their own nuclear families and perhaps a few dependent elders. But the size of the camp depends on the season: in times of easily gathered plant food, large groups may come together for ceremonies such as puberty rites. At other times, the constituent families may scatter widely because food and water are scarce. Patrilineally related men and their families, scattered or not, commonly regard themselves as a group with rights over a particular territory and may be distinguished from neighbours on a territorial basis as well. Marriages are often arranged among territorial groups so that contiguous groups tend to be related, or at least certain members of different groups will be related. But this is the only organizing principle that extends beyond the territorial band. Each band may be thought of as part of a larger society comprised of distant as well as close relatives—a "tribe" in one of the original meanings of the word.

The social organization looks as though it had been built up from within, so to speak. Family-like statuses and roles, alliances by marriage, and systems of "social distance" based on family relationships are the bones and connective tissues of the society. These are all ingredients of the family itself, however extended or metaphorically construed; it is as though these societies were simply the result of the growth of individual families. But this is only appearance; such societies also grow by accretion. But inasmuch as alliances and the compounding of different groups normally are brought about by arranged marriages, the familistic appearance of the whole is therefore maintained.

Almost all status positions rest upon the same criteria of age, sex, and kinship distance. The only achieved status is that of the magical curer, the "shaman." Again, with the exception of the shaman, the only division of labour in these societies is on the basis of age and sex—

just as in the individual nuclear family unit. Among adults, the hunting of big game is confined to men, whereas the gathering of vegetable foods or small animals, birds' eggs, and so on are women's tasks. This division of labour seems obviously related to men's propensity and relative ability to range far from camp, women being too burdened with the problems of motherhood to track animals wherever they may lead. But the separation of tasks is usually more rigid and confining than the physical and circumstantial differences between men and women dictate, since these would vary among individuals and from society to society—and for that matter, from day to day. Domestic tasks are strictly defined as female and are undertaken only by women even when they seem exceptionally taxing, as attest the following remarks by Lewis Garrard, who travelled with a Cheyenne Indian camp in 1846:

> After a ride of two hours, we stopped, and the chiefs, fastening their horses, collected in circles, to smoke the pipe and talk, letting their squaws unpack the animals, pitch the lodges, build fires, arrange the robes, and when all was ready, these 'lords of creation' dispersed to their several homes, to wait until their patient and enduring spouses prepared some food. I was provoked, nay, angry, to see the lazy, overgrown men, do nothing to help their wives; and, when the young women pulled off their bracelets and finery, to chop wood, the cup of my wrath was full to overflowing, and, in a fit of honest indignation, I pronounced them ungallant, and indeed savage in the true sense of the word (*Wah-To-Yah and the Taos Trail;* University of Oklahoma Press, Norman, Oklahoma, 1966).

Status within the family is based on age, sex, relationships by blood, or marriageability. Males are regarded as superior to women in most activities; the elders are respected as repositories of both secular and spiritual wisdom; and people, such as cousins who may be of the same genealogical distance, are frequently divided into "marriageable" and "nonmarriageable" groups, with consequent differences in their interpersonal behaviour. But in all other respects hunting-gathering societies are profoundly egalitarian, especially in intergroup relations.

Outside the family there is no system of coercive authority. Some persons may, by their wisdom, physical ability, and so on, rise to positions of leadership in some particular endeavour, such as a raiding party or a hunt. But these are temporary and variable positions, not posts or offices within a hierarchical structure. Social order is maintained by emphasizing correctness in conduct—etiquette—and ritual and ceremony. Ceremonies bring together the scattered members of the society to celebrate birth, puberty, marriage, and death. Such ceremonies have the effect of minimizing social dangers (or the perception of them) and also of adjusting persons to each other under controlled emotional conditions. (It may very well be true that "the family that prays together, stays together.")

The passage rites at birth, marriage, and death are universal in human society, though puberty celebrations are less common in the modern world, except for such survivals as the Jewish Bar Mitzwa. In most hunting-gathering societies, however, male puberty rituals take up more social time and engage more people than do the other three ritual occasions. They may last as long as a month, food supplies permitting. Almost universally puberty rites include a period of instruction in adult responsibilities, rituals dramatizing the removal of boys from the mothers' care and signallizing the changed social relations between boys and girls of the same generation, and physical ordeals, including scarification or some other mark that will permanently demonstrate the successful passage to manhood.

The Plains Indians

The mounted buffalo hunters of the North American Great Plains, common in popular literature and cowboy movies, constituted a type of nomadic hunting society. But they represented a brief and very special development: an interaction and amalgamation of elements of Indian culture to Spanish horses and the training of them, as well as to metal and guns. The Indians, once mounted, could follow, surround, and kill tremendous numbers of buffalo, where previously the Indians had found the buffalo herds nearly impregnable. So productive was mounted buffalo hunting that tribes of diverse languages and customs were quickly drawn into the Great Plains from all directions. A distinctive, picturesque culture arose among them, reaching its peak about 1800. But from 1850 through the 1870s the tide of white settlers virtually wiped out the buffalo. By that time most of the Indians had been defeated in battle and confined to reservations.

Equestrian Indians can be regarded as a special form of nomadic hunters rather than as a form of pastoralists. Pastoral culture is dominated by the requirements of domesticated livestock and by the relation of herds to pasture. The Plains Indians' nomadism, however, was determined by the habits of the wild buffalo herds. The natural cycle of the buffalo was to concentrate in huge herds in summer and disperse into smaller groups in winter and spring. The Indians accordingly travelled in small camps of a few related families in winter and formed huge encampments in summer and fall for tribal ceremonies and organized cooperative hunts. The summer camps sometimes numbered several thousand people.

The continual intrusion of new groups into the Plains —first Indians, then whites—and the introduction of new weapons constantly altered the balance of power and kept the region in a state of belligerent turmoil. Equestrian bow-and-arrow Indians were superior militarily to those on foot; Indians with guns, of course, were superior to bow-and-arrow Indians; but Indians with both guns and horses—as happened in the Central Plains first—were vastly superior to the others. But the supply of horses and guns and especially ammunition continued to fluctuate wildly as access to sources varied greatly from place to place and time to time.

Nomadism places limitations on property and material technology, and the Plains Indians consequently manufactured no pottery, cloth, or basketry, although leatherwork and beadwork were highly developed. On the other hand, being equestrian, they could carry far more goods than nomadic hunters on foot. Perhaps the most notable thing they carried was the large conical tent (tepee) of decorated buffalo hide.

Socio-political organization was informal, probably because of the fluidity of the population. On the other hand, some tribal cohesion and systems of alliance were required because of the constant raiding. Consequently a large number of pan-tribal associations arose, especially military societies and male age-graded societies.

Religion among the Plains Indians reflected the varying sources of the original religions of the prehorse tribes. Some elements, however, became widespread in the Plains. The folk-hero of a great many myths was the trickster Old-Man-Coyote. There was a widespread concept of Manitou, the pervasive spirit. Most notable was the nearly universal importance attached to the sun—but without the notion of sun as a supreme deity. Ordeals and self-torture and mass ritual self-torture were common Plains religious practices. The Indian tortured himself and fasted in order to suffer hallucinations that would reveal a personal guardian spirit for his protection in the hunt and in battle.

As in other nomadic hunting-gathering societies, principal ceremonies were related to the life cycle, with special prominence given to male puberty rites to instill bravery, endurance, and hunting and raiding skills.

SETTLED HUNTING AND FOOD-GATHERING SOCIETIES

Outstanding examples of the settled hunters and gatherers were the peoples of the North Pacific Coast of North America, roughly from Oregon to southern Alaska. The resources of the sea and inlets and rivers were in astonishing variety, and some, like the salmon during their runs, were so easy to catch that the word "harvesting" seems more appropriate than "fishing" for this activity. In central and northern California there were numerous sedentary Indian groups, such as the Pomo, Wintun, and Yurok. Their basic food was the acorn, which was

ground and stored as flour. Many of the streams had salmon, and the Indians also gathered roots and berries and hunted wild fowl and deer. Other sedentary hunter-gatherer societies are rare and scattered. The most prominent of these are in southwestern New Guinea, as represented by the Asmat. These groups rely on the sago palm, whose starchy pith is easily reduced to flour. Fish, wild birds, and semidomesticated pigs supplement the basic sago.

The effect of food on population and culture

The basic foods of these sedentary peoples had one common characteristic: they were reliable, and they could be stored, much as can the products of agriculture. Salmon were smoke-dried and stored in wooden boxes by the Northwest Coast Indians, and acorn flour obviously could be stored just as can grain flour. Sago flour can also be stored, but it has no season; a palm can be cut at any time the food is required. So abundant and reliable are these resources that such peoples are said to practice a "natural agriculture."

Sedentary life makes possible many improvements in material culture. Houses become larger and more elaborate and are improved over time. The Asmat of New Guinea and the Northwest Coast Indians make huge houses of planks, and are among the best wood-carvers of the primitive world.

Permanent villages and a consistent abundance of food make possible high population densities. The California tribes are estimated to have reached 11 or 12 persons per square mile, as did those of the Northwest Coast. The Asmat of New Guinea have villages ranging up to 2,000 people, which is from 10 to 20 times the size of the average hunting-gathering settlement. Usually such large villages remain politically independent. Intermarriages occur, of course, and some local cohesion is achieved by secret societies and other clublike associations. But such integration is only incidental.

The Northwest Coast Indians

The Northwest Coast Indians elaborated a hierarchical form of organization, or chiefdom. They are the only hunter-gatherers to have done so. Chiefs or nobles occupied positions of high status, inherited in a single descent line by primogeniture. Secondary lines of descent, collateral to the above, were of lesser status. Finally, there were the commoners.

Along with chiefly status went the socio-economic institution of redistribution. Surplus products of family production were passed on to the chief, who in turn gave a large feast (or "potlatch"), during which he distributed gifts to those who needed them. This process of redistribution had the economic function of encouraging specialization and division of labour. The potlatch in late times on the Northwest Coast became famous for its competitiveness. A chief of a lineage or longhouse, for example, would amass as much food and material goods as he could in order to lay on a feast and give presents lavishly in hopes that the guest lineage would be unable to reciprocate on the same scale. Thus one lineage, house, or perhaps village might "defeat" the others.

The Northwest Coast Indian type of chiefdom is primarily social and economic. It can be called political only to the extent that a certain amount of personal authority for decision making may reside in a high social status. This authority can serve a purpose, however. The egalitarian nature of hunting-gathering bands tends toward anarchy, which becomes perilous in populous societies. Quarrels can turn into feuds for lack of a higher authority to settle them.

HORTICULTURAL SOCIETIES

Primitive agriculture is called horticulture by anthropologists rather than farming because it is carried on like simple gardening, supplementary to hunting and gathering. It differs from farming also in its relatively more primitive technology. It is typically practiced in forests, where the loose soil is easily broken up with a simple stick, rather than on grassy plains with heavy sod. Nor do horticulturalists use fertilizer intensively, or crop rotation, terracing, or irrigation. Horticulture is therefore much less productive than agriculture. The villages are small—some no larger than many hunting-gathering settlements—and the overall population density is low compared with farming regions.

Forest horticulturists use fallowing techniques variously called "slash-and-burn," "shifting cultivation," and "swidden cultivation" (a northern English term now widely used by anthropologists). After about two years of cropping a plot is left fallow for some years and allowed to revert to secondary forest or bush. Before resuming cultivation the bush may be cut, left to dry, and then burned. The ashes bestow some fertilization, but the foremost benefit of this procedure is that the plot will be relatively weed free at first.

Since the fallowing periods of the plots are much longer than the planted periods, the swidden horticulturalists must gradually encroach on more distant land. Sometimes this results in semisedentary villages when the newly arable plots finally are so distant that a few horticulturalists must start to build huts near the newer fields, to be joined later by others. Such a land-hungry system, in a region of competing populations, greatly increases the chances of conflict. Population dispersal thus becomes a grave threat in horticultural regions. Land for expansion inevitably must be found at the expense of neighbours or by shortening the fallowing periods—which eventually results in lower production.

Many forest tribes—typical are the horticulturalists of the South American tropical forest—constantly maintain a military posture. Large-scale warfare is not usual (because of the lack of political leadership) but raids, cannibalism, torturing of captives, and other forms of belligerence are.

Horticulturalists have more material goods than most hunter-gatherers, though not more than such societies as the Indians of the Northwest Coast. This suggests that the accumulation of domestic goods is related not so much to the higher productivity of the horticulturalists as to their greater stability of settlement.

The most highly developed of aboriginal slash-and-burn horticulturalists were undoubtedly the Maya of Guatemala and Yucatán, who had a chiefdom or primitive state. But this was most exceptional, for almost all other primitive horticulturalists did not go beyond simple tribes with egalitarian and nearly autonomous communities. Any regional confederation was likely to be only on the basis of intermarriage and clanship. Sometimes an ephemeral sort of near-chiefdom arises, founded on the capabilities of a charismatic leader. In Melanesia, where a well-established form of personal politics thrives, the leader is called Big-Man or Centre-Man.

The Big-Man of Melanesia

The Big-Man in Melanesia is big because he has a following. He begins with his own family and near relatives and friends, who provide goods that he, on behalf of his group, gives away to other groups at a feast on some ceremonial occasion. He and his faction are feasted reciprocally by others at other times. His ability to redistribute on an increasingly lavish scale to larger groups expands his following. He thus amasses what the anthropologist Bronisław Malinowski, in reporting on the Trobriand Islanders, called a "fund of power." With the public esteem gained in this economic contest, the Big-Man is sought out for giving advice, adjudicating quarrels, planning ceremonies, admonishing and conciliating. But this is influence, not the true authority that inheres in a status or office in an established hierarchy. A really big Big-Man may succeed in integrating a region of several villages, but when he loses to a rival or dies, the unity of the region dissolves until some other unusually influential man unites it again.

In many respects the religion of horticultural peoples resembles that of hunting-gathering peoples. Shamans, life-crisis ceremonies—especially puberty rites—totemism (ceremonies for plant or animal species believed to be ancestral to particular human groups like clans or lineages), and the worship of animistic spirits are common in the religion of many kinds of primitive societies. The egalitarian society does not usually practice ancestor worship, as does the hierarchical society. Among horticultural peoples with chiefdoms, the chief's ancestors, in

time, become gods. The most remote ancestors, the founders of the chiefly lineage, are the most important gods; more recent ancestors and those of related but collateral lines have a lesser status. The result is a hierarchy of gods resembling the political hierarchy on earth. Furthermore, the chiefdoms tend to be theocracies, with the hierarchy of priests closely and functionally related to the political hierarchy.

HERDING SOCIETIES

Herding societies are in many respects the direct opposite of forest horticulturalists. They are usually the most nomadic of primitive societies, they occupy arid grasslands rather than rain forests, they have a nearly total commitment to their animals, and their socio-political system is nearly always that of a true hierarchical chiefdom rather than of egalitarian villages and tribal segments.

The military advantages of herders

A society largely committed to herding has military advantages that a settled agricultural society does not have. If military power is important to survival, it will increase the commitment to the herding specialization, mainly because of the advantage conferred by mobility. This increased commitment, however, will result in the gradual loss of certain previously acquired material developments such as weaving, metalworking, pottery, substantial housing and furniture, and, of course, variety in the diet. Wealth is a burden in such societies. Successful nomadic pastoralists normally have some kind of symbiotic relationship to a settled society in order to acquire goods they cannot produce themselves. The symbiosis may be through peaceful trade. But often the military advantage of the pastoralists has led to raiding rather than exchange.

The best known and purest pastoral nomads are found in the enormous arid belt from Morocco to Manchuria, passing through North Africa, Arabia, Iran, Turkistan, Tibet, and Mongolia. They include peoples as diverse as the Arabized North Africans and the Mongol hordes. Other less specialized and successful pastoralists include the Siberian reindeer herders, cattle herders of the grasslands of north-central Africa, and the Hottentot and Herero of southern Africa.

Classic, full pastoralism with its powerful equestrian warriors seems to have developed around 1500 to 1000 BC in inner Asia. This relatively late full-scale pastoralist specialization may have resulted from population pressure. Horticulture mixed with domestication of animals seems to have predominated until even the least cultivable zones were filled. When warfare became endemic in such zones, many groups were forced to become fully nomadic in the arid grasslands. They might have been the losers, pushed out of their homelands, only to discover later the military power that accrued to the pastoral way of life. Thus the victims became victors.

The full pastoralism of inner Asia requires the care of animals—including varying combinations of horses, cattle, camels, sheep, and goats—kept in ecological balance with grazing conditions over an enormous area. In some regions, however, the people may depend on a single animal species. In Arabia it is the camel (although most tribes in Arabia or North Africa also keep horses), and in east-central Africa some peoples specialize exclusively in cattle. Among these there is sometimes found a complete symbiosis between a tribe of herders and an adjacent tribe of horticulturalists to the point that they resemble a single society composed of two specialized castes, the herders occupying the superior position.

There are many part-time herding societies, and many that have merely borrowed herding techniques. The reindeer breeders of the Siberian tundra have learned to apply to reindeer some of the methods of the horsemen farther south, but hunting remains their main subsistencial activity. In the Argentine pampas, Indians learned to domesticate and ride the Spanish horse, which they used to hunt the rhea bird and the wild herds of Spanish cattle. The Navaho and other Indians of the American Southwest have exploited the sheep brought in originally by Spaniards, but mostly as a source of wool for blanket and rug weaving. Llamas and alpacas were domesticated in the South American Andes, but no independent pastoral society ever emerged there.

Fully committed pastoralists manifest a considerable degree of cultural uniformity. In economics, social organization, political order, and even in religion, their livelihood with its functional requirements has ironed out what must have been considerable cultural differentiation among such disparate peoples as Mongols, Arabs, and African Negroes.

Cyclical routes of the pastoralists

The wanderings imposed on pastoralists by the necessities of forage and water tend to be cyclical and to follow long-established routes. The cycles are usually seasonal: to the lowlands in winter, to highlands in summer in temperate zones, to more arid areas in the wet season, to more watered regions in the dry season, and so on. Frequently the seasonal moves are accompanied by cultural and organizational changes. For example, a large group may draw together to pass through hostile territory and disperse later when in their own land; frequently the lush (normally wet) season brings the pastoralists together for ceremonies, trade, and fun, while the dry season requires dispersion and arduous work (as in digging deep wells to water the animals). Anthropologists call such established cyclical movements "transhumance orbits."

Since pastoralists live in so many different environments, and since even the same society varies from season to season and in response to wider spaced drought cycles, it is reasonable to expect great variations in population density. These factors, of course, affect political organization. Nomadic pastoralism lends itself to wide fluctuations in the size of political units, political cohesion, and degree of centralization.

The elementary unit of organization is the patrilineally extended family, frequently an elder patriarch and his sons and their families. In addition, if some degree of primogeniture (*i.e.*, the eldest son inheriting most of the decision-making power for the group) prevails, and if it is extended to include other groups in terms of putative birth order and patrilineal descent, the basis of the pastoral social organization is established. This social structure has been called the "conical clan" (for its hierarchical shape). It is a characteristic social organization of chiefdoms everywhere. Its capacity for waxing and waning, fusion and fission, has obvious advantages, especially when a brief makeshift political ordering of a very large horde is militarily necessary. But the large organizations cannot maintain themselves for very long, since they easily split into their parts.

It has been noted that herders are likely to raid settled villages. But herders frequently raid each other as well. Livestock is wealth and can be exchanged for other forms of wealth—including wives. Stock raiding, like most forms of aggression, has two facets: one seems to be to replenish one's wealth at a stranger's expense; the other is to warn strangers against encroachment. But a raid frequently leads to retaliation and then to counterretaliation, until such raiding societies gradually become hereditary enemies.

The militarism of herding societies has played a major role in history. As wealthy agricultural civilizations developed in the Fertile Crescent of the ancient Near East, in the Indus River Valley, and at the middle bend of the Yellow River in China, they became easy prey for nomads. Indeed, it is likely that urbanization was stimulated for defensive reasons because of the dangers posed by nomads. These dangers may also have stimulated the formation of legal and governmental institutions in sedentary societies threatened by the pastoral raiders.

Trend to ancestor worship

To the extent that pastoral nomadic societies achieve wealth and success in herding and in war, they tend to solidify and extend their chiefdom structure. They also add to their religious organization a hierarchical principle together with the content known as ancestor worship. Much of the mythology by which a primitive people ex-

plains itself and its customs comes in this way to have an ingredient familiar to readers of the Old Testament—the lengthy story of who begat whom and in what order.

Increased dependence on herds, particularly dependence on one particular species, such as cattle, horses, or camels, is reflected in much of the ideological and ritual content of religion. Sometimes the significance of herding leads not only to the glorification of herds and herding but even to a religious taboo against planting. Some Mongols, so quintessentially pastoral, believe that plowing and planting defile the earth spirit. Among the Nuer, as among other African cattle herders, horticulture may be practiced in time of need, but it is considered degrading toil whereas herding is a very prideful occupation. The ethnologist of the Nuer, E.E. Evans-Pritchard, wrote:

> They are always talking about their beasts. I used sometimes to despair that I never discussed anything with the young men but livestock and girls, and even the subject of girls led inevitably to that of cattle . . . . (From E.E. Evans-Pritchard, *The Nuer;* Oxford University Press, London, 1940.)

Peasantry

The one remaining category of non-urban society is the peasant. Peasants are not nomadic but sedentary (thus distinguishable from both hunting-gathering societies and pastoralists); they are not horticultural tribal societies but more intensively and fully agricultural; and neither are they urban, like populations who lived in the centres of the classic civilizations. Except for a few other characteristics, however, writers on peasantry have not agreed on a precise definition. Everyone notes that peasant communities tend to be small, and most also agree that they tend to be tradition bound and resistant to change, also, that they are "part-societies" or "part-cultures" in relation to some larger civilization, colony, urban centre, state, or elite class.

PEASANT COMMUNITIES

The best way to proceed, then, is to consider, however briefly, all the kinds of societies that have been conventionally called peasant in their full historical, geographical, and cultural diversity. One way to decide the question of peasant society's structural relationship to a larger society is to review a wide variety of examples.

The first agricultural revolution

By about 4000 BC in Mesopotamia and about 1000 BC in Middle America some agricultural villages had begun to become cities. This was part of the evolution of intensive agriculturalists that was to lead gradually to the empires of classical civilization. One major significance of the classical empire was that it could protect, and thus politically incorporate, at least in some areas, scattered villages of simple cultivators; it could tax them (exact tribute in some form), even conscript them, and, thus, make a census of them. The invention of writing may well have gotten its start in such record keeping. At least, it was developed greatly in urban centres.

Outlying agricultural villages could not have survived without internal pacification (in the law-and-order sense) and without some kind of frontier militia. The rulers of the urban centre would soon find that the price of the protection they extended to the peasants could easily be increased to the point of virtual expropriation. As a result, the peasants were reduced to the barest subsistence level. But even without exploitation, the peasant acquires a lowly status. Dependence often creates servility in manner, rural isolation tends to make the peasant ignorant and no match for the sophisticated urbanite, and poverty keeps him dependent generation after generation. It is not simply that the peasant is somehow "exploited" (a difficult point to determine in most cases) but that his village is normally small, poor, ignorant, and generally backward compared with the urban centre or feudal manor. Horticultural tribes, of course, live in even smaller villages and are poorer and more ignorant, but they in some sense possess a complete culture, while a peasant villager has only a part-culture. What seems to be universal is the peasant's low status, with its concomitant ascription of poverty and ignorance, in contrast to other parts of the same culture.

Peasants distinguished from farmers

The pre-industrial urban centre with its statelike protection and intervention in the lives of peasants made it possible to extend agricultural domains, usually as peasant holdings, into pastoral or other "wild" territory. An enormous part of the world's population became peasants as primitive peoples and nomads were dominated and displaced or transformed. Today, with the rise of modern science and industry, peasants are being rapidly displaced in all of the so-called developing or modernizing parts of the world.

To understand what peasants are it is helpful to contrast them with what they are not—farmers in an industrialized world. The modern farmer, operating on a cost-accounting basis, has little in common with the peasant family or villager. Industrialization increased the need for agricultural produce, which stimulated production for the market. The application of science—with such aids as artificial fertilizers, pesticides, farm machinery, and hybrid grains—enabled large-scale production with relatively little labour.

Peasants could not simply turn into farmers because they lacked the capital. Some became hired workers in the fields. But most migrated to the cities or overseas to the Americas. In some parts of Great Britain, central France, and the Low Countries, however, peasant villages have managed to survive by maintaining a low standard of living and long working hours, and often by developing some special handicraft for sale as folk art.

Along with the rise of commercialism and industrialization in modern times, certain changes of a political nature have affected the rural areas. Strong national states have replaced old feudalistic dependencies, many nations have even expropriated the holdings of large absentee landowners, and civil rights and legal protection have been extended to peasants. Everyone has become, to some degree, a citizen of a modern state. Yet peasants are still handicapped by their relative isolation and poverty, and they tend to be the last group (other than primitives) to become full participants in modern national cultures.

The distribution of peasant societies

Western Europe (including Great Britain) can be treated as an entity, since its feudal epoch, rise of national states, and finally its full and earliest participation in the commercial and industrial revolutions sets it apart as a region of distinctive similarities among its diverse parts and certain important dissimilarities from other large world regions.

The industrialization of eastern Europe and western Asia occurred later than that of western Europe, and its effect on the demography of peasant villages was much less. The Soviet Union's collectivization of agriculture in the late 1920s had the greatest influence on the lives of the peasants in that region (and in the Soviet bloc of eastern Europe after World War II). The ancient classical peasantry of Russia and adjacent regions thus merits a separate classification.

Japan is unique among modern nations in the character of its peasantry and the changes from its original feudal-like system to rapid industrialization and transformation of its traditional countryside. Certainly Japan cannot be simply lumped with China and perhaps some smaller entities as "the Far East."

Similarly, China is China. China was a "sea of peasants" over such an enormous span of time that the peasants' was more than a "part-culture" in the ancient Chinese civilization. The peasants themselves became distinctive and "classic."

Southeast Asia, especially Indonesia, qualifies as a separate peasant region not only because of its ancient cultural distinctiveness but also because of its epochs of Islāmic and European colonialism, which set it off from Japan, especially, and from China.

India experienced various foreign invasions and finally British colonialism, itself an administration of a distinctive sort. Next to China, India has the largest peasantry and the most problems in modernizing its agriculture.

The complexity resulting from Hindu, Islāmic, and British policies has resulted in a mosaic of local practices in land tenure, and the residue of the caste system causes special difficulty in modernizing most rural villages.

Latin America, together with the Caribbean islands, is the only region in the Western Hemisphere to contain sizable numbers of peasants. There are no peasants in the United States or Canada, except, perhaps, for a few widely scattered villages of French Canadians, Appalachian mountaineers, southern sharecroppers (rapidly disappearing because of mechanization), and perhaps Pueblo and Navaho Indians. These all seem like peasants in their poverty, the backwardness of their isolated village life, and their dependent status; but these groups are not numerous and are disappearing rapidly, and it has been argued that their members should not be considered peasants at all.

Latin America's distinctiveness in this respect stems from two historical factors. (1) The Spanish and Portuguese colonized the area so long ago that the inhabitants have had time to develop a stable, though complex, amalgam of Iberian and Indian culture. (2) During the pre-Columbian epoch in Middle America, in parts of the circum-Caribbean region, and in the Andean highlands, the aboriginal Indians achieved a form of state-empire with a stratified society of the same general order that existed in early classical civilizations of the Old World. Thus a class of native peasants was already in existence in several large regions when the Spaniards arrived. In other areas —Amazonia and Argentina, for example—quite different developments resulted from the nonpeasant Indian societies there: tribal horticulturalists in Amazonia and nomadic hunters in Argentina, neither of which became dependent on the new colonial regimes.

In Africa, scattered peasant communities occur in the upland areas of the Mediterranean shore, but they are closely related in origin to Arabic culture and European colonialism and are not numerous. Sub-Saharan Africa, despite the colonialism of many of its areas, is distinctively African. Nevertheless, there is wide disagreement on how to classify its rural inhabitants, who, for the most part, make up the basic citizenry of its new nations and are not so much members of a dependent part-culture.

TYPES OF PEASANT SOCIETIES

The community of self-serving households. Though peasants are usually thought of as living in small, close-knit communities huddled against outside danger, they sometimes are so well-protected in mountainous or insular isolation that they feel secure enough to live more independently. In such circumstances, they dwell in scattered households in close proximity to the land they cultivate. They require a market centre of some sort where they may exchange goods and services. The Irish countryman, the isolated *fermier* of the French Massif Central, the Scottish crofter, the Paraguayan *campesino,* and the Brazilian *caboclo* are examples of such independent peasants. Occasionally the same people will be found living in close communities and also in scattered, more self-sufficient households. In the state of Michoacán in Mexico, for example, some of the tightest and closest knit communities to be found anywhere on earth ring Lake Pátzcuaro, in immediate proximity to the large modern market town and tourist centre of Pátzcuaro. These are the fishing-, agricultural-, and handicraft-specialist villages of the Tarascan Indians. But many more thousands of Tarascans also live scattered in the adjacent mountains, making only infrequent visits to the market centres.

The village with internal specialization and exchange. In certain times and places peasant villages have developed considerable self-sufficiency by creating part-time specialists and even full-time professional occupations. Such a development, however, presupposes an intensive agriculture in support of a fairly large population, in order that the specialists may be kept fully occupied. The best examples of this kind of village are found in India, in the European medieval manor, and in some Latin American haciendas.

Hindu villages

The most distinctive, as well as the most clear-cut, specialization occurs in Hindu India, where a typical village may contain as many as 2,500 people. The professional specialties are pottery manufacturing, stoneworking, barbering, trading, weaving, laundering, and herding. All of these occupations are carried on by separate castes, to which should be added the "twice-born" caste, the Brahman, or wise-man priest, though this is more of a status than an occupation.

The specialized services of the various castes often are rendered without any immediate payment or return service. The occupational castes all have an obligation to provide their services. The full-time peasant agriculturalist, for example, expects a new plow or hoe from the carpenter, a pot from the potter, haircuts from the barber, and so on. After the semi-annual harvest, the peasant distributes appropriate shares of produce to those who have served him.

The caste system of occupations largely determines the status of individuals, but there are ways to attain higher status by acquiring wealth or political office. A wealthy landowner of low caste will continue to observe all the traditional attitudes of deference to those of higher caste; yet his opinion may be important and his power considerable in other than direct interpersonal dealings. And, of course, high and low status may be earned within a given caste depending on individual skill and personality.

Land ownership and tenure patterns are variable and complex. There are large owners who hire labour by wage or by shares. The majority are family-owners and workers of small plots, but large numbers of agricultural workers are landless, working only for others. Many families own some land and at the same time work other plots by shares or for wages. The usual peasant holding is worked jointly by a father and his sons. When the father dies, the land, stock, and implements are distributed equally among the sons. This practice is the major cause of the small size of the individual peasant holdings.

Many Indian peasant villages are exogamous (marrying outside), which results in ties among several villages as a consequence of giving and receiving wives. In such cases, every person participates in a social network outside his village to a greater extent than he associates with persons of other castes within his own village. These regional relationships are the means by which a common culture is diffused over a wide area. Hindu peasant villages are less alike the farther they are from each other, yet vast areas of rural India are remarkably homogeneous in culture.

The medieval manor

The European feudal estate also tended toward economic self-sufficiency in its local specialized occupations but was unlike the Hindu peasant village in several respects. For one thing, there were no castes. The aristocrats considered both their own and the peasant class to be permanent, God-given arrangments of hereditary status. Thus, to the extent that membership was in fact static, these classes were like Hindu castes (which have frequently been defined as "frozen classes"). But the other occupational classes of medieval times were not so castelike, although a tendency existed for son to succeed father. The occupational guilds resembled, to a certain extent, the wide geographical relationships of the Hindu castes. At the time of greatest stability in the European system, the social and political differences from Hindu practice were perhaps largely those of degree.

Other differences were enormous. India, no one needs to be told, is overcrowded. Medieval Europe, between the 11th and 15th centuries, was almost a wasteland by comparison. In fact, the existence of vast tracts of forest lands gave urgency to the problem of law and order; large groups of outlaws and predators could easily hide out. The essentials of the feudal system were master-client relationships: between kings and nobility, and between the nobility. The superior individual gave protection to his clients, who in turn provided crops or services (especially labour and military duty).

The institution most typical of medieval society was the local *seigneurie,* which may be defined as an estate com-

prising a group of people subjected to a single master. The land in such a system was of two kinds. One was the large home-farm, cultivated under the immediate direction of the *seigneur* or his supervisors. The other part of the *seigneurie* consisted of various small to middle-sized holdings whose tenants occupied and cultivated them freely under the *seigneur's* protection in return for helping him in the cultivation of his demesne.

There were essentially three kinds of labourers: (1) the tenure holders, who owed regular services to the demesne; (2) wage labourers, normally paid in kind, unless they were imported labour to help out at some crisis like the grape harvest; and (3) workers housed and provisioned by the demesne. These workers are called "prebendal" in English (French *provendiers*) because they were provisioned and housed at the master's expense. The only difference between a prebendal worker and a slave was the freedom of the prebendal worker to leave if he was dissatisfied.

The tenure holder, or peasant, owed the *seigneur* two basic obligations, rent and services. Rents were highly variable, but services were usually still the greater burden. The basic services were agricultural labour on the demesne land, military duty, and craft work. Agricultural labour, for example, might be calculated as three mandays a week per tenure holding. Since a family on a holding might be quite large, the three days could be divided among several men. Military duty would be highly variable because it would be a simple response to emergency —ordinarily an "all-hands" response.

Craft work was divided among peasants who had some skill passed on from father to son, especially metalworking. Spinning and weaving, wine making, carpentry, and sometimes milling and baking were duties divided among certain, but not all, peasant families. Probably most craftsmen worked at these tasks only part of the time in addition to the basic form of work.

The crafted products did not pass from peasant to peasant or between different specialists but were usually paid to the *seigneur* who reassigned them to others. This kind of indirect passage of goods from producer to centre and thence to ultimate consumer is the essence of the redistributional system described earlier as characteristic of chiefdoms. It is a way by which a power-holder can muster goods and serve his people at the same time. In political as well as economic structure, the resemblance of the *seigneurie* to a primitive chiefdom is remarkable.

The *seigneur* was thus not simply a landowner or an exploiter of labour. He was a leader of men whose political-military status was highly direct and personal. He had command over his tenants, and the system would not have worked unless his subjects generally believed in and accepted him. His protective function gave rise to the seignorial court, which was the recognized place for a hearing of pleas and complaints. All in all, the seigneur served his people in many necessary ways, and they served him in others.

Naturally the asymmetrical power relationship between seigneur and peasant sometimes resulted in attempts by seigneurs to multiply the services or benefits due them. On the other hand, the peasants, if numerous enough, often found ways to resist. But the power of the seigneurs presumably lay originally in their ability to allow or prevent the occupation by peasants of land under their military power. Similarly, peasants at certain early and insecure epochs might want the security of hereditary tenure; at other more secure and prosperous times, they might want freedom to leave.

By the 12th and 13th centuries in France every tenant was either free or a serf. The norm among free tenants was to be bound to the *seigneur* only because of their occupation of the land. If the tenant left, all obligations both ways were broken. On the other hand the serf was not free to leave the land. Otherwise mutual dues and obligations were the same.

The hacienda

Another form of agricultural self-sufficiency is exemplified by the hacienda. In the early colonial period of Latin America the hacienda combined the Iberian and American Indian systems of land use. Pre-Columbian Indians in large areas of Latin America (from Chile north through the Andes, and in Middle America) were densely settled on communal village holdings under the suzerainty of absentee aristocratic Indians. Other areas of Latin America were inhabited by more primitive tribes of slash-and-burn horticulturalists and nomadic hunter-gatherers. During colonial times in the areas of densely settled Indian population, the leading Spaniards were granted political control over designated villages. They were allowed to tax the Indian families and in return were supposed to protect them and educate them in the Catholic faith. Sometimes Spaniards were rewarded by the crown with enormous tracts of land, *latifundios,* usually in areas of lesser population where large-scale herding would be the primary economic resource. Indian labour was also exploited in gold and silver mining and in workshops (*obrajes*).

The economy based on the exploitation of unskilled Indian labour was eventually disrupted by disease. Indians had no immunity to several commonplace European afflictions such as smallpox, typhoid fever, measles, and malaria. Numerous disastrous epidemics occurred, and by about 1600 both Spain and its richest New World possessions were in rapid economic decline.

Meanwhile, a new form of rural estate came into being as the economy of town and city, workshops, mines, and commerce was depressed. A large, privately owned estate could withstand monetary and commercial crises by becoming increasingly self-sufficient. The estate was manned by impoverished Indian workers who needed security and protection. The workers were usually paid in kind, enough for bare subsistence and given credit (against the promise of future labour) for the purchase of other necessities. This debt peonage was the foundation of a permanent labour supply, resembling the serfdom of medieval Europe.

The hacienda had a permanent group of peons settled on its lands, allowed to farm small plots for themselves. There were also house servants, some of whom might reside in the master's home. Other Indians might be residents of neighbouring villages but dependent on the hacienda for protection and often for grazing rights on fallowed rangeland claimed by the hacienda. A hacienda with numerous dependent villages on its periphery could muster a large labour force when needed and not employ it when not needed. The permanent debt peons, however, were more closely bound up in the everyday life of the hacienda. Like the European serf, the peon in difficult times probably welcomed the security of such an arrangement.

The hacienda probably was never completely self-sufficient, but it could take care of its own people in many ways. Large haciendas, some with thousands of peons, could afford numerous specialists, such as metalworkers and leatherworkers, weavers, bakers, masons, carpenters, and sometimes even a resident priest. There might also be a jail and a whipping post. And just as in European seignorial law, the master adjudicated disputes and meted out punishment. The economy of the specialized crafts resembled the European redistributional system in so far as the planning, commissioning, and delivery of all benefits were centralized under the hacienda master and his agents.

Although debt bondage no longer exists in Latin America, the tenant worker on the remaining large haciendas in some of the Andean areas seems as closely bound to the soil as peasants ever were. The Chilean tenant is legally free to move as he pleases, but he cannot, in fact, usually do so. He works his ancestral land, which he understands belongs to the hacienda, whose owner he has been conditioned all his life to regard as his master and protector. Were the worker and his family to leave, the other haciendas would not accept him. And since there is no vacant fertile land, as in Paraguay and parts of Brazil and Argentina, he could not become a squatter. Most peasants fear the city, which is already filled with the unemployed younger sons of peasants.

In Mexico, it was not until well into the 20th century that the hacienda system began to yield to modernism and more liberal laws, and the hacienda became increasingly commercialized. But earlier the peasant could not improve his position, legally or economically. By the end of the Díaz regime in 1910, the concentration of ownership of land in the hands of a few *hacendados* was greater than in any other Latin American country. But the payment for agricultural labour had not risen appreciably since 1792. Over the same period the price of maize had increased 179 percent and that of beans 565 percent.

The closed regional market system. A kind of regional self-sufficiency may be seen among peasants in the Middle American highlands and the Andes, in parts of Indonesia, and in West Africa. These are nearly self-sufficient regions embracing peasant hamlets and villages that trade with each other, usually on a periodic market day or fair. These villages are typically cohesive and tend to be self-governing through ritual and religion. The relations of peasants with the outside world are usually mediated by the community (or its officials), except in the peasant market, where frequently some kind of middleman or representative of a store sells items not produced in the region itself.

The highland Indians of Mexico and Guatemala

The highland Indian communities, especially in Mexico and Guatemala, are quintessentially of this type. The inhabitants of the region are Indians (although there is a heavy overlap of Old Spanish custom, dress, and a folkish variant of Spanish Catholicism). They all speak the same Indian regional dialect. They occupy such a distinctly inferior and helpless relationship to the outer society that intermediaries of some sort are needed, and as a consequence of the felt inferiority they act toward outsiders in an extremely withdrawn manner. This withdrawal trait of the Indian peasantry has been appropriately labelled the "*encogido* syndrome," meaning a nearly utter lack of self-confidence.

The Indian communities which have legal ejidos (communal holdings) as well as small family properties, are not usually subject to outside landowners. Thus the *encogido* syndrome derives not from any inferior position of the peasantry to a resident owner (as in the hacienda system) but from the simple fact of ethnic stratification. These Indians feel themselves in an inferior position to everyone else in the whole outside world, that is, inferior to everybody except others designated as Indians. And they are usually also inferior in visible ways: extremely poor, uneducated, and ignorant of the manners and customs of the nation's urban citizenry. They have responded by withdrawing into their own community, which only serves to continue or to exacerbate the inferior condition, since much of it is caused simply by economic and social isolation from the main currents of national life.

'Within the community of Indians extreme egalitarianism prevails. Prestige is won only in community service, which means giving more than receiving—very much like the general reciprocity of primitive society. Giving takes the form of subsidizing one of the traditional fiestas of the community, an obligation that may be expensive in food, liquor, candles, and fireworks. Families work for years to get up the capital to sponsor a magnificent festival, at which they not only spend their savings but go into considerable debt. Lavishness in money and labour equals love (of the community); the poverty of an individual family may be accompanied by the highest prestige. This bears a resemblance to the Big-Man system of Melanesia, where prestige is also linked with lavish giving.

The economic egalitarianism of the village does not rule the periodic regional market. There the family, or its agent, tries to maximize its gains. Outsiders who see more of the marketing behaviour than of the more pervasive social economy within the individual villages are often misled by the ferocious haggling and cheating. Many regional markets are held in a smallish city of the national, non-Indian sort. The economically important market is the one in which the Indians exchange their own products with each other, yet it may be overshadowed by the market in which Indian handicrafts are sold to outsiders. Town stores may display their own wares in stalls at the market, and the carrier-middlemen may have brought materials from a considerable distance to sell where such items are rare and will bring a better price. And the same middleman (often simply the owner of a pickup truck) may buy materials or handicrafts that are abundant at one market for sale in another. In Mexico the weekly Tarascan market in Pátzcuaro, or the Otomí at Ixmiquilpan, or in Guatemala the Mayan markets at Chichicastenango or Panajachel, are large and complex, incorporating all of the above elements.

In these highland Indian communities the emphasis is on self-sufficiency, at least for the regional economy, instead of on the personal dependency and exploitation that characterize the hacienda and seignorial communities. The Indian peasant commune is of interest because the tradition-bound folk culture, so dominated by the sacred, so unindividualized, and so homogeneous, contrasts strikingly with modern urban societies.

Close-knit peasant communities do change, of course; but it is their resistance to change that commands attention in modern times, when virtually all other institutions are committed to rapid change. Because the peasant community is a rigid structure, the individuals who want change and have the means or capacity to carry it out simply leave the village to find their place in some mixed community. In some peasant communities this is not possible. Thus in densely settled central Java opportunities for the surplus peasant population to adapt to other modes of life were few. As population increased, putting stress on the traditional network of communal villages, the community tried to fit more people into the traditional system. The social patterns grew more elaborate but remained traditional; cooperative labour and tenancy institutions became more intricate but otherwise did not change. Just as in the Latin American Indian commune, egalitarianism continued on a basis of what has been called shared poverty.

In West Africa peasant society is based on full-time intensive agriculture, with a considerable amount of craft specialization and a large amount of trade carried on in great markets. In ancient times the more populated areas were organized as chiefdoms and primitive states, which exacted a tribute or tax from the agriculturalists in exchange for their protection and regulation. Intensive cultivators produce mainly for their own consumption, but with a frequent surplus or specialized handicraft to trade in the market. In modern times, they may produce surpluses to trade with a middleman representing an exporter. African villagers traditionally have been politically dependent on some hierarchical chieftains or patrimonial retainers. But they do not, like peasants elsewhere, feel themselves to be different from or inferior to any other class or culture. Nor are they so regarded by the urbanites with whom they come in contact.

General characteristics of peasant societies

One characteristic of peasant economy is that the production unit is normally the family. But this does not mean that families are all the same size. Mainly, technical and economic requirements tend to govern the size of the family, which ranges from large three-generational extended families down to the nuclear unit of one set of parents and their unmarried children. Inheritance patterns tend to reflect the requirements of the agricultural operation. Whether the land is split equally among the heirs or passed on as a single unit (commonly through the eldest son) will depend on whether farming requires large holdings or whether a small, intensively farmed area is sufficient. In some historical instances, the ecological determinant of the size of holdings has been contravened by ideology or law. For example, the Code Napoléon required that agricultural holdings be inherited equally, with the result that when the fragments of land were obviously becoming too small the French peasants responded with one of the most drastic reductions in the birth rate in all of recorded demographic history.

Some villages, notably among Latin American Indians,

are quite communalized. Individual families work plots of land, but it is the community as a whole that makes the important decisions. Other peasant societies with more independent households find numerous occasions for cooperation and labour exchanges among families. One widespread means of establishing a network of reciprocal obligations and trust within a peasant community is through ritualized ties of fictive kinship, such as the godparent-hood common throughout most of peasant Europe and Latin America (in Spanish it is coparenthood—*compadrazgo*). Other forms of fictive kinship are the familiar blood brotherhood of Balkan Europe, the *mit* of Nepal, and the *oyabun-kobun* of rural Japan.

THE POLITICAL ROLE OF THE PEASANTRY

Peasants have sometimes exploded with pent-up mass violence. Such revolts occurred in late medieval times in western Europe, feudal Japan, and in the 19th century in tsarist Russia and Manchu China, to cite only the more well-known ones. Peasants in modern times have also fought valiantly and stubbornly in China, Algeria, Cuba, Mexico, Vietnam, and elsewhere. These latter movements were not all as independent and spontaneous as the earlier ones, however; they were part of non-peasant, national political movements, with leadership, ideology, and revolutionary objectives far from the peasants' immediate aims or comprehension.

Marx and Engels considered the peasantry to be a passive if not reactionary force in revolutionary situations. Marx thought of peasant life as "rural idiocy," which he contrasted with the revolutionary education undergone by the urban proletariat. He believed that the continued growth of capitalism would destroy the peasantry and increase the numbers of wage labourers. Revolutionary tactics should therefore allow the peasantry to continue its inevitable decline and demise.

Revolts aimed at restoring past

Marx may have been influenced by the history of the peasant revolts of 1524–25 in Germany, which his colleague Engels wrote about in *The Peasant War in Germany*. The revolts were caused by a sharp deterioration in both the economic and legal position of the peasants. The commercial revolution of the late 15th century had impoverished the landholding nobles, who in turn expropriated more and more of the peasants' produce. At about the same time, the introduction of Roman law sped the decline of the peasants into serfdom. The brutal repression of the peasants after their uprising prolonged serfdom in many parts of Germany until the French Revolution. The significance of the peasant rising to Marxists and other theorists of revolution, however, was that it represented an attempt to restore deteriorating conditions to their earlier state, not to make any basic change in the structure of society.

In tsarist Russia, from around 1870 until the Revolution of 1917, the powerful and numerous Populists (Narodniks) held that a Socialist revolution could come about in the countryside patterned after the historic model of the *mir,* a communal village of medieval times. Lenin and other Marxists opposed this idea vigorously, arguing that the industrial proletariat was the only class that could be relied upon.

Peasants and guerrillas

The attitude of the Chinese Communists was completely different. Much of the population of rural China had been engaged for hundreds of years in intensive agriculture, and at the same time it had developed craft specialties and a great deal of commerce and wage labour as well. In China there was little of the stagnant rural idiocy of Marx's peasantry. The necessity for working among peasants led the Chinese Communists to argue that in an area like China the only way to continue revolutionary struggle was through guerrilla warfare on a rural base.

The success of Mao's tactics in China was paralleled in some areas of Mexico during its revolution in 1910. The peasants led by Emiliano Zapata were united in a struggle to regain the lands that had been gradually taken from them by the *hacendados.* Such aims are not different from the peasant struggles of late medieval Europe, but the point is that the Zapatistas waged successful guerrilla warfare that helped to overthrow a 20th-century political state.

The Castro-led revolution in Cuba provides a spectacular example of the success of rural-based guerrilla warfare, and was probably as powerfully influenced by earlier Mexican revolts as by Chinese-Marxist ideology. Similarly, the success of peasant-based warfare in Vietnam against Japanese, French, and American troops is familiar. And the French could not hold their colony in Algeria with regular troops against the guerrillas.

These modern rural revolts differ from the earlier revolts in that rural peoples have become involved in modern nation-states, which themselves are undergoing vast changes. The rural cultivator no longer is simply a member of a small isolated community, protected or exploited by landowners or nobility. He is more often a rural proletarian than a peasant, notably so in the case of Cuba. And his allies, friends, and enemies are numerous.

All of the modern revolutionary nations share one historical fact in common: they have all been colonialized in one way or another by powerful industrialized states. In consequence, various anticolonial (nationalistic or patriotic) elements in the population aligned themselves with the rural people. Modern agriculturalists in those former colonial nations have become, in one way or another, much more of a citizenry than ever was visualized in the conceptions of peasantdom as a tradition-bound, isolated folk society.

BIBLIOGRAPHY. R.B. LEE *et al.* (eds.), *Man the Hunter* (1968), is an intensive survey by professional anthropologists of nomadic hunting-gathering society; L.H. GARRARD, *Wah-To-Yah and the Taos Trail* (1938; new ed. 1955, reprinted 1966), is a classic account of Plains Indian life; F.R. SECOY, *Changing Military Patterns on the Great Plains* (1953), is a superb historical analysis of the rise of Plains Indian nomadic equestrian societies. The best general account of the sedentary fishing societies is P. DRUCKER, *Cultures of the North Pacific Coast* (1965); and an interesting analysis of the forms of economic exchange among those societies is by H. CODERE, *Fighting with Property* (1951). Pastoral societies are depicted in three different regions by O. LATTIMORE, *Inner Asian Frontiers of China* (1951); L. KRADER, "Ecology of Central Asian Pastoralism," *Swest. J. Anthrop.*, 11:301–326 (Winter 1955); and for Africa, E.E. EVANS-PRITCHARD, *The Nuer* (1940, reprinted 1968). A grand survey of the consequences of the domestication of plants and animals in early times is found in R.J. BRAIDWOOD and G.R. WILLEY (eds.), *Courses Toward Urban Life* (1962); and many of the modern primitive horticulturalists are described analytically by M.D. SAHLINS, *Tribesmen* (1968). The rise of peasantry in relation to urban centres or other outside powers is analyzed neatly in relation to other evolutionary levels by W. GOLDSCHMIDT in *Man's Way: A Preface to the Understanding of Human Society* (1959). Important general works on peasants are R. REDFIELD, *Peasant Society and Culture* (1956); and a compilation by J.M. POTTER, M.N. DIAZ, and G.M. FOSTER (eds.), *Peasant Society: A Reader* (1967). Peasantry in medieval Europe is described and analyzed by M.L.B. BLOCH, *Les Caractères originaux de l'histoire rurale française* (1966; Eng. trans. *French Rural History,* 1966); J.H. CLAPHAM and E. POWER (eds.), *The Cambridge Economic History of Europe from the Decline of the Roman Empire* (1941); G.C. HOMANS, *English Villagers of the Thirteenth Century* (1941); and H. PIRENNE, *Histoire économique de l'occident médiéval* (1933; Eng. trans. *Economic and Social History of Medieval Europe,* 1936). Studies of peasantries in different world areas have sometimes led to important theoretical arguments, perhaps most notably in the Indonesia area, where J.H. BOEKE proposed the important concept of the "dual society" in his *Economics and Economic Policy in Dual Societies* (1953); which was followed by J.S. FURNIVALL, *Netherlands India: A Study of Plural Economy* (1939); and C. GEERTZ, *Agricultural Involution: The Process of Ecological Change in Indonesia* (1963). In Latin America R. REDFIELD innovated comparative field studies involving a conception of historical stages from a "folk" primitive society through intermediate peasant-like villages toward an urban pole, described in *The Folk Culture of Yucatan* (1941). Other important analyses of Latin American Indian peasantries are: C.J. ERASMUS, "Community Development and the Encogido Syndrome," *Human Organization,* 27:65–74 (Spring 1968); G.M. FOSTER, *Tzintzuntzan: Mexican Peasants in a Changing World* (1967); G.M. MCBRIDE, *Chile: Land and Society* (1936); and C. WAGLEY, *The*

Latin American Tradition (1968). C.M. ARENSBERG initiated anthropological study of Irish peasantry in his classic *The Irish Countryman* (1937). In the Far East, a classic was written by FEI HSIAO-TUNG, *Peasant Life in China* (1939); and S.C. DUBE wrote interestingly on modern changes in *India's Changing Villages* (1958); while T.C. SMITH traces peasant history in his *Agrarian Origins of Modern Japan* (1966). Modern political problems are discussed in C.C. BRINTON, *The Anatomy of Revolution* (1965); in D. BROKENSHA and M. PEARSALL (eds.), *The Anthropology of Development in Sub-Saharan Africa* (1969); and by G. DALTON, *Economic Anthropology and Development* (1970).

(E.R.Se.)

Human-Factors Engineering

The term human-factors engineering is used to designate equally a body of knowledge, a process, and a profession. As a body of knowledge, human-factors engineering is a collection of data and principles about human characteristics, capabilities, and limitations in relation to machines, jobs, and environments. As a process, it refers to the conception of designing machines, machine systems, work methods, and environments to take into account the safety, comfort, and productiveness of human users and operators. As a profession, human-factors engineering includes a range of scientists and engineers from several disciplines that are concerned with man at work.

The terms human-factors engineering and human engineering are used interchangeably on the North American continent. In the U.S.S.R. the term engineering psychology is preferred; in Europe, Japan, and most of the rest of the world the prevalent term is ergonomics, a word made up of the Greek words, *ergon,* meaning "work," and *nomos,* meaning "law." Despite minor differences in emphasis, the terms human-factors engineering and ergonomics may be considered synonymous. Human factors and human engineering were used in the 1920s and 1930s to refer to problems of human relations in industry, an older connotation that has gradually dropped out of use. Some small specialized groups prefer such labels as bioastronautics, biodynamics, bioengineering, manned-systems technology, and many more; these represent special emphases whose differences are much smaller than the similarities in their aims and goals. A 1970 survey of the field revealed that human-factors engineering is the term most widely used and accepted.

The term ergonomics

The data and principles of human-factors engineering are concerned with human performance, behaviour, and training in man-machine systems; the design and development of man-machine systems; and systems-related biological or medical research. Because of its broad scope, human-factors engineering draws upon parts of such social or physiological sciences as anatomy, anthropometry, applied physiology, environmental medicine, psychology, sociology, and toxicology, as well as parts of engineering, industrial design, and operations research.

The man-machine model. Human-factors engineers regard man as an element in systems, and a man-machine model is the usual way of representing that relationship. The simplest model, a man-machine unit, consists of an individual operator working with a single machine. In any machine system, the human operator first has to sense what is referred to as a machine display, a signal that tells him something about the condition or the functioning of the machine. A display may be the position of a pointer on a dial, a light flashing on a control panel, the readout of a digital computer, the sound of a warning buzzer, or a spoken command issuing from a loudspeaker.

Having sensed the display, the man interprets it, perhaps performs some computation, and reaches a decision. In so doing, he may use a number of human abilities, including the ability to remember and to compare what he perceives with past experiences, to coordinate what he perceives with strategies he may have formed in the past, and to extrapolate from his perceptions and past experiences to the solution of novel problems. Psychologists commonly refer to these activities as higher mental functions; human-factors engineers generally refer to them as information processing.

Having reached a decision, the human operator normally takes some action. This action is usually exercised on some kind of a control—a pushbutton, lever, crank, pedal, switch, or handle. Man's action upon one or more of these controls exerts an influence on the machine and on its output, which in turn changes the displays, so that the cycle is continuously repeated.

A man-machine system does not exist in isolation; it exists in an environment of some sort. Since the nature of this environment influences man's efficiency and performance, the human-factors engineer must be concerned with such environmental factors as temperature, humidity, noise, illumination, acceleration, vibration, and noxious gases and contaminants.

Driving an automobile is a familiar example of a simple man-machine system. In driving, the operator receives inputs from outside the vehicle (from traffic, obstructions, and signals) and from displays inside the vehicle (from the speedometer, fuel indicator, and temperature gauge). The driver continually evaluates this information, decides on courses of action, and translates those decisions into actions upon the vehicle's controls—principally the accelerator, steering wheel, and brake. Finally, the driver is influenced by such environmental factors as noise, fumes, and temperature.

The simple man-machine model provides a convenient way or organizing some of the major content areas of human engineering: the selection and design of machine displays and controls; the layout and design of workplaces; design for maintainability; and the work environment.

Human factors in large systems. No matter how important it may be to match an individual operator to his machine, some of the most challenging and complex human problems arise in the design of large man-machine systems and in the integration of man into these systems. Examples of such large systems are a modern jet airliner, an automated post office, an industrial plant, a nuclear submarine, and a space vehicle launch and recovery system. In the design of such systems, human-factors engineers study, in addition to all the considerations previously mentioned, three factors: personnel, training, and operating procedures.

Personnel, training, and operational procedures

Systems are generally designed for particular kinds of operators. A space vehicle, for example, is designed for a highly select handful of astronauts. Automobiles, on the other hand, are designed to accommodate a wide spectrum of people. In large systems, the specification of personnel requirements is an integral part of systems design.

Personnel are trained, that is, they are given appropriate information and skills required to operate and maintain the system. System design includes the development of training techniques and programs and often extends to the design of training devices and training aids.

Instructions, operating procedures, and rules set forth the duties of each operator in a system and specify how the system is to function. Tailoring operating rules to the requirements of the system and the people in it contributes greatly to safe, orderly, and efficient operations.

Applications. The basis of human-factors engineering —the consideration of information about human users in the design of tools, machines, jobs, and work environments—has always been present. One of the oldest and most efficient of human implements, the scythe, shows a remarkable degree of human-factors engineering, undoubtedly reflecting modifications made over many centuries: the adroitly curved handle and blade and the peg grasp for the left hand. All of this is in sharp contrast with the conventional snow shovel, a modern implement of generally poor design that has been blamed for many a wintertime heart attack.

The need for a more formal approach to these human problems was created when machines became vastly more complex than they had ever been. High-speed jet aircraft, computers, radar, nuclear submarines, communication satellites, space vehicles—all these are products of the last few decades. The fantastic growth in the number and complexity of machines has created entirely

new problems about the use of man and the way he can be integrated into systems. Moreover, the solution to these new problems can not be found in the collective wisdom of society. For example, only a few years ago no one had any way of predicting with any certainty how man could or would function in a weightless environment. Human-factors engineering is, therefore, a child of the times, born of a mechanized civilization.

Applications of human-factors engineering have been made to such simple devices as highway signs, telephone sets, typewriters, and stoves, and to a host of modern, sophisticated complexes such as data-processing systems, automated factories, and space vehicles.

The experience gained in devising these systems has contributed to the realization that even relatively simple devices raise unexpectedly important questions on human use—questions that conventional engineering practice frequently cannot answer.

Push-button telephone. The modern push-button telephone handset provides a good example of a relatively simple device that has required a great deal of human-factors engineering. The layout of the keys in the four rows of three buttons, for example, was selected only after extensive tests on a variety of arrangements: circular, two vertical rows of five buttons, two horizontal rows of five, and a diagonal pattern; the arrangement of the numerals and letters on the keys, in the order of left to right and from top to bottom, was chosen as superior to other arrangements such as that used on many adding machines and desk calculators, in which the numbers increase from bottom to top. The top-to-bottom design decision was not simply a matter of logic; tests showed that people actually made fewer errors and took less time with that arrangement than they did with the adding-machine arrangement. Other human-factor considerations in the design of the push-button keyset were the size and style of numerals and letters for maximum legibility, the optimum sizes and spacing of the keys, and the proper force-displacement characteristics of the keys to provide tactile feedback or "feel" with the button fully depressed.

Similar factors were considered in designing the shape of the handset itself. The locations, separations, and angles between the earpiece and mouthpiece were determined so that the assembly would fit comfortably around the greatest number of different human faces; and the weight of the handset was designed to be neither too light nor too heavy.

Space suit. The designing of a much more complicated device, such as a space suit, presents more intricate problems. A space suit is a complete miniature world, a self-contained environment that must supply everything needed for an astronaut's life, as well as comfort. The suit must provide a pressurized interior, without which an astronaut's blood would boil in the vacuum of space. The consequent pressure differential between the inside and the outside of the suit is so great that when inflated the suit becomes a distended, rigid, and unyielding capsule. Special joints were designed to give the astronaut as much free movement as possible. The best engineering has not been able to provide as much flexibility of movement as is desirable; to compensate for that lack, attention has been directed toward the human-factors design of the tools and devices that an astronaut must use.

In addition to overcoming pressurization and movement problems, a space suit must provide oxygen, a system for removing excess products of respiration, carbon dioxide and water vapour; protection against extreme heat, cold, and radiation; protection for the eyes in an environment in which there is no atmosphere to absorb the sun's rays; facilities for speech communication; and facilities for the temporary storage of body wastes. This is such an imposing list of human requirements that an entire technology has been developed to deal with them and, indeed, with the provision of simulated environments and procedures for testing and evaluting space suits.

Social problems. The telephone and the space suit are but two out of thousands of examples that might have been selected to show where human-factors engineering has been consciously applied to solve technological problems. The same human-factors principles and methods have lately been applied to a variety of social problems: education, medicine, law enforcement, architecture, city planning, highway and transport design, and pollution control. The modern concern with man's relationship to the total environment around him implies a much-broadened definition of human-factors engineering.

Applications to social problems

The human-factors approach to design. Two general premises characterize the approach of the human-factors engineer in practical design work. The first is that the engineer must solve the problems of integrating man into machine systems by rigourous scientific methods and not rely on logic, intuition, or common sense. In the past, the typical engineer was baffled by the complex and unpredictable nature of human behaviour; not understanding it, he was often inclined to ignore or to deal with it summarily by using his own best guesses. Human-factors engineers have tried to show that with appropriate techniques it is possible to identify man-machine mismatches and that it is usually possible to find workable solutions to these mismatches through the use of methods developed in the behavioral sciences.

The second important premise of the human-factors approach is that, typically, design decisions cannot be made without a great deal of trial and error. There are only a few thousand human-factors engineers out of the thousands of thousands of engineers in the world who are designing novel machines, machine systems, and environments much faster than behavioral scientists can accumulate data on how man will respond to them. More problems, therefore, are created than there are ready answers for them, and the human-factors specialist is almost invariably forced to resort to "trying things out" with various degrees of rigour to find solutions. Thus, while human-factors engineering aims at substituting scientific method for guesswork, its specific techniques are empirical rather than theoretical.

The profession. As a profession, human-factors engineering is very young. The first known course in engineering psychology was offered at Johns Hopkins University in 1947. In the 1970s, hundreds of educational institutions in the world offered human-factors programs, most of them either in schools or departments of engineering or in departments of psychology.

In the United States, the main professional societies are the Human Factors Society, the Society of Engineering Psychologists of the American Psychological Association, the Institute of Environmental Sciences, the Aerospace Medical Association, and the Man-Machine Systems Group of the Institute of Electrical and Electronics Engineers. Organizations outside the United States include the Ergonomics Research Society (with headquarters in Great Britain), the Ergonomics Section of the Czechoslovak Committee for Scientific Management, the Gesellschaft für Arbeitswissenschaft (Association for the Science of Work; Germany), the Human Factors Association of Canada, the Japan Ergonomics Research Society, the Nederlandse Vereniging voor Ergonomie, the Nordic Ergonomic Society, the Polish Committee on Ergonomics and Labour Protection, the Société d'Ergonomie de Langue Française, and the Societa Italiana di Ergonomia. These are, in turn, federated with the International Ergonomics Association (IEA), an organization (in 1970) with members in some 30 countries around the world. The principal journals in the field are *Human Factors* (published in the United States), *Ergonomics,* and *Applied Ergonomics* (published in Great Britain).

BIBLIOGRAPHY. A. CHAPANIS, *Man-Machine Engineering* (1965), a small, readable, but authoritative survey of the field; K.B. DE GREENE (ed.), *Systems Psychology* (1970), a discussion emphasizing those aspects of human factors that relate particularly to modern systems, both technological and social; E.J. MCCORMICK, *Human Factors Engineering,* 3rd ed. (1970), a large textbook covering all aspects of human factors engineering—the best of its kind; C.T. MORGAN *et al.* (eds.), *Human Engineering Guide to Equipment Design* (1963), a highly technical designer's guide; W.E. WOODSON and D.W. CONOVER, *Human Engineering Guide for Equipment Designers,* 2nd ed. (1964), a less technical guide.

(A.I.C.)

Humanistic Scholarship, History of

Humanistic scholarship has been of decisive importance in the history of Western culture. For 25 centuries over a vast area of the earth, the study of Greek and Latin and knowledge of the ancient literature of the Mediterranean world have formed the core of all education. This ideal model has at times inspired an almost spell-like religious veneration. It has comprised a complex ensemble of knowledge and faith; of reason, emotion, and passion; of past, present, and eternity—and the living presence of a 1,000-year tradition suggests that it will endure. As an amalgam of facts and values, the system of humanistic scholarship employs three kinds of data:

Kinds of data employed

1. The history of the culture of antiquity (literature, fine arts, philosophy, customs, and monuments): its origins, development, and preservation.
2. The history of pedagogy: institutions, courses of study, programs, and relations between teachers and pupils.
3. The history of the humanist ideal: the ideal of humanist studies fits in with a view of man's nature that is based on dogma or even with a mythology of human personality.

This article is divided into the following sections:

The humanities as a discipline
  Meaning of humanistic scholarship
  Nature of humanistic scholarship
The humanities in antiquity
  Ancient Greece
  The Hellenistic period
  Latin culture
The medieval period
  The compromise of the Church Fathers
  Byzantium
  The Western tradition
The Renaissance: Humanism reborn
  Greek culture as another sacred history
  The dawn of a new culture
From the 18th to the 20th century
  From belles lettres to archaeological reality
  Philology and humanistic scholarship as modern disciplines
Conclusion

THE HUMANITIES AS A DISCIPLINE

Meaning of humanistic scholarship. The English expressions "humanistic scholarship" or "classical scholarship" denote a many-faceted idea. According to the British historian John Edwin Sandys, scholarship may be defined as "the sum of the mental attainments of a scholar. . . . It is sometimes identified with 'learning' or 'erudition'; but it is often contrasted with it." By "classical scholarship" Sandys means "the accurate study of the language, literature and art of Greece and Rome, and of all that they teach us as to the nature and the history of man." The French scholar Salomon Reinach wrote that "classical philology"—as such scholarship is called on the Continent—"is the knowledge of the intellectual, moral, and material life of the Greeks and Romans." Concerned with culture and civilization as a whole, philology is different from linguistics, the study of language and languages, which limits itself to the scientific study of their vocabularies and grammars. The Humanists of the Renaissance used the Latin expression *humaniores litterae* in the sense of humanistic scholarship. This phrase was deemed incorrect, however, by Friedrich August Wolf (died 1824), a great German philologist, who proposed the German expression *Altertumswissenschaft* ("knowledge of antiquity").

Nature of humanistic scholarship. The diversity of terms designating humanistic scholarship reflects the ambiguousness of a discipline that closely links the knowledge of facts with the honouring of specific values. As early as the 2nd century AD, Aulus Gellius, a Latin rhetorician, stressed the need for literary studies as alone able to endow man with the fullness of his humanity, in accordance with the ideal of the Greek *paideia* (education) and the Roman *humanitas*. And even today, the *studia humanitatis* still constitute the privileged area of pedagogy, like some ancient revelation that has even resisted the revolutionary affirmation of the Judeo-Christian revelation. Literature and art, mythology and thought, and the history of classical civilization, consecrated by tradition, still comprise the mental decor of European education. According to the process metaphysician Alfred North Whitehead, all of "the European philosophical tradition . . . consists of a series of footnotes to Plato." Similarly, until the 19th century, the intellectual and moral consciousness of the West was little more than a never-ending commentary on classical culture.

Goethe wrote:

> May the study of the Greek and Roman Antiquity always remain the basis for all higher culture! The antiquities of China, India, and Egypt will never be more than curiosities; . . . they will never bear a great deal of fruit for our moral and aesthetic culture.

For Westerners, however, humanistic scholarship has always comprised their apprenticeship. Just as the Jews were molded by the sacred writings of the Torah and Talmud and the Muslims and other Oriental peoples by their own sacred texts, Europeans have discovered their spiritual identity and the secret of their humanist calling through the masterpieces of ancient culture.

Humanistic scholarship teaches no specific philosophical doctrine. It inspires, however, a mental and moral attitude that is characteristic of the West—an attitude that makes human consciousness the alpha and omega of all thinking. "Man is the measure of all things," said the Sophist Protagoras in the 5th century BC; the individual is a centre of values. Whereas the wise men of the Orient have tended to view the human personality as an illusion and an evil, Westerners have made it a major virtue.

Nevertheless, the teachings of humanistic scholarship have undergone declines and rebirths in the course of the past few thousand years. Their very meaning has been altered, and they have even on occasion been called into question. Since the scientific revolution of the 17th century, the primacy of literary values has been subjected to increasingly violent criticisms. Today classical studies no longer occupy their traditional place of honour. One of the major problems of the 20th century is to ascertain whether the old humanism is really outmoded and whether a new humanism can be developed.

THE HUMANITIES IN ANTIQUITY

Humanistic scholarship has been sharply circumscribed in both duration and extent. In the geography of humanism, the two focuses were at Athens and at Rome, and historically two centuries—the age of Pericles (mid-5th century BC), in which the Greek miracle occurred, and the age of Augustus (at the turn of the Christian Era), the high point of Latin culture—represent the most brilliant moments of 1,000 years. In view of the fact that the Homeric poems were composed around the mid-9th century BC, Emperor Hadrian—ruling from AD 117 to 138—was already an archaeologist, a lover of masterpieces of the distant past; and in his villa at Tibur replicas of classical monuments were built. His contemporary Plutarch meditated on the greatness and decline of the Greeks and Romans, while Pausanias wrote his *Description of Greece*, a detailed guidebook for sightseers. The age of creation was over, and a classicism that highlighted the riches of a culture consecrated by the veneration of generations had begun. Thenceforth the road to beauty and truth, to moral values, went by way of memory—drawing upon an inexhaustible reservoir of precedents, models, and examples to be imitated.

Historical focuses of classical creativity

Ancient Greece. From Homer through Aristotle (died 322 BC) humanistic scholarship revealed its rich resources.

The Homeric poems. The Homeric poems formed the ultimate backdrop for humanistic scholarship. The *Iliad* and the *Odyssey* mirror a more remote prehistory but give it a false appearance. It was only after AD 1870 that the stirring researches of Heinrich Schliemann, a German archaeologist, lifted the veil from the peoples of Troy, Mycenae, and Tiryns and revealed the specific details of those forgotten civilizations. Throughout the history prior to these findings, however, the masterpieces of the

blind Ionian poet Homer marked the magnificent starting point at which, in a single thrust, the treasures of humanistic scholarship arose.

The Homeric poems, which are viewed today as the culmination of a long, slow evolution, were long considered to have been the point of departure for a tradition that asserted itself all at once, like a gift from on high, fully ripened in all its wisdom and beauty. Actually, they perpetuate the memories of battles and adventures that occurred three to five centuries before the first elements of the traditional epic poem came into being. That vast saga was built by degrees by the gathering and organizing of dissimilar elements transmitted through memory by narrators and singers, probably at the beginning of the 9th century BC. These first epics, presented at civic and religious festivals throughout the Greek world, soon became praiseworthy models. In 594 BC a decree by Solon gave rules for the public recital of the Homeric poems. A little later, Peisistratus, who was in power from 560 to 527, set up a committee to codify Homer's text, to arrange its episodes in order, and to determine the exact wording, and he incorporated recitals from Homer into the Panathenaean festivals. The *Iliad* and *Odyssey* thus became part and parcel of the religious traditions, and their texts were endowed with canonical authority.

Textual redactions by the Alexandrian school

In due time, new revisions in the text were made by the critics and philologists of the Alexandrian school, particularly in the 3rd and 2nd centuries BC. These specialists, focussing their attention on verbal corrections and scientific interpretations, found factual errors and unwarranted additions, which they sought to correct. They explained the text and commented on it, clarifying the meaning or reconstructing it when it was lost. Zenodotus of Ephesus (died *c.* 234 BC) was probably the first to divide the epics into 24 books, which profoundly altered their rhythm. The late-3rd- and 2nd-century grammarians Aristophanes of Byzantium and Aristarchus worked tirelessly over the texts and issued greatly revised editions. It was this Hellenistic vulgate version that the West received, thanks to manuscripts from no earlier than the 13th century AD.

The story of the Homeric texts is as complicated as that of the Bible. Plato said of Homer that he had been the educator of Greece. By reading the poems and reciting them from memory, the feelings, imagination, and even the speech of the children and young people were molded. Thus, the Homeric ethic kept constantly alive a sense of chivalry in which the hero vied with the gods in glorifying their virtues. The characters in the *Iliad* teach a moral code of greatness and honour, in defeat as in victory, that seems to have survived all of the changes in civilization. Alexander the Great took the ancient poems with him during his military campaigns and sought zealously to reproduce the feats and the generosity of Achilles, son of Peleus. The ideal model of Homeric society—which was both aristocratic and fictitious—was to inspire university youth of the West for centuries; and the tales of the *Odyssey*, the first adventure novel in history, were to provide similar fascination.

Dionysian importations and early poets and dramatists. The limited horizon of the Homeric world was subsequently filled out, however, by various contributions from traditional wisdom and religion. The worship of Dionysus, an emotional god of fruitfulness, coming from Phrygia, Thrace, and the interior of Asia Minor and later giving rise to the Orphic mysteries, was grafted onto the Homeric poems. Onomacritus, a collector of the oracles of Orpheus and Musaeus, was a member of the committee appointed by Peisistratus to codify the Homeric texts; he is said to have been condemned for trying to insert into them some Orphic-like revelations.

Hesiod, Aeschylus, and Pindar

Hesiod's poems, said to have been written around 700 BC, flow naturally from the spiritual landscape of the Homeric epic. In his *Theogony*, which relates the genealogy of the gods of Mt. Olympus beginning with the primal chaos, the interplay of relations among the divine beings ends with the victory of Zeus, who maintains peace among gods and men according to the divine laws of the cosmos. As for the reality of human experience on earth, however, its principles are set forth in Hesiod's *Works and Days*, a poem on agriculture in which the cosmic rhythm of the seasons, marked by recurring festivals, determines the timetable according to which the earth is to be worked. Maxims and adages and proverbs and precepts set forth the rules of peasant wisdom, which ensure peace and prosperity.

The early-5th-century writer Aeschylus, first of the great Greek dramatists, wrote edifying plays that embodied this heritage of religion and myth and lyrical poems that paved the way for the humanization of mythology. The works of his contemporary Pindar, author of odes that extolled the victors in the Greek national games, developed the Homeric ethic of the hero who struggles and wins for the honour of his city, as well as his own glory, under the kindly gaze of the gods on Mt. Olympus. This wisdom, born of conflict and patriotism, had a high educational value; it furnished a repertory of themes and images that taught the youth the value of striving and the virtue of sacrifice, which led—through renunciation—to the highest achievements.

Thus, by the 5th century BC, humanistic scholarship possessed its epic framework, centring around the Homeric revelation. At that point, however, it was still simply an ensemble of mythological data, traditional precedents, and mental habits and attitudes that shaped the memory, sensitivity, and behaviour of the Greeks without rational reflection and expression.

The Sophists and Socrates. It was in Athens, thanks to the rise of the Sophists, the founding fathers of pedagogy, that the decisive step was made from *mythos* to *logos* (rational principles), from tradition to intellect. Thus, the Sophists—among whom must be included even their adversary Socrates—were the patriarchs of all Western intellectuals.

The rise of rationalism

The Sophists were humanists who sought to transmute the age of theology into that of rationalism. Curious about every aspect of humanity, they undertook a systematic study of human reality without any mythical or religious preconceptions. They analyzed traditional myths to reveal analogies and allegories of human existence. From that time on, pedagogy became the precondition for every reform in human understanding. According to the ancients, Socrates deserved the credit for having brought philosophy down from heaven to earth —*i.e.*, to the earthly abode of man. As teachers of wisdom, the Sophists prided themselves on preparing men to accomplish their human tasks within the framework of the city-state.

The art of pedagogy had as its goal the education of the political man. The technique of oratory, which enabled a citizen to manipulate his fellow citizens in accordance with the laws of public opinion, presupposed intellectual mastery. The concept of *paideia*,' as the Sophists conceived it, denoted a complete education based on grammatical correctness and intellectual accuracy in dialectical thinking, with the molding of the adult's personality as a goal. The Sophists were the first to organize a course of systematic studies, the seven arts (later called the "liberal arts"), which formed the harmoniously united, well-rounded, thoroughgoing, and complete education that was transmitted by succeeding generations of rhetoricians. Quintilian, a Roman rhetorician who inherited this tradition in the 1st century AD, expressed its essence in the formula *orbis doctrinae*—*i.e.*, the circle of knowledge that brings the sciences together in perfect unity.

The pedagogical work of the Sophists brought about a synthesis of the achievements reached by the Greek genius in the age of Pericles, and the intellectual tradition of the West rests on that foundation. Thenceforth, the ideal of an encyclopaedic *paideia* was destined never to be wholly lost. The first enlightenment in European history, that of the 5th century BC, was carried over into the 4th century by philosophers of genius who, with their systematizations of knowledge, added an extra dimension of doctrinal depth and metaphysical clarity to the somewhat empirical and technical work of the Sophists.

Plato and Aristotle. Socrates' pupil Plato (died 348 or 347 BC) was a master of inner contemplation, attained through the conversion of the soul, which detached itself from perceptible realities and united with the transcendent Ideas that conferred eternity. Though in Plato's view the scientific disciplines are privileged—mathematics, for example, opening the way to a liberating asceticism—the arts of public speaking and poetry are fraught with suspicion because they expose thought to the dangers of illusion and of enslavement to the perceptible. Plato excluded the poets from the ideal city he conceived, and he retained the ancient myths only to the extent that allegorical interpretations could reveal the supreme intelligence in its mystic essence. The school founded by Plato, the Academy, was both a small university, a brotherhood, and a religious sect the members of which sought to realize an ideal of inner perfection. Throughout the history of the West, Platonism—ever and again reborn—dominated the royal road of mysticism.

Plato's suspicion of the arts

Aristotle (died 322 BC) represents the other pole of European thought. He was a genius in systematic thinking and an encyclopaedic mind keenly aware of the empirical realities of nature and humanity. As a theoretician he drew up systems of rules for good reasoning; as an observer he brought biological knowledge to a level not surpassed until Darwin; as a social scientist he described the various forms of the human community; as a critic he defined a new science for judging the creations of art and imagination. So great was his analytical genius that it visited a kind of paralysis upon those who for centuries thereafter looked to him for authority and were content simply to repeat his teachings. The break with Aristotelianism, on the threshold of the modern age, was seen by some as a symbolic murder of the father—with a resultant unleashing of passionate emotions.

The Hellenistic period. From Homer to Aristotle, the masterpieces of classical culture came forth: writers of comedies and tragedies rubbed shoulders with the historians Herodotus and Thucydides and the great orators Demosthenes, Aeschines, and Isocrates. But, with the death of Alexander the Great in 323 BC, cultural history came to a temporary halt, or rather changed its direction. The age of creators was over; now came the continuers, who inherited and exploited the riches that had been created. Since they could not equal the masters of the past, the inheritors were content to study their works and imitate them. Henceforth, these works were held up as objects of admiration for the generations to come. It was not, strictly speaking, a time of decadence but, rather, a new way of asserting intellectual energies.

Transition from creators to continuers

Alexandria and Pergamum as cultural centres. At the same time the centre of gravity of culture shifted. "The monarchical city of Alexandria," wrote Sandys, "took the place of democratic Athens as the literary centre of the Greek world." The Ptolemies, who ruled on the Nile, and the Attalids, who ruled in Pergamum, made science and literature an attribute of the power and glory of the state. For the first time knowledge took the form of a public institution subsidized by the prince. Scholars and scientists were considered to be government officials and led a privileged life. The Mouseion (Museum), or House of the Muses, in Alexandria was the prototype of Oxford and Cambridge. It was also a huge library with hundreds of thousands of volumes; it was an architectural complex in which porticoes alternated with gardens; it had a botanical garden, a menagerie, and an anatomical institute. Thus, a cultural utopia was actually built of stones, books, and men 2,000 years before Francis Bacon, in his *New Atlantis*, described the "Treasure Island" of scientific research.

Alexandria remained a cultural centre for at least seven centuries: pagans, Jews, and Christians with a thirst for knowledge all lived there until Christian intolerance and then the Arab conquest extinguished its glorious history. Euclid's *Elements* (300 BC), Ptolemy's great work on astronomy, the *Almagest* (AD 135–150), Strabo's *Geography* (*c.* AD 23), and Galen's medical works (AD 130–200) —all came from the scientific capital of Egypt. Literary studies also were written there. The very notion of humanistic scholarship—*i.e.*, the systematic study of a body of ancient works raised to the level of classical masterpieces—was developed by the scholars of the Museum and their colleagues in Pergamum. In the words of the French writer Raymond Queneau,

> In three centuries, in the literary field, the classical writers invented everything. In three more centuries, the Alexandrians made another, equally far-reaching discovery: they invented the Classics.

Literary and textual scholarship and criticism. The Hellenistic period was the age of scholars, librarians, commentators, compilers, lexicographers, and archivists of the human spirit. At Athens the concrete living culture had been public property, a product of everyday consumption. At Alexandria and Pergamum, culture was enclosed in libraries, like a dead plant in a herbal; by dint of close study one had to revive the letter as well as the spirit of the work. The critics were editors, concerned with restoring and perpetuating the original words. At the very moment (3rd century BC) when the definitive text of Homer was being prepared in Alexandria, the Greek version of the Hebrew Old Testament, called the Septuagint, was also being completed in that same city and with the same care. Textual criticism broadened out into literary criticism; philology was born as a science of interpretations; and new methods in grammar, analysis, and prosody made possible a new approach to great works of literature. And the canon of classical literature was formed along with that of the Bible. Texts were carefully evaluated according to the criteria of taste, along the lines of Aristotle's *Rhetoric* and *Poetics*. "Rhetoric, from this time onwards," wrote George Saintsbury, a literary historian and critic, "more and more tends to become the Art of Literary Criticism generally, and to absorb Poetics within itself." In the libraries of Alexandria and Pergamum, culture became what it has since remained: an imaginary museum of hallowed masterpieces, tirelessly commented upon by scholars who watch jealously over correctness of wording and accuracy of interpretation. Hellenistic classicism was, thus, the first version of Scholasticism.

Development of a canon of definitive editions

Although Alexandria was also the centre of Neoplatonic philosophy, the Alexandrian philologists were especially proficient in textual criticism. At Pergamum, whose parchments competed with Egypt's papyruses, the ideas of the Stoic school of philosophy were dominant. There a cosmopolitan spirit encouraged freer interpretation, tinged with allegory. Reaching out from the Mediterranean centres, the new culture spread eastward—*e.g.*, to Syria, where (beginning with the 2nd century BC) Antioch became a centre of study and a transmission point for Greek thought—and it remained in the Mesopotamian region and beyond until the Arab conquest. Muslim culture then fell heir to this culture; thus the works of Aristotle, Galen, and the rest were first transmitted to the West by way of translations made in Muslim countries.

Latin culture. Roman culture was merely the extension of Hellenic culture into the western part of the Mediterranean Basin; the Latin language never made its mark in the East. Victorious on land and sea, the people of Rome quickly found themselves at the height of power, even though they remained intellectually and artistically underdeveloped. Miraculously, the decline of Greece in civil and military matters did not signify the end of Hellenism. Roman supremacy went hand in hand with a genuine cultural conversion of the conquerors in the course of the 2nd century BC: Rome became the pupil of Greece, and all of the masterpieces of Latin civilization—from Cicero to Horace, Livy, Virgil, and Ovid—were simply imitations or pastiches of classical models, a kind of Western postscript to the Greek miracle. The Roman national epic, the *Aeneid*, acknowledged this connection by borrowing its hero from the camp of the vanquished in the *Iliad*. Similarly, the great Latin writers were simply epigones (imitators) of Hellenism.

The pedagogic model of Latin *humanitas*, brought to Rome by Greek professors, was only a renewal of the liberal arts program along the lines of *paideia*. The culture of the Romans bore a pragmatic stamp; it developed

the intellectual tools needed by an elite of jurists and officials and administrators and soldiers who devoted themselves to organizing and governing the world according to universal norms. Whereas the Greek genius was one of contemplation, the Roman genius proved itself in action. Thus, the creative spontaneity of the great writers of the classical age of Rome—the Augustan age—was under the discipline of a political operation of national recovery.

The ideal of *humanitas* and rhetoric

Rhetoric, as it has come down to the West, was codified by Quintilian, a Roman rhetorician and writer, in his *Institutio oratoria* (Eng. trans., *The Institutio Oratoria*, 4 vol., 1920–22) around AD 95. In the art of oratory the elements of general culture, from poetry to philosophy, were gathered together in a synthesis as a way for men to influence their fellow humans. According to Saintsbury:

> The *Institutes of Oratory* contain the . . . most satisfactory applications of criticism to literature, as it presented itself to an intelligent and thoroughly educated person . . . when, except for a few belated authors, mostly of curiosities, the list of the great writers of antiquity was all but closed.

Thenceforth, the system of humanistic scholarship constituted a private preserve that the teachers of rhetoric exploited in order to train scholars. The platitudes (*topoi*) of literature were used as themes for discussion, whose various stages and technical procedures were strictly defined and expressed. Rhetoric, an epitome of classicism, dominated medieval Scholasticism; it inspired university teaching until the end of the 18th century.

Varro, Pliny the Elder, and Plutarch

Apart from rhetoric, Rome's contribution to the Western tradition remained limited. The Latin compilers and encyclopaedists were distinctly inferior to those of Alexandria. The works of Marcus Terentius Varro, an immensely learned and prolific scholar, are almost all lost, except for several precious books of his *De lingua Latina*. The *Historia naturalis* of Pliny the Elder (died AD 79), a scholar and civil servant, reveals by contrast Aristotle's genius. Nevertheless, the classical heritage did include the works of the great Roman historians in which memories of the Res Publica Romana and the Imperium Romanum were kept alive as models. The genius of Rome, which was great in action rather than in meditation, was embodied in the legendary virtues of its great men: Sallust, Livy, Tacitus, and Suetonius were teachers of a concrete morality that honoured self-sacrifice and greatness of soul. To these must be added the writings of Plutarch (died *c.* AD 119), a Greek biographer and moralist, who compared the heroes of Greek history with those of Rome. Plutarch was largely instrumental in keeping alive the century-long worship of energy and the public-spirited republicanism that, paradoxically, the later Scholastic civilization nostalgically associated with its evocations of Roman grandeur.

THE MEDIEVAL PERIOD

Cultural divergence into Byzantium and the Roman West

In the history of classical scholarship, the medieval period witnessed the divergence of the two branches of the culture of antiquity—the Greek and the Latin—that coincided with the long decline and final fall of the overexpanded Roman Empire. It was the period in which Christianity, having become the state religion, caused its values to take precedence over pagan teachings, which involved a complete restructuring of mental and spiritual attitudes.

In AD 330 the emperor Constantine chose Constantinople (Byzantium) as the new capital of the Roman world. Byzantium was destined to become the sanctuary of Greek-language culture and to be separated thereafter from the Latin culture that continued in the West, a separation that defined the intellectual order of the medieval world. The Middle Ages ended, however, in 1453, with the fall of Constantinople to the Turkish army. As a result, Byzantine scholars returned to the West with their manuscripts, which in turn kindled a rebirth of literature and a new ardour for the humanities, with emphasis on their literal integrity and spiritual inspiration. The Christianity that was based at Rome had retained the common heritage of classicism, but in its Latin form it had been weakened and diluted by Scholasticism. The return to Greek sources, however, signalled the advent of the humanities in their modern form.

The compromise of the Church Fathers. Early in the Christian Era, the inevitable problem of the new faith's accommodation with classical culture arose.

The antipathy between Athens and Jerusalem. The consolidation of Christianity, which prolonged the Hebrew–Christian tradition, led eventually to a fundamental questioning of the traditional pagan culture of Hellenic orthodoxy. A polemic ensued, one that lasted for several centuries. It was a dialogue fraught with emotion that marked the death throes of paganism. In the course of the controversy, professors, writers, even emperors—such as Marcus Aurelius and Julian the Apostate, upholders of a glorious past—confronted the learned doctors who had received the teachings of Christ. The masterpieces of ancient art and thought clashed with the revolutionary challenge of the Nazarenes; and to the Greeks, the notion of the good tidings of the crucified Jesus was scandalous and insane. The English historian Edward Gibbon summed it up in his famous statement: "I have described the triumph of barbarism and religion." In AD 391, under Theodosius I, the Sarapeum of Alexandria, a library and research centre, as well as a temple to Sarapis, was turned into a monastery; in 415 Hypatia, daughter of the last known member of the Mouseion who was herself a noted philosopher and mathematician in the pagan tradition, was murdered by a fanatical mob. Finally, in 529, the emperor Justinian ordered the philosophical school of Athens, the last refuge of the pagan mind, to be closed; and in that same year St. Benedict, patriarch of the monks of the West, founded the monastery of Monte Cassino in Italy.

The closing of the last pagan school

Thus, between Athens and Jerusalem and between humanistic scholarship and Jesus Christ, no understanding seemed possible; the radical alternative would have entailed the total loss of classical culture, which was incompatible with the new religious requirements.

The accommodation between Athens and Jerusalem. But the victorious new faith did not possess a system of Christian scholarship that could be counterposed to classical scholarship. Lacking such a system of their own, the Christians realized the need to utilize pagan doctrine if they were to avoid slipping into barbarism. Thus, as early as AD 96, the formula *en Christō paideia* appeared, which avowed the possibility of a cultural conversion by which works of Hellenic inspiration could be subordinated to religious faith. The Christian Church now took charge of the intellectual and aesthetic values of the past.

The principal achievement of the Church Fathers was to bring about this cultural updating. They soon appropriated, like war booty taken from a defeated enemy, not only the language and literature but also the aesthetic structures and ideas of ancient paganism. Humanistic scholarship was thus saved and put to new uses to serve the Christian revelation. The bases of Western culture are the result of this alliance between two spiritual outlooks, alien to each other.

According to the Church Fathers, the Judeo-Christian tradition, because it alone went back to the creation of the world, was the source of all culture and could thus absorb the pagan tradition. In their view the genealogical tree of knowledge passed from Moses through the Egyptians (of whom he was the teacher) to the Greeks, whose works could therefore be considered divinely inspired. Such was the claim of the Christian scholar Justin Martyr (died *c.* 165), who acknowledged that the Word of God could have inspired the great pagan authors. About the turn of the following century, Clement of Alexandria, the founder of Christian philosophy, asserted that pagan wisdom to the Greeks was the equivalent of the Old Testament to the Jews prior to the revelation of the Son of God. The new culture represented a fulfillment of, not a breach with, the old culture. Around the turn of the 5th century, Jerome (*c.* 347–420) and Augustine, the most prestigious figures among the Christian intelligentsia and scholars of the highest quality, had a profound respect for classical humanism; and, at a time when barbarians were besieging the Roman Empire on

every side, these men smoothed the transition from the Imperium Romanum to the medieval Romania by seeking to save what could be preserved of the venerable treasure of antiquity.

The Church Fathers played a key role in the history of humanistic scholarship. They were responsible for the peaceful coexistence of the double heritage—the pagan and the Christian—that molded the consciousness of the West. Thanks to them, Christianity entered into the domain of pagan culture, which thus ceased to be viewed as a danger to human minds and emotions. In the 15th century, the tradition of Christian humanism was continued by Marsilio Ficino, founder of the Florentine Academy, and by Giovanni Pico della Mirandola, notable for his *Oratio de hominis dignitate;* in the 16th century, by Erasmus, John Colet, and Juan Luis Vives; and in the 17th century, by the Cambridge Platonists.

Byzantium. It is to the scholars of Byzantium that modern scholars are indebted for the preservation of Greek culture.

Political and religious conditions. During the Middle Ages the unity of the Imperium Romanum, at the height of its expansion, was split into three distinct areas, which differed in religion and language, as well as in politics. All during the Middle Ages (*i.e.*, until 1453), Romania, Byzantium, and Islām were in a continuous state of action and reaction, with military expeditions going hand in hand with cultural and commercial exchanges.

Moreover, the three cultures, despite their religious differences, possessed in common the spiritual tradition of religions of the book. They also shared the treasures of the profane writings of antiquity. Arab scholars and their Jewish colleagues translated, copied, and commented on the great Greek works and were largely responsible for preserving and transmitting them to the farthest reaches of Christendom. Thus, the medieval period, though beset by military struggles and internal contradictions, kept its unity as far as respect for the cultural tradition of humanistic scholarship was concerned.

The political unity of the West was the first of the three to be broken—as a result of the barbarian invasions but also in large measure because of its own inner decay. This great region now found itself wrenched apart and reduced to a state of cultural stagnation. For more than 1,000 years the principal role of Byzantium was to ensure the survival of the Greek culture and its language (which had become virtually forgotten in the West). Christianity had undoubtedly triumphed, and its spiritual and theological preoccupations left their mark on the civilization as a whole. The controversy over Iconoclasm (727–850)—*i.e.*, opposition to the religious use of likenesses of the divinity—was a Christian backlash against pagan culture, the effects of which were also discernible in the literary field. During these many centuries, cultural life had its peaks and valleys, but the updating carried by the Church Fathers generally remained valid; the most brilliant moments of the Byzantine Empire were its revivals of Greek studies.

The two golden ages of Byzantium

The development of Byzantine scholarship. The first great century of Byzantium, that of the renowned emperor Justinian (reigned 527–565), is associated with his success in re-establishing the unity of the empire by reconquering provinces from the barbarians. The dedication of Hagia Sophia, the cathedral in Constantinople (537), was one of the highlights of his reign. As a Roman emperor, Justinian upheld the rights of the Latin language in Byzantium. It was basically in Latin that the major task of codifying Roman law—the Codex Justinianus, or Justinian Code—was accomplished. The Hellenization of the empire was achieved in the 7th century. Renouncing the West, Byzantium then became the capital of Greek power in the East, at the very moment when Islām began its rise. Byzantium's second golden age, which occurred between 867 and 1081 under the Macedonian dynasty, was highlighted in its earlier years by the philological studies of Patriarch Photius and later by the literary, philosophical, and historical writings of Michael Psellus, both of whom were staunch upholders of the pure Hellenic tradition.

Byzantium also had its dark centuries and its long, slow agony. It witnessed wanton destruction and sometimes irreparable losses in the legacy of ancient writings. In this respect, the capture of Constantinople by Western crusaders in 1204 had the most appalling consequences. But by and large, the role of the Byzantine scholars, who were not conspicuously creative, was that of being conscientious preservers of the great works of the classical canon established in Alexandria. Like the teachers of Alexandria, they became the respectful heirs of the texts, which they studied with care, making abridged editions and anthologies of them. Since a manuscript's life is limited in time, the mere fact that they served as copyists was salutary. The only field in which they were innovators was that of history. Eusebius of Caesarea (died *c.* 340), Constantine's historian, was also the author of *Historia ecclesiastica* (a Christian church history), the first of its kind and source of all the others. Eusebius' work was based on a chronology of world history beginning with the creation. This type of overall design, carried over to the West by Jerome, a Latin Church Father, served as a guide for viewing European history in perspective. Even in Byzantium this historical tradition was preserved for many centuries, and works of this genre were written by a historian, Leo Diaconus, near the end of the 10th century and continued by a philosopher, Michael Psellus.

In the history of humanistic scholarship, the Byzantine scholars were above all preservers of Greek culture. In the words of Frederic Harrison, an English writer:

> The peculiar, indispensable service of Byzantine literature was the preservation of the language, philology, and archaeology of Greece. It is impossible to see how our knowledge of ancient literature or civilisation could have been recovered if Constantinople had not nursed through the early Middle Ages the vast accumulations of Greek learning . . . ; if [they] had not poured out their lexicons, anecdotes, and commentaries; if the *Corpus scriptorum historiae Byzantinae* had never been compiled; if indefatigable copyists had not toiled in multiplying the texts of ancient Greece. Pedantic, dull, blundering as they are too often, they are indispensable.

The Western tradition. While Byzantium maintained the continuity of political power in the Hellenized East, the Romania of the West saw its cultural heritage, like its civil authority, shattered by the barbarian onslaught.

Political and cultural conditions. The power groups that slowly emerged with the advent of feudalism did not include a lettered aristocracy. Brought to power in a time of dire distress, the new authorities found themselves in a cultural void. Consequently, responsibilities were transferred to the only power capable of rescuing what was salvageable—the church—which, with the withering of secular authority, represented the only cultured class, as well as the only power with wide-ranging authority. Thus, the church became the sole patron of all the fields of art, knowledge, and education; theology now became the science of sciences, and its demands in matters of learning became law.

Nostalgia for the Imperium Romanum in the Middle Ages

Under those conditions the Latin Middle Ages, even in their darkest periods, were haunted by the nostalgic myth of the Imperium Romanum and subjected to a feeling of inferiority with regard to a grandiose past. Characteristically, throughout the Middle Ages the task of renewing the empire was linked with that of renewing learning. A cultural renaissance was included along with the political renaissance in the same plan of restoration. Thus, Theodoric the Great, king of the Ostrogoths who ruled over Italy through the early 6th century, undertook to repair the ancient monuments; in about the same period Boëthius, a philosopher, and Cassiodorus, a historian and statesman, translated and digested in Latin the great works of ancient Greece. Charlemagne, emperor of the West from 800 to 814, took steps to remedy the wretched state of culture: his Carolingian renaissance brought to the fore the Anglo-Saxon scholar Alcuin and his disciple Rabanus Maurus, and Einhard the Frank. A short time later, Emperor Otto I, called "the Great," who ruled from 936 to 973 and who, in turn, sought to rebuild the Roman Empire, similarly tried to foster a cultural re-

vival, in which the outstanding name was that of Gerbert of Aurillac, later Pope Sylvester II. Gerbert was a great scholar: proficient in Greek and Latin, he was fond of ancient manuscripts; and, as tutor to the young Otto III, he dreamed along with his imperial pupil of restoring the splendid empire of Constantine the Great. The flowering of medieval culture in the 12th and 13th centuries was characterized by the same desire to revive the past.

Medieval culture was rooted in the memory of a golden age. Its men were like dwarfs perched on the shoulders of the giants who had preceded them. Progress could only mean recouping the lost treasures of classical scholarship. In the meantime, men had to be satisfied with manuals and digests and collections of selected passages. Thus, the authorities became such men as Macrobius (start of the 5th century), a Neoplatonist, who prepared a commentary on Cicero's *Somnium Scipionis* in which Homer, Plato, Cicero, and Virgil are canonized, and Martianus Capella, a North African pagan, who composed between 410 and 439 an allegorical novel, *De nuptiis Philologiae et Mercurii*. Other compilers, who mingled pagan wisdom with Christian doctrine—thus ineptly perpetuating the memory of the vanished high-classical culture—enjoyed widespread renown. In the West, as in Byzantium, scholars dwelt amid a humanistic scholarship that was revised by the Church Fathers. But the West suffered under two severe handicaps: (1) the major setback caused by the barbarian invasions and (2) the fact that not a single scholar had access to the Greek originals, which were known solely through partial Latin translations, some of which had reached them by way of the Arab world. Thus, the Western feeling of inferiority was based on reality. It is therefore understandable why they idolized Aristotle, whose genius ranged far above their own.

Creative scholarship in the universities. There was, nevertheless, one field in which medieval civilization demonstrated its creative ability: in the founding of the university. In the West, schools had previously developed in the monasteries and cities under the sponsorship of the ecclesiastical authorities, who saw that the priests were educated in church doctrine. Some of these centres were for specialized studies: Salerno, for example, was a centre of medical education from the 9th century, and Bologna was the legal capital of the West, long before its status as a university was established. After protracted struggles between the local civil and religious authorities, the universities received their written charters from the Holy See. Thus the University of Paris received its charter in 1215 and thereby initiated a new era in the intellectual history of the West.

In the medieval system, the university was a corporation, but it was also a seat of culture in which knowledge was gathered and enunciated. In this manner its autonomous intellectual function was recognized and consecrated; thus, alongside the religious *sacerdotium* and the political *imperium*, the *studium* was the site reserved for the full pursuit of knowledge. Undoubtedly, the uses of reason were subordinated to the preconditions of theology, but the latter was itself thoroughly permeated with the rational spirit as refined by the practices of Scholasticism.

The Scholastic method

Despite its eventual decline, Scholasticism represents the first attempt to set up the general principles of knowledge as a whole, thanks to the establishment of a universe of systematic discourse. Within the confines of the university, learning was regulated according to the rituals and prescriptions of the community of scholars. Before the advent of the printed book, truth was established orally; it arose in conformity with fixed rules amid the exchange of arguments, of which Scholasticism defined the code. In the ceremony of the *disputatio* ("debate"), the themes of the *quaestiones disputatae* were presented, and the systematic analyses of these themes enabled one to arrive at *summae* (syntheses), in which the mastery of the great doctors of learning shone forth.

The new contributions—the university as an institution and the Scholastic method—still retained the preconditions of humanistic scholarship. The intellectual dialectic of Scholasticism—in matters of rhetoric, science, and philosophy—was based upon themes formulated by masters of the past, whose teachings were authoritative; and the process of argumentation quickened as soon as it fastened upon the text of a good writer. Truth was a truth of repetition and commentary; it was contained, essentially, in the works of the ancients, specifically in the writings of Aristotle—a profane revelation comparable in authority to that of the sacred religious texts. As for the system of knowledge, which formed the essentials of the university, it carried on the ideal of the *enkyklios* ("comprehensive") *paideia* as defined by the Sophists and developed by the Roman rhetoricians. The faculty of arts —charged with the studies prerequisite to those in theology, medicine, and law—included in its program the seven liberal arts: the literary *trivium* (grammar, rhetoric, dialectics) and the scientific *quadrivium* (arithmetic, geometry, harmony, astronomy). Thus, the pattern of the *paideia* that prevailed in Plato's time survived for 2,000 years.

The *trivium* and the *quadrivium*

The Romania of the West—vaster and more cohesive and surer of its fate than the Byzantine Empire—constituted a unified cultural whole. Under the authority of the Roman Catholic Church, the network of European universities dotted Europe like magical mountain peaks. Professors and students travelled all over the West, thus ensuring the free circulation of ideas despite political barriers. Unity of language guaranteed the community and communication of thought everywhere. From one end of Christendom to another, as the language of the church, Latin modulated the communion of prayer and the liturgies of faith, as well as those of reason.

THE RENAISSANCE: HUMANISM REBORN

Greek culture as another sacred history. As a capital event in the development of Western culture, the Renaissance represents a dramatic turn in the history of humanistic scholarship.

Impact of manuscripts from the East. A transformation of the European consciousness was seemingly made imperative by the establishment of a new and direct relationship with the heritage of Hellenic antiquity. Even before the fall of Constantinople to the Turks in 1453, cultural exchanges had occurred between Eastern and Western Christianity. During the first half of the 15th century, the political and religious powers of Byzantium, faced with the extreme gravity of the Turkish peril, had sought aid from the West, even if the price of it had to be an adherence of the Eastern Church to Roman orthodoxy. The occasion of negotiations to this end brought a certain number of Byzantine scholars to the West; others came to escape the Ottoman menace. First in Italy and then in western Europe, they re-established the knowledge of the Greek language, which had become quite rare. With these professors came the texts: many manuscripts passed from East to West, carried by refugees or purchased at the source by specialists sent as emissaries by potentates now seized with a passion for Humanism.

The Renaissance comprised a return to the Hellenic sources of Western culture, whose authentic meaning had been deformed and diminished by Scholastic Latinism. The doctors of the Middle Ages had paid tribute to the ancients, but they knew little or nothing of their works, except through recent adaptations or selections done by mediocre minds in an uncertain vocabulary. The Latin translations acted as a screen masking the Greek original; and medieval Latin itself, moreover, was corrupted by ecclesiastical and Scholastic overlays and thus was only distantly akin to the language of the masters of Rome. The rediscovery of the authentic texts meant the revelation of a hidden treasure that Byzantine scholars had held in charge for 1,000 years. It was the accident by which the texts were transferred from Constantinople to the West that brought about the resurrection of the ancient ideals of *paideia* and *humanitas* in the form of Renaissance Humanism.

Resurrection of *paideia* and *humanitas*

A complete break now took place with medieval culture, which had been discredited by the degeneration of

Scholasticism, which, beginning in the 14th century, had turned more and more into an artificial kind of dialectics. The sanctification by means of culture now tended to replace the religious sanctification. The ancient masterpieces brought to mankind new reasons to live and to hope and a new ideal model of personality. And there now began to take shape the modern type of cultivated man, distinguishable by the fact that, in his youth, he had visited the imaginary museum of the masterpieces of art. Thus consecrated, culture became the faith of the 15th- and 16th-century Humanists: their life was a passionate search for available manuscripts, statues, and coins—with, at the same time, an indefatigable delving into their meanings. Greek and Roman antiquity took its place as a second sacred history, not obliterating the Bible but sometimes overshadowing it. This devotion to the past at times reached the proportions of an abdication of personality: the old became a parasite on the new; humanity was living a double life, and the more important of the two was not that of the here and now. As emphasized by the Italian Leonardo Bruni, an early-15th-century Humanist: "I have the feeling that the days of Cicero and Demosthenes are much closer to me than the sixty years just past."

The dignity of the free man. Humanistic scholarship thus became the innovative source of a spiritual conversion. The Renaissance scholar, in the mirror of the ancient texts, discovered an unsuspected humanity. The clerics of the Middle Ages had known their Ovid and Virgil; but they had read the *Ars amatoria* (*Art of Love*), as they did the Song of Solomon, as mystically embodying divine love. Renascent philology, however, exposed the all-too-human evidences. The literary and philosophical treasure of antiquity now offered for investigation a humanity free of Christian dogmatism, as well as of Scholastic axiomatics.

The literature of antiquity, purged of its Christian overlays, revealed the worth of the human being in the free expansion of his nature no longer weighted down by a corrupting heritage of original sin. If Renaissance Humanism in its highest embodiments—as in Erasmus or John Colet—aroused the enthusiasm, as it were, of a new golden age, it was because it celebrated the sacrosanct worth of man. Man thus took unto himself the attributes of divinity, in the very first instance the power of creation. And his most wonderful creation was man himself in the full excellence of his humanity. The praise of the human condition was the theme of the *De dignitate et excellentia hominis* (written 1452) by Giannozzo Manetti, a Florentine philologist, and the *Oratio de hominis dignitate* (1486) by the young and brilliant Humanist Giovanni Pico della Mirandola. Christian faith, however, was not necessarily repudiated. The devotees of Humanism such as Marsilio Ficino and Erasmus himself were also Christian believers—but believers in a Christianity that had seen a rebirth of the spirit and had been aerated and put on its guard against the perils of superstition. The association of Socrates with Christ, implicit in Erasmus' temptation to canonize Socrates, implied a profound alteration in the traditional image of Christ, as well as a new reading of the gospel—often in a Platonic spirit.

The universal Renaissance scholar

Humanistic philology. The Renaissance *homo universale* (or "universal man")—at the same time philologist, artist, creator, and encyclopaedic genius in the spirit of the ancient *paideia*—appeared as a new ideal type of the human condition brought into being by the renewal of humanistic scholarship. The knowledge of humanistic scholarship was the privileged path to spiritual formation; philology as science became associated with philology as value. Apprenticeship to the ancient languages became the surrogate for ascetic monasticism. Philology then passed over into hermeneutics—*i.e.*, into the search for meaning according to the various modes of approach offered by the study of antiquities. The Humanists were engaged in the huge task of establishing authentic texts through the collation of manuscripts. The invention of the printing press in the middle of the 15th century provided for the rapid circulation of the ancient literature. Philology, as an interdisciplinary science, now covered the history of men and institutions, the study of civilization in its various aspects—religion and philosophy but also mores and customs, fine arts, techniques, public works, weights and measures, coinage, and so on. The fervour of the philologists, such as the Italians Lorenzo Valla and Marsilio Ficino and the Frenchmen Robert Estienne, Guillaume Budé, and Julius Caesar Scaliger, was sustained by the devotion of the printers of Venice, Lyons, and the Netherlands. A common market of the humanities thus arose outside the network of the universities, which, for the most part, were reluctant to welcome the new learning.

Repudiation of medieval Latin and Scholasticism

As a source of values and a way of life, humanistic scholarship rejected the barbarous and degenerate medieval Latin of the church and the university. The scholars effected a return to the classical language, modelled after Cicero and the poets of the Augustan century. This development was decisive for the future of culture: through it, humanistic philology, breaking away from the living continuity of Scholastic expression, prepared the death of the Latin language, which occurred in the 15th century. Ciceronianism, with its servile imitation of the classical models, entered into the pedagogical customs, as new exercises in Latin speech, Latin composition, and Latin verse were adopted.

The dawn of a new culture. Thus, a new definition of culture was formulated.

The common culture of European schools. Education, having escaped from the control of the Scholastic system, now took on a new structure. The colleges disseminated the curriculum of Humanist education through all of Europe. By action of the Jesuits the number of such institutions increased. Everywhere they established the same prototype, consisting of an education in Latin based upon the study and imitation of classical models. Within the shelter of the school, the educated youth of the West lived an anachronistic existence in the imaginary setting of a stereotyped antiquity. The myths of Homer's Greece, Pericles' Athens, Alexander's epic campaigns, and the Roman Republic all molded their personalities and excited their imaginations according to the basic requirements of style considered as a capital virtue. Romulus and Remus, Paris and Helen, Dido and Aeneas, Augustan clemency, Nero's evil ways, and the heroism of Regulus replaced the golden legend of medieval saints with that of ancient gods, heroes, and great men. Such was the imaginary backdrop against which, in modern Europe, drama and tragedy and opera and comic opera would be projected. Such schools were the crucibles for both authors and audiences of Shakespeare's *Julius Caesar* and *Coriolanus*, such operas as Claudio Monteverdi's *Orfeo* and Christoph Gluck's *Alceste*, the tragedies of Pierre Corneille and Jean Racine, Goethe's *Iphigenie auf Tauris*, Jean Anouilh's *Antigone*, and Eugene O'Neill's *Mourning Becomes Electra*, not to mention Jacques Offenbach's *Belle Hélène*.

This restoration of humanistic scholarship, essentially in its Latin form, helped to establish a new common cultural feeling in the West. The Humanists ensured the unity and universality of European education for centuries to come at the very time when the Reformation was dividing Christendom against itself. Educated men, despite their religious differences, could find communion in the educational ideals that had formed them all equally. Though members of the Reformed churches of Germany, England, and France would henceforth read Holy Scripture in their native tongues, they would continue to read the sacred writings of Humanism in the original Latin.

Exposure of classicism's provincialism

Broadening horizons of discovery, invention, and science. Through a paradoxical coincidence, the provincialism of Latin classicism was increasing at the very time that the limits of the Old World were being surpassed: for the 15th century saw the beginning of the great adventure by land and by sea, undertaken by the Portuguese and the Spanish, that opened up new worlds in distant places. The growth of knowledge, both through technical inventions, as well as through explorations, went hand in hand with a growth of power. The dis-

covery of the horizons of the globe had as a consequence the discovery of mankind, in its full variety of forms, as represented by the "savages" of the world. And the scientists studying the faraway reaches of sky and earth and the human body and the ordering of nature were continually discovering new facts that gave the lie to the Scholastic order but that had also escaped the notice of antiquity.

Thus, a veneration for the prestigious genius of the ancients developed at the very time when the intellectual situation clearly revealed the shortcomings of ancient culture in terms of the understanding of the real world. According to Francis Bacon (died 1626), author of the inductive method, the geographical New World demanded the elaboration of a new mental world. Unquestionably the classical heritage retained its value on the aesthetic and moral levels, but in Bacon's view modern men, writing in their own tongues, could also compose masterpieces to set alongside the classics. And on the scientific level, the superiority of the moderns was apparent without argument. Bacon clearly described the need for a new culture and a plan for it that, while giving the ancients their due respect, would prepare a synthesis of the newly acquired knowledge and organize its future development. The 17th century would be, in Bacon's perspective, the century of scientific academies and societies: the Accadèmia del Cimento in Florence, the Royal Society in London, and the Académie des Sciences in Paris.

According to Galileo (died 1642), the book of nature was written in mathematical signs—implying that it was not written in Latin. Bacon's *Novum Organum* (1620) was to replace Aristotle's *Organon*, using different principles. Mechanistic thinking portended a swift cultural revolution. Humanistic scholarship, being backward looking, stood in grave danger for want of due attention to the present. That was the true significance of the quarrel of the ancients and the moderns, which broke out in 1687 with the reading of a poem by Charles Perrault, author of several favourite fairy tales, at the Académie Française and spread throughout Europe. It was the opening blast of the great crisis of Western education, for which no solution has yet been found.

FROM THE 18TH TO THE 20TH CENTURY

From belles lettres to archaeological reality. During the 18th century, humanistic scholarship passed from the sacralized ideal of belles lettres to the scientific investigation of what antiquity really was.

Perpetuation of traditional classical schooling. In the century of the Enlightenment, humanistic scholarship continued to define the common pedagogical law of European teaching.

Jesuit, Pietistic, and national schools

The Jesuit type of school—duplicated hundreds of times in Europe, America, and Asia and imitated by responsible non-Catholic authorities—remained in force everywhere. It might have seemed that the persecution of the Jesuits, their expulsion from Europe, and the banning of their society by the Holy See (1773) should have been the occasion for an active reconsideration of teaching problems in view of the need to fill the resulting pedagogic vacuum. But the primacy of classical Latinism remained essentially untouched. The only significant innovation was the proliferation in Germany of *Realschulen* ("practical schools")—technical institutes or modern secondary schools developed under the influence of Lutheran Pietism. These institutions were meant for the children of the lower bourgeoisie and of craftsmen, to whom they offered a concrete curriculum in which geography, mathematics, experimental and natural sciences, and modern languages predominated, as befitted the current utilitarian concerns. At the end of the 18th century, the authorities created by the French Revolution set up a system of modern "national education," which was also freed of the fascination of the classical humanities. The latter regained the major part of their prestige, however, with the educational reforms instituted by the Napoleonic regime and then maintained by the monarchy restored in the 19th century.

Thus the social elite of European youth continued to study a curriculum inspired by the Humanists of the 16th century and codified by the Jesuit masters, despite the rapid transformations of the technological and industrial revolution. The intellectuals of the Enlightenment, themselves trained in the selfsame schools, were on familiar terms with the ancient authors, who continued to mold their intellectual sensibilities—as is amply proven by the examples of Alexander Pope, David Hume, and Samuel Johnson in England; Montesquieu and Voltaire in France; and Gotthold Lessing, Immanuel Kant, and Friedrich Klopstock in Germany. Jean-Jacques Rousseau was much absorbed in the reading of Plutarch, and the republican civic feeling of the French revolutionaries never ceased to imitate textbook oratory or to depend upon ancient examples.

Desacralization and rediscovery of antiquity. But while the classical humanities retained their status as teaching models, the progress being made in philological studies conducted on a scientific level was little by little eroding the system of myths and legends on which the model rested. The critiques of the scientists destroyed the pretty pictures perpetuated by the professors for their pupils. As portrayed by the teachers, the distant past melted into a single panorama; the heroes, playing out their roles in the same settings and in the same costumes, seemed as interchangeable as the characters of tragedy. From the time of the Renaissance, however, another school of Humanists had been rediscovering actual antiquity in its historical reality. Archaeology challenged the classicism of the professors, re-examining all of the trite old images and received ideas.

Parallel between classical and biblical scholarship

There is a striking parallel between the evolution of classical scholarship and that of biblical scholarship. In both areas, a venerable tradition with canonized texts and themes, supported by temporal authorities, dictated an overall view and an understanding that were fixed once and for all. But the biblical literature, translated into the Latin Vulgate of Jerome, was called into question again at the time of the Reformation, when scholars insisted on a return to its Hebrew and Greek sources. This thaw in the studies brought with it a new reading of the received texts, a critique of their authenticity, and a revision of their interpretation. The work done on the Old Testament by Benedict de Spinoza, a mid-17th-century Rationalist philosopher and a good Hebraist, and on the whole of Scripture by Richard Simon, a Catholic religious of the later 17th century, brought about a sort of desacralization of the holy text, which—even if it had been inspired by God—carried many traces of historical alterations and human intervention. The same method of philological criticism accomplished a concomitant desacralization of classical writings. The ideal of belles lettres had hidden and distorted the authenticity of texts, works, and men. Thus, the cult of a transcendental revelation of the True, the Beautiful, and the Good in its Greco-Roman guise had to give way to the demand for a historical investigation ready to unmask the best established prejudices and conventions and to lay open positive knowledge of the ancient civilizations.

This patient labour of restitution found its place not in classroom teaching but at a higher level, in universities, where the masters were free to devote a major part of their efforts to research. Inasmuch as the universities of Catholic countries had been held suspect and paralyzed since the Council of Trent (1545–63), which had crystallized the doctrines and reforms of the Counter-Reformation, the Protestant universities would have to serve as the seats of development of modern philology. At Leiden and elsewhere in Holland, numerous Protestant scholars were at work as indefatigable researchers and editors of texts, dictionaries, encyclopaedias, and literary and historical works. Their work renovated the image of antiquity in the complexity of its ages and forms.

Classical research at Dutch and German universities

The 18th century saw the supremacy of the German philologists at the old University of Leipzig as well as at the modern universities of Halle and Göttingen, where several notable scholars edited and commented upon the ancient texts, facilitating their comprehension by every

possible means, and brought about the institution of a seminary of philology in which—in an atmosphere of critical erudition and exact science—professors would be trained who were moved by a new spirit.

Philology and humanistic scholarship as modern disciplines. These studies achieved the status of a modern discipline at the close of the 17th century.

The founding fathers. In England the great name of enlightened philology was that of the royal librarian Richard Bentley, a dignitary both in the church and in science and also a master of Trinity College in Cambridge. Bentley established his reputation during the years 1697 to 1699 by demonstrating, by rigorous methods and in one stunning stroke, that the *Epistles of Phalaris*, a traditionally respected text that depicted the cruel tyrant of ancient Acragas (in Sicily) as a humane and cultured man, were a spurious counterfeit of relatively recent date. An indefatigable (though occasionally mistaken) worker, Bentley took as his field all of Greek and Roman literature. His most remarkable intuitions appeared especially in his Homeric studies. Having worked on biblical texts as well, Bentley drew a parallel between "Moses," author of the Pentateuch (the earliest books of the Bible), and "Homer," author of the *Iliad* and the *Odyssey*. In each instance the work attributed to an author designated by name had to be understood as a regrouping, at a certain time in history, of the elements of an earlier tradition. Behind the author's name lay in reality a collective work. Moreover, in order to resolve certain difficulties of ancient metrics, Bentley restored to the Homeric text a letter, *digamma*, which had been dropped since the classical period. This discovery was one of the starting points of historical linguistics.

Bentley's true successor in this work, at the end of the 18th century, was the German philologist Friedrich August Wolf. German historians like to assign the birthdate of modern philology to April 8, 1777, the day when young Wolf was accepted, not without difficulty, into the University of Göttingen as a *studiosus philologiae;* this was an unprecedented departure, for, until then, philology had not been an independent course of studies; only students in theology had attended the courses of the seminary of philology. Wolf, however, considered philology to be a discipline sufficient unto itself. But his enrollment was only a symbolic gesture; his real glory was assured when he wrote the *Prolegomena ad Homerum* (1795), which transferred Homeric studies from the realm of legend to that of science. The mythical personage of the old blind bard of Ionia was definitively rejected; and the texts that he was supposed to have created were now seen to be compilations based on pre-existing elements, which throughout history had been subjected to complex revisions, elaborations, and alterations in transmission. It is true that the theory itself had already been advanced by Bentley and that the French Hellenist J.-B.-G. d'Ansse de Villoison had discovered evidence that reinforced it. But Wolf radicalized the theory: he presented a demonstration that, going back over the earliest steps of the historical elaboration of the text, forever destroyed the conventions established by the respectful veneration of centuries of Humanism. The scandal was such among the adherents of traditional belles lettres that Wolf's theory was given the name of "Homeric atheism."

New spirit in contemplating antiquity

Wolf's critical work inaugurated the era of modern positive philology. It obliterated the conventional image of antiquity that had been perpetuated in the schools and helped to provide a new understanding of the origins of mankind and of the ancient epochs of culture. The Homeric epic was seen as perpetuating the memory of the earliest times of Greek civilization. As early as the beginning of the 18th century, the Neapolitan thinker Giambattista Vico, in his *Scienza nuova* (1725, expanded in 1730 and 1744), was exalting the heroic eras of human society, of which the *Iliad* was one example among others. Although Vico's book had not been widely read, Enlightenment scholars, enthralled by the discovery of savages, in whom they saw a reminder of the beginnings of mankind, established a relationship between these primitives and the persons of the ancient epics, whether classical or barbarian. The popularity of certain old Germanic texts that recently had been uncovered inspired the imitations that the Scottish poet James Macpherson made and falsely attributed to the early Gaelic bard Ossian: *Fingal. An Ancient Epic Poem* (1762) and *Temora* (1763). The immense success of these compositions throughout Europe led to a new reading in which the infancy and youth of the Hellenic world came alive again. As early as the *Geschichte der Kunst des Altertums* (1764; "History of Ancient Art") of the art historian Johann Joachim Winckelmann, there was evidence of a new spirit in the contemplation of ancient works of art. The romantic vision of the world introduced a breath of life into the interpretation of the past, which had to be recaptured in the authenticity of its presence. Poetic license might often interfere with the methodological rigour of philological science, but at the same time it would preserve these studies from the funereal dullness of erudition. The archaeological passion of Heinrich Schliemann (died 1890), excavator of Troy and Mycenae, was romantic in its essence.

The development of Sanskrit studies. Another decisive development in the history of humanistic scholarship was brought on by the study of Sanskrit language and literature, beginning at the end of the 18th century. English Orientalists, encountering the culture of India, became conscious of the fact that the languages of Europe derive from Sanskrit. Beginning in 1784, two British Orientalists, William Jones and (later) Henry Thomas Colebrooke, were responsible for creating an enthusiasm for Orientalism, shared subsequently by many scholars, including the German Friedrich von Schlegel, a Romantic critic. Philology was thenceforth to be a comparative science; thus, classical philology became a branch of Indo-European philology. This was the consummation of the desacralization of Greek and Latin, which ceased now to have an absolute value in themselves and to warrant the respect due to mother tongues. The classical civilizations, now merely two civilizations among others that had also flourished, had to be transferred from an ideal realm to one of historical relativity. Classicism could no longer be viewed as an eternal affirmation of the True, the Good, and the Beautiful; it was now merely an abstraction drawn from the study of certain works produced by the geniuses of a certain place and time.

The final desacralization of Greek and Latin

Specialization and expansion as a technology. Thus deposed from their dignity as an artificial paradise, the classical humanities became, in the 19th century, an ensemble of specialized disciplines, enclosed more and more within their own respective technicalities. German science, with the University of Berlin as its privileged locus, now played a preponderant role. Modern scholars ceased being all-around or general scholars; they became specialists either in Greece or in Rome. And they often narrowed their interests down to a certain period within a civilization. Such a division of labour also applied to the various ways of approaching the study of ancient culture. Henceforth there were often separate departments devoted to literary history, philological criticism of texts, the history of art, political and social history, archaeology (with its adjuncts—such as epigraphy, the science of inscriptions, and numismatics, the study of medals and coins), the history of ancient law, the history of ancient philosophy, and even prehistory, which meant doing research *in situ* on the origins of classical civilizations. Scholars became more numerous, and the great cultural powers increased the numbers of scholarly journals, specialized institutes, and schools of archaeology in Greece, Italy, and the Middle East. This technological expansion was accompanied by a fragmentation of knowledge, however, so that no one today can claim to have a full knowledge of the whole domain of humanistic scholarship. For a review of the humanities as they have existed in the late 19th and 20th centuries, see the article HUMANITIES.

CONCLUSION

The unity and identity of humanistic scholarship are now finally shattered. The myth of antiquity, propagated by

the venerable ideal of belles lettres, remains a forgotten dream. Thus disappeared the program of *paideia*, which, from the Sophists and Socrates onward, had inspired pedagogic uniformity in the West. Unfortunately, this outdated and discredited program has not been replaced by anything else; modern education is in a crisis and has been seeking in vain for a new foundation on which to base the training of individuals in contemporary society. Some theoreticians have attempted to create a substitute for classical studies in "modern humanities," based upon the exact sciences and on techniques. But the positive sciences do not supply that sense of peculiarly human values that was derived from the study of ancient literatures. The present-day proliferation of theories of pedagogy has too often developed within an abstract space from which the face of man has been banished. This pedagogic malaise reflects the crisis of civilization.

The need for a broader humanism

Humanistic scholarship had made the human form the centre and measure of all things. When man ceases to be the principal concern of mankind, barbarism threatens. It is now known that classical studies deal with only one civilization among many. But, even thus relativized, it must still be admitted that Hellenic culture was an unusual success, though it is no longer the only success. Perhaps the solution to the present crisis may lie in a generalization of classical humanism. Instead of tying all teaching to the theme of a conventionalized image of the Greco-Roman world, the effort could be made to reconstruct the various forms of human presence upon the face of the earth. The great civilizations and those not so great would all have a place within such studies, and humanistic scholarship, thus made relative, would become one chapter in the overall study of anthropology. In order to bring together tradition and modernity and to merge classical culture and universal culture, the time has come now for a new updating of the study of man.

BIBLIOGRAPHY

General works: JOHN EDWIN SANDYS, *A History of Classical Scholarship*, 3rd ed., 3 vol. (1958), still the most complete study of the history of classical philology; GEORGES GUSDORF, *Les Sciences humaines et la pensée occidentale*, 6 vol. (1966–73), a comprehensive history of Western culture, including humanistic scholarship; GEORGE SAINTSBURY, *A History of Criticism and Literary Taste in Europe from the Earliest Texts to the Present Day*, 3 vol. (1900–04, reprinted 1961), also a very reliable work. Shorter, but very useful, is ALFRED GUDEMAN, *Grundriss der Geschichte der klassischen Philologie*, 2nd ed. (1909); see also ULRICH VON WILAMOWITZ-MOELLENDORFF, *Geschichte der Philologie*, new ed. (1959), a sketch by one of the German masters in classical studies.

Old textbooks by German scholars of the 19th century: FRIEDRICH AUGUST WOLF, *Vorlesungen über die Altertumswissenschaft* (1831–35); AUGUST BOECKH, *Encyklopädie und Methodologie der philologischen Wissenschaften* (1877).

Historical periods: (*Greek period*): WERNER JAEGER, *Paideia, die Formung der griechischen Menschen*, 2nd ed. (1936; Eng. trans., *Paideia: The Ideals of Greek Culture*, 3 vol. 1939–45), a masterpiece; WILHELM NESTLE, *Vom Mythos zum Logos* (1940); HENRI I. MARROU, *Histoire de l'éducation dans l'antiquité* (1948; Eng. trans., *History of Education in Antiquity*, 1956), very useful; SIR JOHN L. MYRES, *Homer and His Critics*, ed. by DOROTHEA GRAY (1958), a history of the Homeric studies. (*Middle Ages*): ERNST ROBERT CURTIUS, *Europäische Literatur und lateinisches Mittelalter* (1948; Eng. trans., *European Literature and the Latin Middle Ages*, 1953), a wonderful study of the beginnings of European culture; C.H. HASKINS, *The Renaissance of the Twelfth Century* (1927), and G. PARE, A. BRUNET, and P. TREMBLAY, *La Renaissance du XII° siècle* (1933), are good surveys of the time; M.D. CHENU, *Introduction à l'étude de saint Thomas d'Aquin* (1950), a cultural prolegomena to the study of Thomism. (*Renaissance*): ANDRE CHASTEL, *Art et humanisme à Florence au temps de Laurent le Magnifique* (1959); and the old but ever valuable work of JACOB BURCKHARDT, *Die Kultur der Renaissance in Italien* (1860; Eng. trans. from the 15th ed., *The Civilization of the Renaissance in Italy*, 2nd ed., 1890). (*Since the Renaissance*): MAX WENGER, *Altertumskunde* (1951), a selection of texts concerning the discovery of Greece and Rome since the revival of learning; CONRAD BURSIAN, *Geschichte der klassischen Philologie in Deutschland* (1883), old but still valuable; PETER GAY, *The Enlightenment: An Interpretation* (1967), shows the permanent influence of classical scholarship on most of the 18th-century thinkers.

National histories of classical studies: FRIEDRICH PAULSEN, *Geschichte des gelehrten Unterrichtes auf den deutschen Schulen und Universitäten*, 2 vol. (1896–97); EMILE EGGER, *L'Hellénisme en France* (1869).

Biographies of outstanding scholars: PRESERVED SMITH, *Erasmus: A Study of His Life, Ideals and Place in History* (1923); JOHAN HUIZINGA, *Erasmus* (1924; Eng. trans., 3rd ed., 1952); PIERRE DE NOLHAC, *Pétrarque et l'humanisme*, 2nd ed., 2 vol. (1907); R.C. JEBB, *Bentley* (1887); KARL JUSTI, *Winckelmann und seine Zeitgenossen*, 4th ed., 3 vol. (1943); MAX HOFFMANN, *August Boeckh* (1901), in German.

(G.P.G.)

Humanities

The humanities is a term that refers to one of the administrative divisions of the college and of the graduate school of arts and sciences, as organized especially in the large, modern American university.

IDENTITY AND SCOPE OF THE HUMANITIES

This division brings together those educational disciplines not included within the divisions of the natural sciences and of the social sciences and combines with them to comprise the arts and sciences taught in the nonprofessional schools of the university (outside of such schools as those of law and medicine). In a less pragmatic but also less definite sense, the term is used for a division of knowledge itself apart from the simple concern for administrative convenience. The humanities are considered to constitute a distinct kind of knowledge that is humanistic—*i.e.*, which is concerned with human values and expressions of the spirit of man—and are thus appropriately grouped together in one administrative division separate from that of the sciences. It has been found difficult, however, to establish identifying criteria for the humanities, and disagreement has frequently occurred about whether or not a given discipline, such as history or the fine arts, should be included within the humanities.

The wide scope given to the term can be seen from the language used by the United States Congress in the law establishing the National Endowment for the Humanities, which declares that

> The term "humanities" includes, but is not limited to, the study of the following: language, both modern and classic; linguistics; literature; history; jurisprudence; philosophy; archeology; the history, criticism, theory, and practice of the arts; and those aspects of the social sciences which have humanistic content and employ humanistic methods.

This enumeration, however, includes a greater diversity of subjects than is found in the humanities divisions of even the largest universities.

Alternative criteria of the humanities

Although it is not unusual for the humanities to be identified with certain disciplines and subject matters, it is often argued that such a definition is inadequate. Some subjects so identified, it is said, may be pursued and studied in a way that has nothing humanistic about it—so the study of language often carried on in linguistics belongs to mathematics and science rather than to humanities; and on the other hand, subjects not identified as humanities may be studied in a humanistic way—as science is when its history and philosophy is the object of study. This being so, the criteria for distinguishing the humanities must be sought elsewhere than in the nature of subject matters. Some locate their characteristic feature in the fact that they are arts and methods of analysis or of criticism that, because they transcend all subject matters, can be applied to any subject, while others claim that their distinguishing mark lies rather in the end, or purpose, that the scholar has in view in the study and pursuit of his subject, whatever it may be—a method or end that somehow is concerned about human values. Still others hold that they are to be distinguished from other disciplines by the language that they employ, by the special faculty of mind that they use (*e.g.*, the imagination), or by the kind of experience on which they are based and to which they appeal, which is usually particular and nonscientific. Each of these positions or approaches has been taken as a basis for elaborating a general theory of the humanities.

THE HUMANITIES AS AN EDUCATIONAL PROGRAM

Humanitas or *paideia*

The term humanities derives from the doctrine of *humanitas* that Cicero developed in an account of the education of an ideal orator. In Latin, *humanitas* may mean "humanity" or "humaneness"; but as Cicero used it in his work *On the Orator*, it also refers to a special kind of educational program. In fact, the authority of a later Latin grammarian, Aulus Gellius (2nd century AD), is cited for identifying it with the Greek *paideia*—*i.e.*, the general and liberal education used in preparing a free man for manhood and citizenship, which developed from its origin in the mid-5th century BC, when the first itinerant Sophists offered private tutoring to young men in the Greek city-states. As further developed by Greek and Roman rhetoricians, it constituted the basic program of classical education. Then, turned to Christian ends by St. Augustine and other Church Fathers, it became the basic education of the Christian Middle Ages. Thus, *humanitas*, or *paideia*, was originally identical with a program comprising studies called *artes*, *bonae artes* ("good arts"), or *artes liberales* ("liberal arts"), which included mathematical and linguistic arts, as well as some science, history, and philosophy.

The term *humanitas*, which dropped out of educational parlance during the Middle Ages, was revived in the 15th century by Italian Humanists, so called because they claimed to devote themselves to, and teach, *studia humanitatis* (the study of *humanitas*, or the humane disciplines). Coluccio Salutati, a Florentine chancellor who may have been the first to use the term, contrasted it with *studia divinitatis* ("divine studies"). The program of studies pursued by the Humanists included grammar, rhetoric, poetry, history, and moral philosophy, studied in the language and literature of the ancient Greeks and Romans. Such emphasis was placed upon imitating the language and models of the ancient classics that *studia humanitatis* came to mean little more than the study of Latin and some Greek, and, as such, it was criticized and attacked by educational reformers, such as D. Diderot and J. d'Alembert, the 18th-century French Encyclopaedists (see in their *Encyclopédie*, under "College"). The program of classical studies, often combined with mathematics, provided the basic liberal education of the 19th century in the American as well as the English college and, on the European continent, in the *lycée* and *gymnasium*.

General and liberal studies

The fact that, in their origin and for a great part of their long history, the humanities stood for an entire educational program helps explain several features common to the contemporary discussion of the humanities. A program of general and liberal studies has no hard and fast lines demarcating the subjects to be studied from those that are not. According to time and place, different studies have come to be included, and the priority of emphasis has varied from one discipline to another. The program proposed by Cicero as well as by St. Augustine included mathematics and some science, along with the verbal arts of grammar, logic, and rhetoric; whereas the programs recommended by the Renaissance Humanists concentrated upon the study of language and literature. For all, however, the end envisioned was to train a man to be skilled and raised to the peak of his capacities in all that is most distinctively human. It was a program of "more humane letters," as indicated by the title of the Oxford classical degree in litterae humaniores. Although this program might be highly specialized in some respects, its teachers claimed to be educating a person in the best that had been thought and said on almost all of the great concerns of man. As comprising the elements of a liberal education, the humanities are as much a subject of dispute, and provide as wide an area of disagreement, as liberal education itself, and invoke much discussion that is inseparable from issues in general and liberal education; thus, the relevant discussions in the various education articles should also be consulted.

THEORIES OF THE HUMANITIES

Theories of the humanities are usually developed on a basis somewhat narrower than that of the elements of an educational program; thus, the humanities are now considered and analyzed as a fundamental division of the world of knowledge. The Renaissance Humanists, as already noted, contrasted them with divine studies; but since the 19th century they have more commonly been opposed to the sciences.

Nineteenth-century theories. Indeed, the first modern attempts at elaborating a general theory of the humanities are seen in the efforts of certain 19th-century German philosophers to stake out and defend certain areas of knowledge as lying outside of, and beyond, the reach of the natural sciences. Wilhelm Dilthey, a pioneer in this endeavour, called these studies *Geisteswissenschaften*, sometimes translated as "human sciences"; whereas a Neo-Kantian philosopher, Heinrich Rickert, preferred the term *Kulturwissenschaften*, or "cultural sciences," which has also been translated as "humanities." (Thus, the *Logik der Kulturwissenschaften* [1942] of another Neo-Kantian, Ernst Cassirer, was published in English translation as *The Logic of the Humanities* [1961].)

Dilthey's *Geisteswissenschaften*

Rickert held that method, not subject matter, provides the distinguishing feature of the cultural sciences. Unlike the natural sciences, which are "nomothetic" in that they aim to reach general laws, these studies are "idiographic" in that they are concerned with what is individual and with its unique value. According to Dilthey, on the other hand, it is subject matter that makes the decisive difference: the human studies have for their objects the work and actions of men, and these, he claimed, are fundamentally different from natural phenomena and call for a different kind of analysis and explanation. All merely natural entities can be explained from without, as it were; indeed, they must be, since it is impossible for man to know from within what it is, for example, to be a stone or a cat. The works of men are essentially different in that all of them are expressions of human purpose and cannot be fully accounted for unless they are understood as such. To be sure, man is a part of nature and can be studied as such by the natural sciences; but, in addition, man is also a product of his culture and of his own drives and choices, and in these respects he cannot be known without an understanding and comprehension of purpose and meaning.

So far, however, these German scholars have provided no criterion for distinguishing the humanities from the social sciences. In fact, the work of these same scholars, and especially that of Dilthey, has contributed importantly to the development of the social sciences. For this reason, the term human studies, as used in the 19th century, might better be taken to mean not only the humanities but also the social sciences, at least in certain respects, and the term is still used in this sense in continental Europe. More recent theories, on the other hand, are as much concerned to separate the humanities from the social sciences as from the natural sciences.

Twentieth-century theories. Of these theories, four can be singled out as representative of the work done in the third quarter of the 20th century: those that view the humanities as (1) constituting certain general arts and methods applicable to any subject matter; (2) employing and emphasizing certain language function over others; (3) based on a distinct faculty of mind; and (4) based on, and appealing to, a general experience available to any man and not just to the specialist. Each of these theories represents a distinctive approach to the humanities and provides a developed account and defense of them as constituting a unified field of study distinct from both the natural and the social sciences. On one proposition all four theories are in basic agreement, viz., that the humanities form an important and valid area of knowledge distinct from that of the sciences. They differ, however, concerning the characteristics that make the humanities distinctive and set them off from the sciences, and concerning their relation to the sciences. In the special faculty and general experience theories, there is a radical separation, even an opposition, between the two; whereas in the general arts and language function theories, there is continuity and a difference of emphasis rather than a decisive break.

Theories denying a radical difference between the humanities and the sciences. According to the general arts theory, the humanities cannot be identified with any one special method, subject matter, or end inasmuch as they transcend all and yet apply to all. They are "serving sciences," claimed Ronald Crane, a proponent of this view, since they are

General arts theory

> oriented to uses and ideals which can be stated apart from the specific nature of the subject matters or objects of study by which the uses and ideals are to be attained.

Thus, of any subjects whatsoever, a scholar can discuss and analyze the language and literature, the history, and the philosophy. As applicable to a multiplicity of subjects and endeavours, the humanities might thus be looked upon as general arts or techniques.

There are four principal groups of humanistic methods or arts, according to Crane: the arts of language for coping with symbolic expressions; arts of the analysis and appreciation of ideas; arts of literary and aesthetic criticism; and arts of knowing the historical situations. Although these arts pertain to all subjects, they predominate more in some than in others; and such subjects are more humanistic because in them the arts come closer to providing a complete account of their nature.

There is no sharp separation, in this view, between the humanities and the sciences. The sciences, especially in the teaching of them, have to employ the humanistic arts. The difference between them must be sought rather in the direction in which they move in providing analysis and explanation: the sciences move from diversity and particularity toward unity, uniformity, simplicity, and necessity; whereas the humanities stress uniqueness, unexpectedness, complexity, and originality.

According to the general arts view, it is no accident that the humanities are so named: they have the capacity to deal with those aspects of experience that are most distinctively human and are not fully accounted for by the natural processes and social forces and structures to which the sciences appeal. As addressing what remains after the sciences have rendered their account, the humanities are concerned with the nature and value of human achievement—with "that kind of unprecedented excellence," as Crane has expressed it, "that calls forth wonder as well as admiration"—and hence are ineradicably normative as well as descriptive.

Language function theory

According to a second theory, the best approach to an understanding of the humanities lies in an analysis of language. Language as used in any utterance or expression is, in this view, a highly complex structure that satisfies many diverse needs and has many different functions. William T. Jones, for instance, an advocate of this view, names a few typical functions: language, he claims, "may satisfy curiosity, convey information, complement personal deficiencies, relieve anxieties, obviate loneliness." For certain purposes, it may be useful to isolate and consider just one of these functions—for example, the cognitive function of providing information and knowledge. But such an isolation, Jones emphasizes, is an abstraction and simplification, since

> it is not true that some particular segment of discourse satisfies a pure cognitive function, that this other piece of discourse satisfies a pure emotive function, that this third piece of discourse satisfies a pure appeal function;

the most that can be said is that in any given instance one function is more dominant than another. Since language simultaneously fulfills many different functions, it should be envisioned, according to this theory, as forming a spectrum or continuum, ranging from the cognitive, designative function at one end to the noncognitive and expressive function at the other.

It is claimed that reference to this linguistic model provides the best means of distinguishing the humanities from the sciences. Though no use of language ever satisfies only one function, language can be directed more to one than to another. The sciences are characterized by a language that is more designative and cognitive than expressive, the humanities by a language that tends toward the expressive, evaluative, and noncognitive. The sciences and humanities are different languages with different functions. Hence, each has its own standards of adequacy and value and should be judged accordingly. Neither is truer or better or more valuable than the other, but each serves in its own way to form man's view of reality. To be sure, the sciences and humanities are, on this basis, admittedly opposed in many respects to each other. But it is claimed that this opposition rests on the misunderstanding and misapprehension that comes from attempting to judge one by the standards of the other; in principle, they can be reconciled within one world, since they do not contradict but supplement each other.

Theories asserting a radical difference between the humanities and the sciences. In the special faculty and general experience theories, the sciences and humanities may supplement each other, but they do so because they belong to radically different, even antagonistic, ways of approaching and understanding the world. In the special faculty view, they reflect a basic and antagonistic division—in the words of Albert W. Levi, "a real dualism, . . . a split within the structure of the human mind"; they belong to different faculties of the mind, operate with different sets of concepts, and express themselves in different forms of language. The sciences are the expressions and product of the rational understanding, employing concepts such as fact, law, cause, chance, and prediction, communicated through an impersonal, referential, objective language; and the humanities, in contrast, are the work of the imagination, which has its proper concern with the things of man through such concepts as appearance and reality, destiny and free will, fortune, fate, happiness, peace, and tragedy, expressed in a language that is dramatic, emotional, and purposive. Each has its own preferred mode of operation, its own values, its own particular contribution to man's knowledge of himself and the world; yet the two are incommensurable as representing basically discordant approaches to human experience.

Special faculty theory

The humanities, in this view, are arts and approaches rather than subject matters. In fact, they are identified with the liberal arts, but these arts are redefined as those of communication, continuity, and criticism, which, in turn, can be correlated most closely with the disciplines of language and literature, history, and philosophy. Levi claims, however, that they must not be identified with these disciplines as such, since to each of them there is also a scientific side. Language, for example, studied and used as an art of communication, is a liberal art; taken as a physical phenomenon and analyzed mathematically, however, it is a scientific study. In this view, no subject or discipline is intrinsically scientific or humanistic; there is a humanistic tendency in every science and a scientific tendency in every humanity.

The world of learning thus divides into the arts on one side and the sciences on the other. Each of these also divides into two: the arts into the fine arts, concerned with making, and the liberal arts, which are those of learning and teaching; and the sciences into the natural and social sciences, concerned, respectively, with natural and cultural phenomena.

General experience theory

In the general experience theory, the characterizing feature of the humanities is their everyday, "natural" logic. The mathematical form of logic that has proved highly successful in mathematics and the sciences has been of little help to the humanities. Though some would conclude that these disciplines fail to achieve any real knowledge worth the name and amount to little more than a gratification of taste, defenders of this fourth theory conclude instead that they operate with, and appeal to, a different kind of logic, which they call a "what logic" because it deals mainly with the "whats" and "whys" of everyday experience, as contrasted with the "relating logic" characteristic of science, in which the main interest lies in relations, patterns, and structures.

The difference between the humanities and sciences, in this view, is not restricted to matters of method but extends also to the kinds of concern that they reflect and the objects of these concerns, as well as to the source of the evidence to which they appeal. Unlike the sciences,

the humanities are based on the common experience of men that is open and available to all; and they aim, according to the champion of this view, Henry B. Veatch, at "simply a knowledge of man, of human nature, of ourselves, and still more broadly of our human situation and of the real world in which we find ourselves." The data of the humanities are given in ordinary experience, in which nothing is so obscure, recondite, or uncommon as to call for specialized investigative techniques and highly artificial and contrived experiments. Such common experience is primordial; it is the most basic experience of all, underlying all of man's knowledge, constituting what he knows best, and providing him with the criteria of his surest knowledge. These features also explain why the humanities, unlike the sciences, continue to cherish and use the great works of the past: as monuments of human achievement, they speak to, and reflect, a concern with elements that are permanent in human experience and thus retain an abiding relevance.

Clearly, these four theories represent different approaches to the humanities and diverse ways of defending their value and importance. The conflicts among them reflect differences more of emphasis than of irreconcilable opposition and rest ultimately upon different philosophies. The patterns of agreements and disagreements among these theories help point out where the major problems concerning the humanities are to be found.

PROBLEMS ABOUT THE HUMANITIES

The role and function of the humanities in education. It is in their role as an educational program that the humanities have the longest tradition, and it is upon this basis, too, that the proponents of the humanities are most agreed, since they hold that the humanities should provide the basis of a general and liberal education. Education in the humanities is nonvocational, nonprofessional, aiming at the maturation of the person as man and citizen, not as a worker in some specialized field; and, for that reason, humanistic education is concerned with preserving and developing the arts and skills that find their expression in the great objects, problems, and values of human interest.

The study of language and literature has long been a strong point in education in the humanities. There are disagreements, however, regarding the form that such a study should take. Greek and Latin have long lost the monopoly that they once enjoyed; and, although emphasis is still placed on the expressive as well as on the cultural value of acquiring some command of a foreign language, it is ceasing to be required as extensively as it once was in the college curriculum. The proponents of foreign language study hold that it is needed not only to gain an entry into another culture and to study its literary monuments at first hand but also to expand the mind beyond the limited scope of one language and especially to gain a knowledge of language so as not to mistake words for things and become the dupe of one's native tongue. Against this view it is argued that especially in the United States, where there is little need for any language but English, the time and effort spent on a foreign language in high school and college are largely wasted as judged by the meagre results that are achieved; any command of a foreign language is much more readily obtained by living where the language is spoken, and hence "study years abroad" have been instituted by many American universities. The classics of Greece and Rome —in translation now rather than in the original—continue to exert their pervasive influence. The study of literature, however, has been broadened to include the major achievements in other languages; and comparative and world literature now command considerable interest.

The extensive use of "great books" in humanities programs represents not only the perennial importance of literature but also a reaction against an overspecialized and professionalized approach to its study. Literature, understood in an extended sense as including history, philosophy, and theology as well as imaginative literature, becomes a meeting place of subject matters and hence of interests that extend beyond any one academic department. So, too, unified courses in the humanities are sometimes offered as a better way of meeting the needs of liberal education than through departmental courses in single disciplines. The demands of students for "relevance" in the courses that they take are often, in part, a protest against the professionalization of the curriculum that converts it into preparation for the academic profession. When taught and studied professionally as preparation for an academic career, literature, history, and philosophy can be as specialized and narrow as any professional course in science. For the humanities no less than for the sciences, there is need for studies and courses for the nonspecialist and nonprofessional student.

A concern with expression and aesthetic value is generally agreed to be a feature of the humanities. Much disagreement exists, however, regarding the way of studying and of teaching this feature. At issue, in particular, is the place and role of the practice of the fine and performing arts. Some teachers maintain that the actual practice of these arts is a necessary and essential part of the study of the "expressive function." Others claim, however, that, so far as their humanistic values are concerned, the fine arts, on their practical side, are in the same position as the sciences; for both, it is only the study of their history and philosophy and of their contribution to, and their place in, the total human achievement that belongs to the humanities.

Interdisciplinary graduate programs

The interdisciplinary interest that has long been characteristic of the humanities has received renewed emphasis in recent years in the development of graduate programs in the humanities. Although these programs have different names—such as "humanistic studies," "joint program in the humanities," "institute of the liberal arts," "humanities program for the preparation of college teachers"—all have two features in common: one is the claim that humanistic disciplines, such as history, philosophy, and literature, while distinct as disciplines, are so closely related that the study of any one of them requires some insight into the other two; and the other is the belief that a broader study of this kind comprises a better preparation for undergraduate college teaching than a narrow and intensive specialization in one discipline can provide. At that level of education the student is endeavouring to discover his way and find his bearings within the world of learning. One of the requirements of the graduate program is, then, that he achieve some competence in mapping that world and showing what are the interrelations between the various areas of knowledge.

The humanities and the conflict of cultural ideals. The attention now being concentrated upon the humanities is characteristic of those times in history in which cultural ideals have come into open conflict. Such times occur when there are differing and opposed ideals about the kind of knowledge that is most important and about the form and method that express the highest cultural achievement. Western history contains a number of such conflicts: the fight that occurred in the ancient world between philosophy and poetry and that between rhetoric and philosophy, described, respectively, by Plato and Cicero; the struggle between pagan and Christian learning in the time of the Early Church and the battle of the liberal arts in the Middle Ages; the Humanists' attack upon Scholasticism in the Renaissance; the battle of the ancients and moderns in the 17th and 18th centuries and of religion and science in the 19th century. In all of these conflicts, the humanities have been present in one guise or another.

Humanities versus the sciences

In the contemporary world, the conflict of cultural ideals has come to the fore as that between the humanities and the sciences. Perhaps the greatest public notice has come from the controversy aroused by the English scientist and novelist C.P. Snow, who, in the essay *The Two Cultures and the Scientific Revolution* (first published in 1959 but still continuing with new additions in 1970), described, criticized, and lamented the polarity that he observed between a literary and a scientific cul-

ture, each spurning the other but each also unable to communicate with the other. Though Snow called for establishing new lines of communication between the two disciplines to improve understanding in both directions, his own sympathies were unmistakably with the scientists. For a century and more, science has provided what might be called the cultural paradigm of the day, and the defense of the humanities has been, at most, a rear guard action. There have recently been increasing signs, on the other hand, of malaise and of doubt at the high price that is being paid for the achievements of science and of its offspring, technology, and the current interest in the humanities is one manifestation of this development.

BIBLIOGRAPHY

The recent sources: The most recent full-length accounts of the humanities as a unified field of studies are those of R.S. CRANE, "The Idea of the Humanities" and "Shifting Definitions and Evaluations of the Humanities from the Renaissance to the Present," in *The Idea of the Humanities, and Other Essays Critical and Historical*, vol. 1, pp. 3–170 (1967); W.T. JONES, *The Sciences and the Humanities* (1965); A.W. LEVI, *The Humanities Today* (1970), and a longer philosophical justification of his position in *Literature, Philosophy, and the Imagination* (1962); and H.B. VEATCH, *Two Logics: The Conflict Between Classical and Neo-Analytic Philosophy* (1969).

The older German tradition: This tradition is well represented by ERNST CASSIRER, *Zur Logik der Kulturwissenschaften* (1942; Eng. trans., *The Logic of the Humanities*, 1961). H.M. JONES provides a good analysis of the social and practical importance of the humanities in *One Great Society: Humane Learning in the United States* (1959). One of the best of many symposia is *The Knowledge Most Worth Having*, ed. by W.C. BOOTH (1967).

The conflict between the sciences and the humanities: Much recent consideration of this conflict has resulted from the writing of the English novelist C.P. SNOW, *The Two Cultures and the Scientific Revolution* (1959), to which "a second look" was added (*The Two Cultures: And, A Second Look*, 2nd ed., 1964), which gave rise to a lively controversy, initiated by the English literary critic F.R. LEAVIS, *Two Cultures? The Significance of C.P. Snow* (1962), which is well documented in D.K. CORNELIUS and E. ST. VINCENT (comps.), *Cultures in Conflict: Perspectives on the Snow-Leavis Controversy* (1964). The controversy was taken up again in the columns of *The Times Literary Supplement* during the spring and summer of 1970 with essays by both Snow and Leavis and a host of letters in support of each. See also ALDOUS HUXLEY, *Literature and Science* (1963); and M.E. PRIOR, *Science and the Humanities* (1962).

Histories: Though many histories of individual disciplines exist, there is no general history of the humanities as such. Crane's book is the only one of those cited that provides considerable history—which, however, is devoted mostly to the modern period. For the ancient period the fullest accounts are to be found in WERNER JAEGER, *Paideia: The Ideals of Greek Culture*, trans. from German, vol. 1, 2nd ed; vol. 2–3, 1st ed. (1943–45); and H.I. MARROU, *Histoire de l'éducation dans l'antiquité* (1948; Eng. trans., *A History of Education in Antiquity*, 1956). For the medieval period, ETIENNE GILSON, *History of Christian Philosophy in the Middle Ages* (1955), with its very extensive bibliographies, provides many leads.

(O.A.B.)

Human Rights

When, in the last third of the 20th century, people speak of "human rights," "the rights of man," or "*les droits de l'homme*," they really mean those rights drawing their formulation from the last decades of the 18th century and the American and French revolutions. But the idea of the inalienable rights of the human being is much older and, in fact, was known to poets, philosophers, and politicians in antiquity and in the Middle Ages. When Sophocles' Antigone—the play was composed about 442 BC—says to King Creon:

> But all your strength is weakness itself against
> The immortal unrecorded laws of God

she invokes the higher law, the law of nature, the natural rights of man. Throughout the centuries there has been a close connection and interdependence between the idea of "natural law" and the idea of the natural rights of man. These ideas may be found in the works of the Stoics, both Greek and Roman, and in the teaching of early Christianity, St. Thomas Aquinas, and medieval English scholars of the law. They are encountered in the writings of Spanish theologian-lawyers of the 16th century and in the 17th century in the works of the Dutch Hugo Grotius, the founder of modern international law, and of John Milton and John Locke, the ideological architects of the English revolution of that century.

HISTORICAL ORIGINS

The first codifications. Ancient legal codes failed to recognize any area of individual freedom from state interference, and the first codifications of something akin to a catalog of rights—if not yet of all men, then at least of the nobles of the land—began to emerge in compacts between princes and feudal assemblies. One of the earliest of these came in 1188, when the Cortes, the feudal assembly of the Kingdom of Leon (on the Iberian Peninsula), received from King Alfonso IX his confirmation of a series of rights, including the right of the accused to a regular trial and the right to the inviolability of life, honour, home, and property. In the Golden Bull of King Andrew II of Hungary (1222), the King guaranteed, among other things, that no noble would be arrested or ruined without first being convicted in conformity with judicial procedure. The most famous and influential commitment of this kind was in the English Magna Carta, accepted by King John at Runnymede in 1215. Though it was exacted by his feudal barons in their own selfish interest and was by no means intended to assert rights and liberties for all, several of its provisions, among them the famous clause 39 stating that "no freeman shall be taken or imprisoned . . . or exiled or in any way destroyed . . . except by the lawful judgement of his peers or (and) the law of the land," gave expression to the idea of individual freedom and became the symbol of this freedom for centuries to come.

Magna Carta: the rights of free men

The English, American, and French revolutions. In 17th-century England the "immemorial rights of Englishmen" were successfully fought for, the landmarks of the struggle being the English Petition of Right (1628) and the English Bill of Rights (1689). The rights set forth in these instruments found their way to the New World colonies. The American Declaration of Independence (1776), the Virginia Declaration of Rights (1776), and the American Bill of Rights (1791) carried not only the ideas of the earlier English documents but in some cases their very text as well. Nor was it only England that exported its doctrine of rights; the powerful influence of the French philosophers of the Enlightenment is visible among all the American revolutionaries. If the French philosophes had helped provide the Colonies with an ideology, the Revolution of the Colonies, in its turn, gave the French an example to follow and had an enormous impact on the events in France. The French Declaration of the Rights of Man and of the Citizen (1789) was directly inspired by earlier American examples.

Constitutions of the 19th and 20th centuries. In the course of the 19th and 20th centuries, the example set by the United States and France adopting bills of rights or otherwise guaranteeing individual liberties was followed on the entire continent of Europe and eventually throughout the Americas, Asia, Africa, and the Caribbean. Great Britain, following the tradition of its unwritten constitution, did not enact constitutional guarantees.

The Russian Revolution of 1917, though following the American and French precedent in the form of its pronouncements, gave them a fundamentally different substance. The difference lay not in the emphasis on economic and social rights in addition to the traditional political and civil rights (for such provisions are found in other constitutions—in the Mexican constitution of 1917, the Weimar constitution of Germany of 1919, and the constitution of the Republic of Spain of 1931) but in the complete transformation of the meaning of political and civil rights.

This basic difference in concepts becomes clear in a

Human rights in the Soviet and U.S. constitutions

comparison of the provisions of the first Soviet constitution. The First Amendment to the American Constitution provides, among other things, that Congress shall make no law abridging the freedom of speech or of the press. The Russian constitution of 1918 "for the purpose of securing freedom of expression to the toiling masses" abolishes all dependence of the press upon capital and turns over to the working people and the poorest peasantry "all technical and material means for the publication of newspapers, pamphlets, books," etc. The U.S. Constitution prohibits abridgment of the right of the people peaceably to assemble, while the Lenin constitution, in order to ensure complete freedom of assembly to the working class and to the poorest peasantry, offers "all premises convenient for public gatherings together with lighting, heating and furniture."

The Soviet constitution of 1936, while changing the wording, has by and large maintained this general approach. Western constitutional ideas were designed to prevent interference with fundamental rights mainly, though not exclusively, by the public authorities. The Soviet concept ignores this aspect and promises to make available technical facilities without promising freedom in the choice of the purposes for which they will be used. In trying to draw conclusions from the existence in so many legal systems of catalogs of rights, these differences of basic concept must be kept in mind.

THE BEGINNINGS OF INTERNATIONAL CONCERN

If these were some of the important landmarks in the effort to enshrine respect for the rights and freedoms of the individual in the various national systems, it is necessary next to examine some of the history of efforts to maintain the human rights of the citizens and inhabitants of states by international action: the so-called humanitarian intervention and the conclusion of international treaties.

Humanitarian intervention. Under traditional international law, the sovereign state had discretionary power in the treatment of its nationals. When, however, the treatment meted out by a state to its own population, particularly to religious or ethnic minorities, was so arbitrary and so persistently abusive and cruel that it shocked the conscience of mankind, other states frequently took it upon themselves to threaten or even to use force in order to come to the rescue of the oppressed minority. A major example of such "humanitarian intervention" was the action, including military action, agreed on in 1827 by Great Britain, France, and Russia against the Ottoman Empire, assertedly to bring to an end the sufferings of the Greek population then under Turkish rule. This eventually led to the independence of Greece in 1830. Similar interventions were undertaken by several European powers to end massacres of Christians in Syria (1860), to bring relief to the persecuted population in Crete (1866–68), and, in the last third of the 19th century, to end the persecution by Turkey of Christian populations in various Balkan countries under Turkish sovereignty. Other examples of humanitarian intervention include the representations made in the last decade of the 19th century and in the first decade of the 20th by various Western powers on behalf of the Jewish population of Romania and representations made by the United States government to the Russian government on behalf of the Jewish citizens of that state.

Limitations of unilateral action

There is a danger of abuse inherent in humanitarian intervention because the intervening state or group of states decides unilaterally whether the intervention is justified and how to intervene. For this reason, the doctrine underlying the concept of humanitarian intervention has never become a fully acknowledged part of international law. Since the coming into effect of the Charter of the United Nations, with its prohibition of the unilateral use of force in international affairs, the status of humanitarian intervention has become even more precarious than it was in earlier periods.

International treaties. The use of international treaties for the protection of the rights of religious minorities can be traced back to the 17th century, when the Treaty of Westphalia (1648), which concluded the Thirty Years' War, established the principle that there should be equality of rights for both the Roman Catholic and Protestant religions in Germany. In the same century, various Catholic governments stipulated in peace treaties for the rights of Roman Catholic subjects of Protestant princes. In 1774 Turkey undertook, vis-à-vis Russia, to protect the Christian religion and its churches within its territory. The Congress of Vienna of 1815 provided for the free exercise of religion and for equality, irrespective of religion, in various cantons of Switzerland as well as for the equality of Christian denominations in Germany. The same congress also agreed on what probably was the first provision in an international treaty aiming at the improvement of the civil status of Jews. When, during the 19th century, Montenegro, Serbia, and Romania achieved their independence from Turkey, they, as well as Turkey itself, were forced to guarantee religious freedom and equality of rights to all inhabitants irrespective of religion (1878).

Prohibition of the slave trade. Throughout the 19th century and beginning with the Peace Treaty of Paris (1814), the universal prohibition of the slave trade had been an object of international effort and concern. Various treaty arrangements to this end were undertaken in 1815, in 1822, and again in 1862, 1885, and 1890. Gradually the movement undertook to combat and suppress slavery as well as the slave trade. In 1926 the Assembly of the League of Nations approved and opened for signature the International Slavery Convention, by which the contracting parties undertook to prevent and suppress the slave trade and to bring about, progressively and as soon as possible, the complete abolition of slavery in all its forms. Since then, the fight against slavery, which persists in one form or another, has continued under the auspices of the United Nations.

International social and labour legislation. Toward the end of the 19th century, a number of philanthropists, social reformers, and economists succeeded in arousing interest in, and enlisting the support of some governments for, the idea of international social legislation. To this end, a conference with a large and ambitious program of international measures of social reform and an impressive representation from European countries, including the great powers of the day, convened in Berlin in 1890. Despite the fact that the congress did not go beyond the adoption of some timid recommendations, its moral effect was nevertheless great. Once again on the initiative of private individuals and associations, governments were persuaded to meet in 1905 and 1906 in Bern, where the first two multilateral labour conventions, which were also the first international conventions for the protection of the human person, were concluded. One of the conventions prohibited night work for women in industrial employment, the other the use of the highly poisonous and inflammable white (yellow) phosphorus in the manufacture of matches. While the actual provisions of these conventions were not in themselves far reaching, the principle involved lent them a symbolic and moral effect that made them landmarks in the development of modern civilization.

The humanitarian law of war. It was also in the second half of the 19th century that the conclusion of multilateral treaties covering aspects of the conduct of hostilities and the protection of the rights of victims of war began. Due largely to the efforts of two Swiss citizens and a Swiss society that took up their ideas and later became the nucleus of what is now the International Committee of the Red Cross, the Convention for the Amelioration of the Condition of the Wounded in Time of War was signed at Geneva in 1864. It provided that military wounded and sick must be cared for whatever their country and that wounded who are captured must be sent home if they are incapable of further military service. The Declaration of St. Petersburg (1868) produced a consensus that the progress of civilization should be used to alleviate as many of the calamities of war as possible. To this end, the signatories undertook to maintain the prin-

Attempts at a minimum standard

ciples then established and to try to reconcile the necessities of war with the laws of humanity. These beginnings of the humanitarian law of war were further elaborated at congresses at Brussels (1874) and at The Hague (1899, 1907). Further legal developments followed during the interwar period (1925, 1929) and after World War II by the four Geneva conventions for the protection of war victims of 1949.

BETWEEN THE TWO WORLD WARS

The Covenant of the League of Nations. When, at the end of World War I, representatives of the victorious coalition assembled in Paris in 1919 to draft the peace treaties with Germany and its allies—treaties that were also to contain the Covenant of the League of Nations—Japan, one of the victorious powers, moved that a provision be inserted obliging the members of the League not to discriminate in law or in fact on the ground of race or nationality. This proposal met with stiff resistance on the part of the British and United States governments and was not approved. Japan later proposed the insertion of a provision on similar lines in the preamble of the Covenant. Although supported by important states, including France, Italy, and China, this proposal was also defeated. As a result, the Covenant of the League of Nations did not deal with or formally recognize the fundamental rights of man and did not lay down any principle of nondiscrimination. The Covenant, nevertheless, did contain certain elements suggesting a recognition of the rights and welfare of man on the international level. Among these were provisions by which members of the League (1) accepted an obligation to secure and maintain fair and humane conditions of labour for men, women, and children; (2) undertook to secure the just treatment of the native inhabitants of their colonies; (3) entrusted the League of Nations with general supervision of the execution of agreements regarding traffic in women and children; and (4) pledged themselves to take steps in matters of international concern for the prevention and control of disease. Those of the victorious powers who as mandatories were entrusted by the League with the tutelage of colonies formerly governed by Germany and Turkey accepted as "a sacred trust of civilization" responsibilities for the well-being and development of the inhabitants. The mandatory powers were thus made accountable to the international community represented by the League of Nations for their conduct of the mandates—an arrangement that has had repercussions even half a century later in regard to the mandate entrusted to South Africa over the territory of South West Africa.

The International Labour Organisation. The Treaty of Versailles and the other peace treaties ending World War I also established the International Labour Organisation, with part XIII of the Treaty of Versailles as its constitution. The aspect that distinguishes the International Labour Organisation from other international governmental organizations is the fact that its organs consist of representatives not only of governments but also of employers and workers. By and large, the structure and aims of the organization are the same now as when they were laid down in 1919. The International Labour Organisation has made a signal contribution to the promotion of human rights through international action not only in fields traditionally associated with labour law and labour relations (such as industrial health, safety, and welfare; hours of work; and annual holidays with pay), but, particularly after World War II, when it became a specialized agency of the United Nations, also in regard to matters at the very core of a system of guarantees of human rights, such as the abolition of forced labour, the elimination of discrimination in employment and occupation, freedom of association, and equal remuneration for work of equal value.

Labour and human rights

The protection of minorities under the post-World War I treaties. In some of the peace treaties, in special so-called minorities treaties, and in declarations made after World War I, a number of states of central and eastern Europe and one state in the Middle East (Iraq) were made to accept a series of obligations toward their racial, linguistic, and religious minorities: all of their nationals were to be equal before the law and were to enjoy the same civil and political rights without distinction as to race, language, or religion. Nationals who belonged to racial, religious, or linguistic minorities in these countries were guaranteed adequate facilities for the use of their languages before the courts and for meeting their educational needs.

The system of minorities protection instituted by these treaties was not general. Germany, although one of the defeated powers, was not among those required to accept provisions for the protection of minorities, apart from an obligation undertaken for a 15-year period in regard to the population of that part of Upper Silesia that at its partition in 1922 fell to Germany. All the relevant treaties provided that their stipulations constituted obligations of international concern rather than domestic matters, and all were placed under the guarantee of the League of Nations.

SINCE WORLD WAR II

Whatever progress had been achieved through the establishment of the League of Nations, the International Labour Organisation, and the other measures taken after World War I was wiped out by the horrors of Fascism, National Socialism, and other totalitarian systems, as well as by the ordeals of World War II itself.

In a series of statements and proclamations (among them the Atlantic Charter of August 14, 1941, and the Declaration by the United Nations signed by all the Allied powers on January 1, 1942), the preservation of human rights and justice was made one of the peace aims of the victorious grand alliance.

Trial of war criminals

At the end of the war, the victorious great powers (Great Britain, the United States, France, and the Union of Soviet Socialist Republics) established the International Military Tribunal for the Trial of German Major War Criminals. Under its charter, the tribunal had jurisdiction to try not only crimes against peace and war crimes but also "crimes against humanity" committed against any civilian population, whether or not in violation of the law of the country where perpetrated. The charter of the tribunal provided that the fact that the defendant acted pursuant to orders of his government or of a superior officer should not free him from responsibility. While the tribunal, which sat in Nuremberg from 1945 to 1946, applied a cautious and restrictive approach to the concept of "crimes against humanity," the treatment by a state of its own citizens was nevertheless made the subject of criminal proceedings in a court that functioned as an organ of the international community. The principles on which the tribunal established at Nuremberg acted were subsequently endorsed by the General Assembly of the United Nations.

Human rights in the Charter of the United Nations. The Charter of the United Nations (1945) contains the reaffirmation of "faith in fundamental human rights, in the dignity and worth of the human person, in the equal rights of men and women and of nations large and small." "Promoting and encouraging respect for human rights" and "assisting in the realization of human rights and fundamental freedoms" are words that appear, with some variations in different contexts, at several places in the Charter. There are two articles in which all members pledge themselves to take joint and separate action in cooperation with the organization for the achievement of universal respect for, and observance of, human rights and fundamental freedoms for all without distinction as to race, sex, language, or religion. The Charter is intentionally vague in its expressions on these subjects, and a proposal at the San Francisco Conference that the United Nations should ensure not only "the promotion" but also "the protection" of human rights was not accepted. Moreover, the Charter contains the often-quoted domestic jurisdiction clause, which says that nothing in it shall authorize the United Nations to intervene in matters that

are essentially within the domestic jurisdiction of any state. The reconciliation of the human rights provisions of the Charter with the domestic jurisdiction clause has created certain difficulties, giving rise to legal argument and political controversy.

The debate about obligations

Because human rights have traditionally come within the domestic jurisdiction of states and because of the vagueness of some of the human rights clauses, some authorities have concluded that in becoming parties to the Charter, states have not, in effect, accepted any obligations as to the human rights field and that the United Nations has no standing to insist on safeguards for human rights in member states. Others insist that the human rights provisions of the Charter are by no means devoid of an element of legal obligation, that the "pledge" made by member states cannot be ignored, and that the domestic jurisdiction clause does not apply because human rights have ceased to be "essentially within the domestic jurisdiction" of states.

It can be said, however, that in the actual practice of the various organs of the United Nations over the first 25 years of its existence, the difficulty of solving the apparent contradiction between the "human rights provisions" and the "domestic jurisdiction clause" has been far less formidable than the differences between the opinions of scholars and the statements by governments would lead one to assume. In practice, neither the vagueness and generality of the clauses of the Charter nor the domestic jurisdiction clause have prevented the United Nations from considering, investigating, and judging concrete human rights situations, provided there was a majority strong enough to force such action.

Though it is undoubtedly true that governments are usually jealous of their own domestic jurisdiction and of the authority of their allies and friends, the General Assembly and other United Nations organs have repeatedly dealt with human rights developments occurring within individual states, such as the race conflict in South Africa, forced labour in eastern Europe, respect for freedom of association and trade-union rights, and discrimination against the Buddhists in Vietnam (1963).

Whatever the merits of the opposing contentions as to the interpretation of the Charter may be, there has been general agreement from the beginning that an international bill of human rights should be prepared to supplement the human rights provision of the Charter. The Charter provides for the establishment of a Commission on Human Rights to be entrusted with the task of preparing such a document and other international treaties and instruments in this field.

United Nations organs for the promotion of human rights. The Charter of the United Nations vests responsibility for the guarantee of human rights in the General Assembly and, under the General Assembly's authority, in the Economic and Social Council. The Trusteeship Council is concerned with human rights in trust territories, and the Security Council can take jurisdiction over human rights questions when it holds that international peace and security are endangered. The Commission on Human Rights and its sister organ, the Commission on the Status of Women, are subsidiary bodies of the Economic and Social Council. Both commissions consist of government representatives. Proposals that the members of these commissions should be individual experts, independent of governments, were not accepted. Both commissions have, in the course of years, prepared a number of international instruments, some of which will be dealt with below.

The International Bill of Rights. When the Commission on Human Rights was established, it was decided that the preparation of an International Bill of Rights should be its first preoccupation. The plan that was adopted called for the bill to consist of the following documents: a declaration, a covenant, and "measures of implementation." Though the first part, known as the Universal Declaration of Human Rights, was proclaimed by the General Assembly in 1948, it was not until 1966 that the rest of the International Bill of Rights, consisting of three international treaties in addition to the Universal Declaration of 1948, was ready for signature and ratification. The three additional treaties are the International Covenant on Economic, Social and Cultural Rights; the International Covenant on Civil and Political Rights; and the Optional Protocol to the latter creating the important right of communication or petition.

The Universal Declaration of Human Rights. The catalog of rights set out by the Universal Declaration of Human Rights is scarcely less than the sum of all the important traditional political and civil rights of national constitutions and legal systems. Among these are equality before the law; protection against arbitrary arrest; the right to a fair trial and freedom from ex post facto criminal laws; the right to own property; freedom of thought, conscience, and religion; freedom of opinion and expression; and freedom of peaceful assembly and association.

To these lists of civil and political rights the declaration has added economic, social, and cultural rights, such as the right to work and to choose one's work freely, the right to earn equal pay for equal work, and the right to education. The declaration is not an international treaty and was meant to proclaim a common standard of achievement rather than enforceable legal obligations. But to the degree that it has filled a gap caused by the delay in the completion and entry of the covenants into force, the Universal Declaration has acquired a different and more important status than was originally intended. It has been widely used by international organizations, conferences, and governments as a means of judging how well governments have carried out their obligations under the United Nations Charter with respect to questions of human rights.

The uses of the declaration

The International Covenant on Civil and Political Rights and the Optional Protocol. In the International Covenant on Civil and Political Rights, each state party undertakes to respect and to ensure to all individuals within its territory and subject to its jurisdiction the rights recognized in the covenant without any distinction because of race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth, or other status. The catalog of civil and political rights guaranteed by the covenant, though it contains the traditional rights guaranteed in national constitutions and legislation and proclaimed in the Universal Declaration of Human Rights, does not fully coincide with the latter. Some rights listed in the Universal Declaration, such as the right to own property and the right to asylum, are not included among the rights recognized in the covenant. On the other hand, the covenant defines a number of rights not listed in the declaration, among them the right of all peoples to self-determination and the right of ethnic, religious, or linguistic minorities to enjoy their own culture, to profess and practice their own religion, and to use their own language.

Work of the Human Rights Committee

The covenant provides for the establishment of an international organ called the Human Rights Committee. The functions and prerogatives of the committee, which consists of 18 persons elected by the parties and serving as individuals, are of a limited character. The committee is called on to study reports submitted by states on measures they have adopted to give effect to the rights recognized in the covenant. Among states that have expressly accepted this function of the committee, the committee may also respond to allegations by one party that another party is not giving effect to the provisions of the covenant. If the committee is not able to resolve such dispute, the matter is referred to an ad hoc Conciliation Commission, which eventually reports its findings on all questions of fact and its views on the possibilities of an amicable solution. States that become parties to the Optional Protocol thereby recognize, in addition, the competence of the Human Rights Committee to receive and consider communications from individuals who claim to be victims of a violation of the covenant. At the conclusion of these proceedings, the Human Rights Com-

mittee forwards its views to the state concerned and to the individual.

The International Covenant on Economic, Social and Cultural Rights. While the states that are parties to the Covenant on Civil and Political Rights undertake to accept immediately applicable and—within the described limits—enforceable international obligations, the parties to the Covenant on Economic, Social and Cultural Rights undertake only to take steps toward achieving progressively the full realization of the rights recognized in that covenant. These rights are based on those proclaimed in the Universal Declaration and include those economic, social, and cultural rights that were listed above. One obligation undertaken by states becoming parties to the Covenant on Economic, Social and Cultural Rights is, however, of immediate application: the prohibition of discrimination in the enjoyment of those rights because of race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth, or other status. The international measures of supervision that apply to the covenant oblige the parties to report to the Economic and Social Council of the United Nations on the measures they have adopted and the progress they have made in achieving the recognition of the enumerated rights.

Both covenants provide that they shall become effective three months after ratification or accession by the 35th state. By the early 1970s the number of 35 ratifications or accessions had not been reached and the covenants were, therefore, not yet in force.

Other worldwide international conventions under the auspices of the United Nations. *The Convention on the Prevention and Punishment of the Crime of Genocide.* The concept of genocide is closely connected with the principles applied after World War II by the International Military Tribunal that sat at Nuremberg and found some of the major German war figures guilty of crimes against peace, war crimes, and crimes against humanity. In the Genocide Convention of 1948 the contracting parties confirm that genocide, whether committed in time of peace or in time of war, is a crime under international law, which they undertake to prevent and to punish. Genocide is defined as any of several kinds of acts committed with intent to destroy, in whole or in part, a national, ethnic, racial, or religious group, as such. Among the prohibited acts are "killing members of the group," "causing serious bodily or mental harm to members of the group," and "deliberately inflicting on the group conditions of life calculated to bring about its physical destruction in whole or in part." "Imposing measures intending to prevent births within the group" and "forcibly transferring the children of the group to another group" are also punishable as genocide, as is conspiracy, direct and public incitement and attempt to commit genocide, and complicity in genocide. One of the significant results of the convention is that the parties to it place it beyond doubt that genocide, even if perpetrated by a government in its own territory against its own citizens, is not a matter within the domestic jurisdiction of states but a matter of international concern.

Genocide as an international crime

The Supplementary Convention on the Abolition of Slavery, the Slave Trade, and Institutions and Practices Similar to Slavery. Supplementing the International Slavery Convention of 1926, adopted under the auspices of the League of Nations and referred to above, a conference of plenipotentiaries convened by the Economic and Social Council in 1956 prepared and opened for signature a supplementary convention outlawing certain practices similar to slavery such as debt bondage, serfdom, purchase of brides, and exploitation of child labour. It emphasizes the criminality of the slave trade and provides for penal sanctions for some other practices such as mutilating, branding, or otherwise marking a slave or a person of servile status.

The Convention Concerning the Abolition of Forced Labour. In this convention, adopted by the International Labour Conference in 1957, the parties undertake to suppress and not make use of any form of forced or compulsory labour as a measure of political coercion or "education"; as a punishment for holding or expressing political views or views ideologically opposed to the established political, social, or economic system; as a method of mobilizing or using labour for purposes of economic development; or as a means of racial, social, national, or religious discrimination.

The Discrimination (Employment and Occupation) Convention. Three major international instruments are devoted to the fight against discrimination. The first, in time, is the Convention Concerning Discrimination in Respect of Employment and Occupation adopted in 1958 by the International Labour Conference. Under its provisions, states parties undertake to declare and pursue a national policy designed to promote, by methods appropriate to national conditions and practice, equality of opportunity and treatment in employment and occupation and the elimination of any discrimination in that respect. This convention prohibits discrimination, exclusion, or preference on the basis of race, colour, sex, religion, political opinion, national extraction, or social origin that has the effect of nullifying or impairing equality of opportunity or treatment in employment or occupation.

The Convention Against Discrimination in Education. The Convention Against Discrimination in Education adopted by the General Conference of the United Nations Educational, Scientific and Cultural Organization (UNESCO) in 1960 embodies functions similar to those of the 1958 convention in regard to access to education. Parties to the convention of 1960 agree to abrogate any statutory provisions and any administrative instructions and to discontinue any administrative practices that involve discrimination in education or in the admission of pupils to educational institutions. The establishment or maintenance of separate education systems or institutions for pupils of the two sexes does not constitute discrimination if these systems or institutions offer equivalent access to education, provide teaching staffs with qualifications of the same standard, and afford the opportunity to take the same or equivalent courses of study.

Supervision and control of employment regulations

The convention on discrimination in employment is subject to the measures of supervision provided by the constitution of the International Labour Organisation (reporting by states, representations by workers' or employers' organizations, complaints, commissions of inquiry, and reference to the International Court of Justice). The states that are parties to the Convention Against Discrimination in Education undertake to submit periodic reports on the legislative and administrative provisions they have adopted and any other action they have taken for the application of the convention. In addition, in 1962, the General Conference of UNESCO adopted a protocol instituting a Conciliation and Good Offices Commission to be responsible for seeking a settlement of any disputes that may arise between parties to the convention of 1960. The arrangements under the protocol are similar to those provided for in the International Covenant on Civil and Political Rights.

The International Convention on the Elimination of All Forms of Racial Discrimination. In 1965 the General Assembly adopted the International Convention on the Elimination of All Forms of Racial Discrimination. The scope of this convention, which entered into force in 1969, is limited to discrimination based on race, colour, descent, or national or ethnic origin and thereby differs from the two conventions of 1958 and 1960 that also prohibit discrimination on other grounds, such as sex, religion, political or other opinion, etc. On the other hand, the Racial Discrimination Convention is more comprehensive than the two earlier conventions insofar as it is not restricted to any particular fields such as employment and education but prohibits discrimination in regard to human rights and fundamental freedom in all the political, economic, social, and cultural areas of public life. The convention's provisions are far reaching: under them, states undertake not to engage in any act or practice of racial discrimination; to take effective mea-

sures to review governmental, national and local policies; and to amend, rescind, or nullify any laws and regulations that have the effect of creating or perpetuating racial discrimination wherever it exists. The dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, and incitement to acts of violence against any race are made punishable offenses. Organizations and propaganda activities that promote racial discrimination are declared illegal, and participation in them made punishable. These provisions have on occasion been seen to pose constitutional difficulties for some states because of the serious limitations they impose on rights to freedom of expression and freedom of association. A Committee on the Elimination of Racial Discrimination established under the convention consists of 18 experts serving as individuals in their personal capacity. Its functions are similar to those of the Human Rights Committee under the Civil and Political Rights Covenant. In becoming parties to this convention, however, states have automatically accepted the committee's right to consider complaints by any party who alleges that another party is not observing the provisions of the convention. The convention also provides for the appointment of an ad hoc Conciliation Commission with prerogatives similar to those established for the commission under the Civil and Political Rights Covenant. Any dispute between two or more parties with respect to the interpretation or application of the convention that is not settled by negotiation or by the procedures provided for in the convention will be referred to the International Court of Justice for decision, unless the disputants agree to another mode of settlement. Parties to the convention may also recognize the competence of the committee to receive and consider communications from individuals or groups claiming to be victims of a violation of any of the rights set forth in the convention. A special provision contemplates activities of the committee in regard to petitions received by United Nations bodies from non-self-governing territories.

Procedure for complaints on racial discrimination

Conventions on the status of women. In the Convention on the Political Rights of Women (1952), the signatory states have guaranteed that women shall be entitled to vote in all elections on equal terms with men, shall be eligible for election to all publicly elected bodies established by national law on equal terms with men, and shall be entitled to hold public office and to exercise all public functions established by national law without any discrimination.

The Convention of 1962 on Consent to Marriage, Minimum Age for Marriage and Registration of Marriages is of particular importance for developing countries, mainly in Asia and Africa. It provides that no marriage shall be legally entered into without the full and free consent of participants, such consent to be expressed by them in person after due publicity and in the presence of the authority competent to solemnize the marriage and of witnesses. The parties are also obligated to take legislative action to specify a minimum age for marriage and to provide for the official registration of all marriages.

The Convention relating to the Status of Refugees of 1951 and the Protocol of 1967. The Convention on the Status of Stateless Persons, 1954. Two principles form the basis of the Refugees Convention of 1951: first, that there should be as little discrimination as possible between nationals and refugees; second, that there should be no discrimination at all based on race, religion, or country of origin among refugees. The parties have undertaken to apply "national treatment" to refugees—*i.e.*, treatment at least as favourable as that accorded to their own nationals with regard to certain rights such as freedom of religion, access to courts, elementary education, and public relief. With regard to other rights (wages, employment, and the right of association), refugees are entitled to the "most-favoured-nation treatment"—*i.e.*, the most favourable treatment accorded to nationals of a foreign country. In other respects (*e.g.*, self-employment and higher education), refugees receive treatment as favourable as possible and, in any event, not less favourable than that accorded to aliens generally. The Convention relating to the Status of Stateless Persons (1954) provides for protective measures for stateless persons similar to those applicable to refugees under the convention of 1951.

Solemn declarations of the United Nations. In addition to developing the law relating to human rights by the preparation and adoption of international treaties, the most important of which have been described in the preceding section, the General Assembly, impressed by the effect of the Universal Declaration of Human Rights of 1948, has used the technique of proclaiming declarations on other occasions as well. A declaration adopted in the form of a resolution of the General Assembly is not binding upon member states in the sense of a treaty, but the solemnity and significance of such a "Declaration" creates a strong expectation that members of the international community will abide by it, particularly when it is used to enunciate principles of great and lasting importance. The best known of these declarations adopted by the General Assembly subsequent to the Universal Declaration of Human Rights is the Declaration on the Granting of Independence to Colonial Countries and Peoples (1960), which, with the machinery established for its implementation in the following years, had a far-reaching impact on the process of decolonization. Other declarations in the human rights field, which can only be listed here, are the Declaration of the Rights of the Child of 1959, the United Nations Declaration on the Elimination of All Forms of Racial Discrimination of 1963, the Declaration on the Elimination of Discrimination against Women of 1967, and the Declaration on Territorial Asylum of the same year.

The effect of UN declarations

Regional developments in the field of human rights. *The European Convention on Human Rights.* The Charter of the United Nations and the Universal Declaration of Human Rights have inspired action for the international protection of human rights within two important regions of the world. Members of the Council of Europe agreed, in 1950, to a European Convention for the Protection of Human Rights and Fundamental Freedoms. This convention together with five additional protocols represents the most advanced and successful experiment in the field. The substantive provisions of the European convention are based on an earlier draft of what is now the United Nations International Covenant on Civil and Political Rights. The instrumentalities created under the European convention are the European Commission on Human Rights, a body much like the Human Rights Committee under the covenant, and the Committee on the Elimination of Racial Discrimination. Members sit as individuals, and any party to the convention may refer to the commission any alleged breach of the provisions by another party. The commission may (but only when its legal competence to do so has been recognized) receive petitions from any person, nongovernmental organization, or group of individuals claiming to be the victim of a violation of the convention. In such cases, the commission is charged with ascertaining the facts and with placing itself at the disposal of the parties to secure a friendly settlement on the basis of respect for human rights. In the event that no solution is reached, the commission is called upon to draw up a report and to state its opinion as to whether the facts found disclose a breach of the obligations under the convention. In 1971 there were 17 parties to the European Convention, of which 11 had accepted the competence of the commission to examine petitions.

The European convention also established the European Court of Human Rights. Eleven states have agreed to accept its compulsory jurisdiction. Either the commission or a state party whose national is alleged to be a victim of a violation, the state party which referred the case to the commission, or the state party against which the complaint has been alleged can refer a case to the court. The judgment of the court is final.

The Court of Human Rights

If a question is not, or cannot be, referred to the court, then a political organ, the Committee of Ministers of the

Council of Europe, makes the final decision on human rights complaints.

The European Commission on Human Rights has, over the years, developed a considerable body of case law on questions regulated in the convention; and the provisions of the convention are, in some European states, automatically part of the constitutional or statutory law. Where this is not the case, the western European states have taken other measures to bring their internal law in line with their obligations under the convention.

The American Convention on Human Rights. In the Western Hemisphere, the Organization of American States has, for many years, taken comprehensive action in the human rights field on a regional basis. Within the framework of the Organization of American States, a special organ, the Inter-American Commission on Human Rights, was established in 1959. Since that time it has undertaken important activities in regard to human rights situations that have arisen in some of the American states. In November 1969 the Inter-American Specialized Conference on Human Rights meeting at San José, Costa Rica, adopted an instrument entitled American Convention on Human Rights (also called Pact of San José, Costa Rica).

The American Convention makes the existing Inter-American Commission on Human Rights an organ for its implementation and also establishes the Inter-American Court of Human Rights. Both the substantive law and the procedural arrangements of the American convention on Human Rights of 1969 are strongly influenced by the United Nations Covenants and by the European Convention on Human Rights. An important difference between the American Convention on Human Rights and its European and United Nations predecessors consists in the fact that under the American Convention the right of petition of individuals applies automatically while under the European system a special declaration by the states concerned is required; and under the United Nations system, the right of petition is applicable only when the state concerned has become a party to the Optional Protocol (on communications) to the International Covenant on Civil and Political Rights. The system of interstate complaints applies under the American convention only among states that have expressly agreed to accept it.

BIBLIOGRAPHY

General: UNITED NATIONS, *Human Rights: A Compilation of International Instruments of the United Nations* (1967), contains the texts of all human rights treaties and other instruments established under the auspices of the UN. The *United Nations Yearbook on Human Rights* (pub. annually since 1946), documents the national and international developments in the human rights field. The classic works on the subject are H. LAUTERPACHT, *An International Bill of the Rights of Man* (1945), and *International Law and Human Rights* (1950, reissued 1968). Writings on human rights under the Charter of the United Nations, the International Bill of Rights, and other instruments established by the UN include: J. ROBINSON, *Human Rights and Fundamental Freedoms in the Charter of the United Nations* (1946); M. GANJI, *International Protection of Human Rights* (1962); E. SCHWELB, *Human Rights and the International Community* (1964), and "International Conventions on Human Rights," *Int. Comp. Law Q.*, 9:654–675 (1960); M.S. MCDOUGAL and G. BEBR, "Human Rights in the United Nations," *Am. J. Int. Law*, 58:603–641 (1964); A. EIDE and A. SCHOU (eds.), *International Protection of Human Rights* (1968); EVAN LUARD (ed.), *The International Protection of Human Rights* (1967); and M. MOSKOWITZ, *The Politics and Dynamics of Human Rights* (1968).

On specific worldwide instruments: N. ROBINSON, *The Genocide Convention: A Commentary* (1960); P. WEIS, "The International Protection of Refugees," *Am. J. Int. Law*, 48:193–221 (1954); E. SCHWELB, "Marriage and Human Rights," *Am. J. Comp. Law*, 12:337–383 (1963), and "The International Convention on the Elimination of All Forms of Racial Discrimination," *Int. Comp. Law Q.*, 15:996–1068 (1966).

On the International Labour Organisation and human rights: J.T. SHOTWELL, *The Origins of the International Labour Organisation* (1934); C.W. JENKS, *The International Protection of Trade Union Freedom* (1957), *Human Rights and International Labour Standards* (1960).

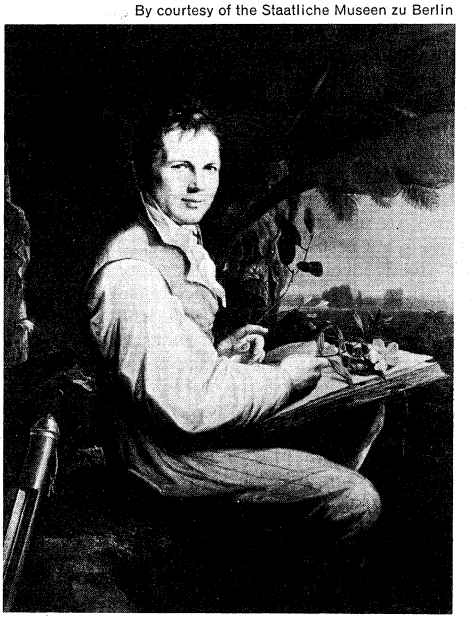
On regional instruments and arrangements: J.E.S. FAWCETT, *The Application of the European Convention on Human Rights* (1969); H. ROBERTSON, *Human Rights in Europe* (1963); G.L. WEIL, *The European Convention on Human Rights* (1962); BRITISH INSTITUTE OF INTERNATIONAL AND COMPARATIVE LAW, *The European Convention on Human Rights* (1965); *Yearbook of the European Convention on Human Rights* (pub. since 1955); J.A. CABRANES, "The Protection of Human Rights by the Organization of American States," *Am. J. Int. Law*, 62:889–908 (1968).

(E.S.)

Humboldt, Alexander von

Explorer and scientist, Alexander von Humboldt gained worldwide renown as the chief propagator in his time of the study of the earth sciences—geography, geology, geomagnetism, climatology—and of the relation of plants and animals to their physical surroundings, now known as the science of ecology. He perceived and identified ecological relationships; he established the connection between geological faults and the related phenomena of volcanoes and earthquakes; he promoted the use of graphic representation of scientific data for ease of comprehension; and through a lifetime of tireless effort he furthered the cause of science by his publications, his support of international collaboration in scientific projects, and the advice he gave young scientists at the formative stages of their careers. With his book *Kosmos*, a survey of astronomy and the earth sciences, he made a valuable contribution to the popularization of science.

By courtesy of the Staatliche Museen zu Berlin



Humboldt, oil painting by Friedrich Georg Weitsch, 1806. In the Staatliche Museen zu Berlin.

Early life. Alexander von Humboldt was born on September 14, 1769, in Berlin, then the capital of the kingdom of Prussia. His father was an officer in the army of Frederick the Great; his mother belonged to a family of Huguenots (French Protestants) who had left France after Louis XIV's revocation, in 1685, of religious liberty for Protestants. After the elder Humboldt's death in 1779, the education of his two sons, Wilhelm and Alexander, was left in the hands of their mother, an unemotional woman of strict Calvinist beliefs. They were privately educated; instruction in political history and economics was added to the usual courses in classics, languages, and mathematics, as their mother intended them to be qualified for high public positions. Alexander, a sickly child, at first was a poor student. He was restless, thought of joining the army, and followed his courses only under parental pressure. After futile studies in economics at the University of Frankfurt an der Oder, he spent a year in Berlin where he obtained some training in engineering and suddenly became passionately interested

in botany. He began to collect plant specimens in the surroundings of Berlin, and learned to classify them. But the poor flora of the province of Brandenburg did not provide much stimulus for an ardent botanist, and Humboldt soon dreamed of journeys to more exotic lands.

Work in mineralogy and geology

A year spent at the University of Göttingen, from 1789 to 1790, finally opened the world of science to him; he became particularly interested in mineralogy and geology and decided to obtain a thorough training in these subjects by joining the School of Mines in Freiburg, Saxony, the first such establishment. Although founded only in 1766, the school had already acquired an international reputation. There, buttressed by a prodigious memory and driven by an unending thirst for knowledge, he began to develop his enormous capacity for work. After a morning spent underground in the mines, he attended classes for five or six hours in the afternoon, and in the evening scoured the country for plants.

He left Freiburg in 1792 after two years of intensive study but without taking a degree. A month later he obtained an appointment in the Mining Department of the Prussian government and departed for the remote Fichtel Mountains in the Margraviate of Ansbach-Bayreuth, which had only recently come into the possession of the Prussian kings. Here Humboldt came into his own; he travelled untiringly from one mine to the next, reorganizing the partly deserted and totally neglected pits, which produced mainly gold and copper. He supervised all mining activities, invented a safety lamp, and established, with his own funds, a technical school for young miners. Yet he did not intend to make mining his career.

Expedition to South America. The conviction had grown in Humboldt that his real aim in life was scientific exploration, and in 1797 he resigned from his post to acquire with great single-mindedness a thorough knowledge of the systems of geodetic, meteorological, and geomagnetic measurements. The political upheavals caused by the Napoleonic Wars prevented the realization of several scientific expeditions in which Humboldt had been given an opportunity to participate. At last, dispirited by his disappointments but refusing to be deterred from his purpose, he obtained permission from the Spanish government to visit the Spanish colonies in Central and South America. These colonies were then accessible only to Spanish officials and the Roman Catholic mission. Completely shut off from the outside world, they offered enormous possibilities to a scientific explorer. Humboldt's social standing assured him of access to official circles, and in the Spanish prime minister Mariano de Urquijo he found an enlightened man who supported his application to the king for a royal permit. In the summer of 1799 he set sail from Marseilles accompanied by the French botanist Aimé Bonpland, whom he had met in Paris, then the liveliest scientific centre in Europe. The estate he had inherited at the death of his mother enabled Humboldt to finance the expedition entirely out of his own pocket. Humboldt and Bonpland spent five years, from 1799 to 1804, in Central and South America, covering more than 6,000 miles (9,650 kilometres) on foot, on horseback, and in canoes. It was a life of great physical exertion and serious deprivation.

Starting from Caracas, they travelled south through grasslands and scrublands until they reached the banks of the Apure, a tributary of the Orinoco River. They continued their journey on the river by canoe as far as the Orinoco. Following its course and that of the Casiquiare, they proved that the Casiquiare River formed a connection between the vast river systems of the Amazon and the Orinoco. For three months Humboldt and Bonpland moved through dense tropical forests, tormented by clouds of mosquitoes and stifled by the humid heat. Their provisions were soon destroyed by insects and rain; the lack of food finally drove them to subsist on ground-up wild cacao beans and river water. Yet both travellers, buoyed up by the excitement provided by the new and overwhelming impressions, remained healthy and in the best of spirits until their return to civilization, when they succumbed to a severe bout of fever.

After a short stay in Cuba, Humboldt and Bonpland returned to South America for an extensive exploration of the Andes. From Bogotá, Colombia, to Trujillo, Peru, they wandered over the Andean Highlands—following a route now traversed by the Pan-American Highway, in their time a series of steep, rocky, and often very narrow paths. They climbed a number of peaks, including all the volcanoes in the surroundings of Quito, Ecuador; Humboldt's ascent of Chimborazo (20,561 feet; 6,265 metres) to a height of 19,280 feet, but short of the summit, remained a world mountain-climbing record for nearly 30 years. All these achievements were carried out without the help of modern mountaineering equipment, without ropes, crampons, or oxygen supplies; hence, Humboldt and Bonpland suffered badly from mountain sickness. But Humboldt turned his discomfort to advantage: he became the first person to ascribe mountain sickness to lack of oxygen in the rarefied air of great heights. He also studied the oceanic current off the west coast of South America that was originally named after him but is now known as the Peru Current. When the pair arrived, worn and footsore, in Quito, Humboldt, the experienced mountaineer and indefatigable collector of scientific data, had no difficulty in assuming the role of courtier and man of the world when he was received by the viceroy and the leaders of Spanish society.

Ascent of Chimborazo

In the spring of 1803, the two travellers sailed from Guayaquil to Acapulco, Mexico, where they spent the last year of their expedition in a close study of this most developed and highly civilized part of the Spanish colonies. After a short stay in the United States, where Humboldt was received by President Jefferson, they sailed for France.

Humboldt and Bonpland returned with an immense amount of information. In addition to a vast collection of new plants, there were determinations of longitudes and latitudes, measurements of the components of the Earth's geomagnetic field, and daily observations of temperatures and barometric pressure, as well as statistical data on the social and economic conditions of Mexico. Whenever Humboldt had found himself in a centre of commerce in America, he had sent off reports and duplicates of his collections to his brother, Wilhelm, who had become a noted philologist, and to French scientists; unfortunately, the continental blockade then enforced by British ships prevented the greater part of his mail from reaching its destination.

Professional life in Paris. The years from 1804 to 1827 Humboldt devoted to publication of the data accumulated on the South American expedition. With the exception of brief visits to Berlin, he lived in Paris during this important period of his life. There he found not only collaborators among the French scientists—the greatest of his time—but engravers for his maps and illustrations, and publishers for printing the 30 volumes into which the scientific results of the expedition were distilled. Of great importance were the meteorological data, with an emphasis on mean daily and nightly temperatures, and Humboldt's representation on weather maps of isotherms (lines connecting points with the same mean temperature) and isobars (lines connecting points with the same barometric pressure for a given time or period) —all of which helped lay the foundation for the science of comparative climatology. Even more important were his pioneering studies on the relationship between a region's geography and its flora and fauna, and, above all, the conclusions he drew from his study of the Andean volcanoes concerning the role played by eruptive forces and metamorphosis in the history and ongoing development of the Earth's crust. These conclusions disproved once and for all the hypothesis of the so-called Neptunists, who held that the surface of the Earth had been totally formed by sedimentation from a liquid state. Lastly, his *Political Essay on the Kingdom of New Spain* contained a wealth of material on the geography and geology of Mexico, including descriptions of its political, social, and economic conditions, and also extensive population statistics. Humboldt's impassioned outcry in this work against the inhumanities of slavery remained unheard, but his descriptions of the Mexican silver mines

led to widespread investment of English capital and mining expertise in the mines.

Influence on scientific community

During his years in Paris, Humboldt enjoyed an extraordinarily full life. He had the ability to cultivate deep and long-lasting friendships with well-known scientists, such as the renowned physicist and astronomer François Arago, and to evoke respect and admiration from the common man, an ability that reflected his generosity, humanity, and vision of what science could do. A gregarious person, Humboldt appeared regularly in the salons of Parisian society, where he usually dominated the conversation. He lived simply, in a modest apartment at the top of an old house in the Latin Quarter. His fortune had been seriously depleted by the cost of his expedition and the publication of his books, and for the rest of his life he was often in financial straits. He was, moreover, always willing and anxious to assist young scientists at the beginning of their careers. Due to his magnanimity, generosity, and wise judgment, poor but promising students were given the necessary encouragement, financial assistance, and introductions to the scientific community to insure a successful start in life. Such men as the German chemist Justus von Liebig and the Swiss-born zoologist Louis Agassiz owed to Humboldt the means to continue their studies and embark on an academic career. The best proof of his wide interests and affectionate nature lies in his voluminous correspondence: 8,000 letters remain.

Return to Berlin. The happy years in Paris came to an end in 1827. Humboldt's means by then were almost completely exhausted; unable to maintain his financial independence, he had to return to Berlin, where the King impatiently demanded his presence at court. Until a few years before his death, Humboldt served as a tutor to the Crown Prince, as a member of the privy council, and as a court chamberlain. He made use of his position to acquaint the young Prince and the royal family with scientific methods and the scientific ideas of his time. His enthusiasm for the popularization of science prompted him to give a course on physical geography to the professors and students of all faculties of the University of Berlin, part of which he repeated in a public lecture to an audience of more than 1,000. In the autumn of the same year, 1828, he also organized, in Berlin, one of the first international scientific conferences. Such large gatherings of possibly liberal-minded people were frowned upon by governments in the wake of the Napoleonic Wars and the attendant rise of democratic expectations, and it is a tribute to Humboldt's adroitness that he was able to overcome the misgivings of official Prussian circles.

Expedition to Russia

In 1829 Humboldt was given the opportunity to visit Russia and Siberia. On the initiative of the Russian minister of finance, Count Yegor Kankrin, he was invited to visit the gold and platinum mines in the Urals, as an adviser to the government on the techniques and organization of mining. But Humboldt had to pledge himself to refrain from commenting on the political situation of the country whose despotism he abhorred. This expedition, lasting only one summer, was very different from the South American journey; the members, Humboldt and two young scientists, were accompanied throughout by an official guard, since they were guests of the tsar. Humboldt and his companions had to endure tiresome receptions at the imperial court and in the homes of provincial governors. They travelled in carriages as far as the Altai Mountains and the Chinese frontier. The resulting geographical, geological, and meteorological observations, especially those regarding the Central Asian regions, were of great importance to the Western world, for Central Asia was then to a large degree unknown territory.

Humboldt passed the last 30 years of his life in Berlin. Once a year he travelled to Paris, where he renewed his contacts with the French scientists, enjoyed daily discussions with his friend, Arago, and breathed the cosmopolitan air he so sadly missed in Berlin.

Even before his visit to Russia, he had returned to an investigation of a phenomenon that had aroused his interest in South America: the sudden fluctuations of the Earth's geomagnetic field—the so-called magnetic storms. With the help of assistants, he carried out observations of the movement of a magnetometer in a quiet garden pavilion in Berlin; but it had been clear to him for a number of years that, to discover whether these magnetic storms were of terrestrial or extraterrestrial origin, it would be necessary to set up a worldwide net of magnetic observatories. The German mathematician Carl Friedrich Gauss had already begun to organize simultaneous measurements of the magnetic field by several observatories in Germany, England, and Sweden. In 1836 Humboldt, still interested in the problem, approached the Royal Society in London with the request that it establish an additional series of stations in the British possessions overseas. As a result, the British government provided the means for permanent observatories in Canada, South Africa, Australia, and New Zealand and equipped an Antarctic expedition. With the help of the mass of data produced by this international scientific collaboration, one of the first of its kind, the English geophysicist Sir Edward Sabine later succeeded in correlating the appearance of magnetic storms in the earth's atmosphere with the periodically changing activity of sunspots, thus proving the extraterrestrial origin of the storms.

During the last 25 years of his life, Humboldt was chiefly occupied with writing *Kosmos*, one of the most ambitious scientific works ever published. Four volumes appeared during his lifetime. Written in a pleasant, literary style, *Kosmos* gives a generally comprehensible account of the structure of the universe as then known, at the same time communicating the scientist's excitement and aesthetic enjoyment at his discoveries. Humboldt had taken immense pains to discipline his inclination to discursiveness, which often gave his writing a certain lack of logical coherence. He was rewarded for his effort by the success of his book, which, within a few years, had been translated into nearly all European languages.

On May 6, 1859, still working on the fifth volume of *Kosmos* with hardly diminished vitality and enthusiasm and with an unimpaired memory, Humboldt died in his 90th year, in Berlin.

BIBLIOGRAPHY

Biographies: HANNO BECK, *Alexander von Humboldt*, 2 vol. (1959–61), the definitive biography (in German), giving a detailed account of Humboldt's life and work with complete bibliography; L. KELLNER, *Alexander von Humboldt* (1963), the first English biography since 1872, puts equal emphasis on life and work and on historical background; C.F. GAUSS, *Briefe zwischen Alexander von Humboldt und Gauss* (1877).

Correspondence: Briefe von Alexander von Humboldt an Varnhagan von Ense, aus den Jahren 1827 bis 1858, 2nd ed. (1860; Eng. trans., *Letters of Alexander von Humboldt to Varnhagan von Ense, from 1827 to 1858*, 1860); *Briefwechsel mit Berghaus*, 3 vol. (1863); *Briefe Alexander's von Humboldt an seinen Bruder Wilhelm* (1880); *Lettres Américaines d'Alexandre de Humboldt*, ed. by E.T. HAMY (1905); *Correspondance d'Alexandre de Humboldt avec Francois Arago*, ed. by E.T. HAMY (1907); *Briefwechsel zwischen Alexander von Humboldt und Graf Georg von Cancrin* (1869), all give a very good idea of Humboldt's personal relations and the development of his scientific views.

Works: Ansichten der Natur, mit wissenschaftlichen Erläuterungen, (1808; Eng. trans., *Aspects of Nature, in Different Lands and Different Climates*, 1849), descriptions of scenic aspects and a sketch of the geography of plants for the general reader; *Voyage de Humboldt et Bonpland aux Regions Équinoxiales du nouveau Continent, fait en 1799–1804*, 23 vol. (1805–34; Eng. trans., *Personal Narrative of Travels to the Equinoctial Regions of the New Continent During the Years 1799–1804*, 7 vol., 1814–29), a very well-known account of the journey through the grasslands of Venezuela and the Orinoco Basin that had great influence on Darwin; *Asie Centrale*, 3 vol. (1843), an important book for its report on relatively unknown regions and for its ideas on comparative climatology.

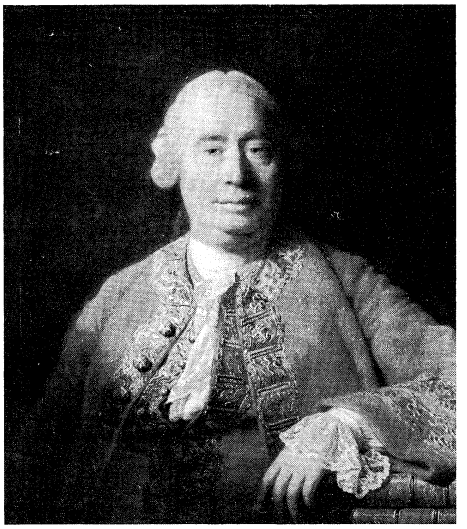
(C.L.K.)

Hume, David

David Hume was an 18th-century British Empiricist philosopher, historian, economist, and essayist, who conceived of philosophy as the inductive, experimental sci-

ence of human nature. Taking the scientific method of the British physicist Sir Isaac Newton as his model, Hume tried to describe how the mind works in acquiring what is called knowledge. He concluded that no theory of reality is possible; there can be no knowledge of anything beyond experience. Despite the enduring impact of his theory of knowledge, Hume seems to have considered himself chiefly as a moralist.

By courtesy of the Scottish National Portrait Gallery



Hume, oil painting by Allan Ramsay, 1766. In the Scottish National Portrait Gallery, Edinburgh.

Early life and works. Hume was the younger son of Joseph Hume, the modestly circumstanced laird, or lord, of Ninewells, a small estate adjoining the village of Chirnside, about nine miles distant from Berwick-upon-Tweed on the Scottish side of the border. The family seems to have had some remote connection with the earls of Home, an influential border family whose ancestral stronghold was in Berwickshire. David's mother, Catherine, a daughter of Sir David Falconer, president of the Scottish court of session, was in Edinburgh when he was born, on May 7 (April 26, old style), 1711. In his third year his father died. He entered Edinburgh University when he was about 12 years old and left it at 14 or 15, as was then usual. Pressed a little later to study law (in the family tradition on both sides), he found it distasteful and instead read voraciously in the wider sphere of letters. Through the intensity and excitement of his intellectual discovery, he had a nervous breakdown in 1729, from which it took him a few years to recover.

Composing the *Treatise*

In 1734, after trying his hand in a merchant's office in Bristol, he came to the turning point of his life and retired to France for three years. Most of this time he spent at La Flèche on the Loire, in the old Anjou, studying and composing *A Treatise of Human Nature*. The *Treatise* was Hume's attempt to formulate a full-fledged philosophical system. It is divided into three books: book I, on understanding, aims at explaining man's process of knowing, describing in order the origin of ideas, the ideas of space and time, causality, and skepticism with regard to the senses; book II, on the "passions" of man, gives an elaborate psychological machinery to explain the affective, or emotional, order in man and assigns a subordinate role to reason in this mechanism; book III, on morals, describes moral goodness in terms of "feelings" of approval or disapproval that man feels when he considers human behaviour in the light of the agreeable or disagreeable consequences either to himself or to others. Although the *Treatise* is Hume's most thorough exposition of his thought, at the end of his life he vehemently repudiated it as juvenile, avowing that only his later writings presented his considered views. The *Treatise* is badly constructed, in parts oversubtle, confusing because of ambiguity in important terms (especially "reason"), and marred by willful extravagance of statement and rather theatrical personal avowals. His mature condemnation of it was only severe, not misplaced. Nevertheless, book I has been more read in academic circles than any other of his writings.

Returning to England in 1737, he set about publishing the *Treatise*. Books I and II were published in two volumes in 1739; book III appeared the following year. The negligent reception of this, his first and very ambitious work, depressed him; but his next venture, *Essays, Moral and Political* (1741–42), won some success. Perhaps encouraged by this, he became a candidate for the chair of moral philosophy at Edinburgh in 1744. Objectors alleged heresy and even atheism, pointing to the *Treatise* as evidence. Unsuccessful, Hume left the city, where he had been living since 1740, and began a period of wandering: a sorry year near St. Albans as tutor to the mad Marquess of Annandale (1745–46); a few months as secretary to Gen. James St. Clair, a member of a prominent Scottish family, throughout an abortive expedition to Brittany (1746); a little tarrying in London and at Ninewells; and then some further months with General St. Clair on an embassy to the courts of Vienna and Turin (1748–49).

Mature works. During his years of wandering, Hume was earning the money that he needed to get leisure for his studies. Some fruits of these studies had already appeared before the end of his travels, viz., a further *Three Essays, Moral and Political* (1748) and *Philosophical Essays Concerning Human Understanding* (1748). The latter is a rewriting of book I of the *Treatise* (with the addition of his essay "On Miracles," notorious for its denial that a miracle can be proved by any amount or kind of evidence); it is better known as *An Enquiry Concerning Human Understanding*, the title Hume gave to it in a revision of 1758. The *Enquiry Concerning the Principles of Morals* (1751) was a rewriting of book III of the *Treatise*. It was in these works that Hume expressed his mature thought.

Theory of knowledge

The *Inquiry Concerning Human Understanding* is an attempt to define the principles of human knowledge. It presents in logical form the significant questions about the nature of all reasoning in regard to matters of fact and experience, and it solves the problems by recourse to association. The basis of his exposition is a twofold classification of objects of awareness. In the first place, all such objects are either "impressions," ultimate data of sensation or of internal consciousness, or "ideas," derived from the data by compounding, transposing, augmenting, or diminishing. That is to say, men do not create any ideas. From this Hume infers a theory of meaning: a word that does not stand directly for an impression has meaning only if it brings before the mind an object that can be gathered from an impression by one of the mental processes mentioned. Secondly, all objects of awareness are either "relations of ideas" or "matters of fact." Ideas can be held before the mind simply as meanings, and their logical relations to one another can then be detected by rational inspection. The idea of a plane triangle, for example, entails the equality of its internal angles to two right angles, and the idea of motion entails the ideas of space and time, irrespective of whether there really are such things as triangles and motions. Only on this level of mere meanings, Hume asserts, is there room for demonstrative knowledge. Matters of fact, on the other hand, objects believed to exist, come before us merely as they are, revealing no logical relations; their properties and connections have to be accepted as they are given. That primroses are yellow, that lead is heavy, and that fire burns things are facts, each shut up in itself, logically barren. Each, so far as reason is concerned, could be different: the contradictory of every matter of fact is conceivable. Therefore, any demonstrative science of fact is impossible.

Hume's doctrine of causality and belief

From this basis Hume develops his distinctive doctrine about causality. The idea of causality is alleged to assert a necessary connection among matters of fact. From what impression, then, is it derived? Hume observes no causal relation among the data of the "external" senses, for, when a man regards any events as causally con-

nected, all that he does and can observe is that they frequently and uniformly go together. In this sort of togetherness, it is a fact that the impression or idea of the one event brings with it the idea of the other. A habitual association is set up in the mind; and, as in other forms of habit, so in this one, the working of the association is felt as compulsion. This feeling, Hume concludes, is the only discoverable impressional source of the idea of causality. But there is also the question of belief. When one sees a glass fall, he not only thinks of its breaking but expects and believes it; or, starting from an effect, when he sees the ground to be generally wet, besides thinking of rain, he believes that there has been rain. Hume claims that he was the first to investigate the nature of belief. He uses this term in the narrowed sense of belief in matters of fact, in anything as existing. He describes belief as a sort of liveliness or vividness possessed by some of the immediate objects of awareness, originally by impressions and the simple memory images of them. How does it come to belong to certain ideas? By association. This is the essence of causal inference: the observer passes from an impression to an idea regularly associated with it. Hume confidently asserts that in the process the aspect of liveliness proper to the impression infects the idea.

Hume has not claimed to prove that the propositions, (1) that events themselves are causally related and (2) that they will be related in the future in the same ways as they were in the past, are false. He firmly believed both of these propositions and insisted that everybody else does, will continue to do, and must do so in order to survive. They are natural beliefs, inextinguishable propensities of man's nature, madness apart. What he has claimed to prove is that they are not obtained and cannot be illuminated either by empirical observation or by reason, whether intuitive or inferential. Reflection shows that there is no evidence for them and shows also both that we are bound to believe them and that it is sensible or sane to do so. *This* is Hume's Skepticism: it is an affirmation of that tension, a denial not of belief but of evidence.

Theory of morals and historical writing

The *Enquiry Concerning the Principles of Morals* is a refinement of Hume's thinking on morality, in which he views sympathy as the fact of human nature lying at the basis of all social life and personal happiness. Defining morality as those qualities that are approved (1) in whomsoever they happen to be and (2) by virtually everybody, he sets himself to discover the broadest grounds of the approvals. He finds them, as he found the grounds of belief, in "feelings," not in "knowings." Moral decisions are grounded upon moral sentiment. Qualities are valued either for their utility or for their agreeableness, in each case either to their owners or to others. Hume's moral system aims at the happiness of others (without any such formula as "the greatest happiness of the greatest number") and at the happiness of self. But regard for others accounts for the greater part of morality. His emphasis is on altruism: of the moral sentiments that he claims to find in man, he traces them, for the most part, to a sentiment for and a sympathy with one's fellows. It is man's nature, he holds, to laugh with the laughing and to grieve with the grieved and to seek the good of others as well as his own. Two years after the *Enquiry* was published, Hume confessed, "I have a partiality for that work"; and at the end of his life he judged it "of all my writings incomparably the best." Such statements, along with other indications in his later writings, make it possible to suspect that he regarded his moral doctrine as his major work. He here writes as a man having the same engagements as his fellows. The traditional view that he was a detached scoffer is deeply wrong: he was skeptical not of morality but of much theorizing about it.

A settled spell (1751–63), spent in Edinburgh, with two breaks in London, followed the publication of these works. It opened with an attempt to get him appointed as successor to Adam Smith, the Scottish economist (later to be his close friend), in the chair of logic at Glasgow. The rumour of atheism prevailed again. In 1752, however, he was made keeper of the Advocates' Library at Edinburgh. There, "master of 30,000 volumes," he could indulge a desire of some years to turn to historical writing. His *History of England*, extending from Caesar's invasion to 1688, came out in six quarto volumes between 1754 and 1762, preceded by *Political Discourses* (1752). His recent writings had begun to make him known, but these two brought him fame, abroad as well as at home. He also wrote *Four Dissertations* (1757), which he regarded as a trifle, although it included a rewriting of book II of the *Treatise* (completing his purged restatement of this work) and an able study of "the natural history of religion." In 1762 James Boswell, the biographer of Samuel Johnson, called Hume "the greatest writer in Britain," and Rome paid him the attention of putting all his writings on the *Index*, the list of forbidden books in the Roman Catholic Church, in 1761.

Hume's reception in Paris

The most colourful episode of his life ensued: in 1763 he left England to become secretary to the British embassy in Paris under the Earl of Hertford. The society of Paris accepted him, despite his ungainly figure and gauche manner. He was honoured as eminent in breadth of learning, in acuteness of thought, and in elegance of pen and was taken to heart for his simple goodness and cheerfulness. The salons threw open their doors to him, and he was even given a distinguished reception at court. For four months in 1765, he acted as chargé d'affaires at the embassy. When he returned to London at the beginning of 1766 (to become, a year later, undersecretary of state), he brought Jean-Jacques Rousseau, the French philosopher and author, with him and found him a refuge from persecution in a country house at Wootton in Staffordshire. This pathetic genius, under one of his delusions, suspected a plot, took secret flight back to France, and spread a report of Hume's bad faith. Hume was partly stung and partly persuaded into publishing the relevant correspondence between them with a connecting narrative (*A Concise and Genuine Account of the Dispute Between Mr. Hume and Mr. Rousseau*, 1766).

In 1769, somewhat tired of public life and of England too, he again established a residence (he never married) in his beloved Edinburgh, deeply enjoying the company —at once intellectual and convivial—of friends old and new, as well as revising the text of his writings. He issued five further editions of his *History* between 1762 and 1773 as well as eight editions of his collected writings (omitting the *Treatise*, *History*, and ephemera) under the title *Essays and Treatises* between 1753 and 1772, besides preparing the final edition of this collection, which appeared posthumously (1777), and *Dialogues Concerning Natural Religion*, held back under pressure from friends and not published until 1779. His curiously detached autobiography, *The Life of David Hume, Written by Himself* (1777; the title is his own), is dated April 18, 1776. After a long illness he died in his Edinburgh house on August 25, 1776, and was buried on Calton Hill.

Adam Smith, his literary executor, added to the *Life* a letter that concludes with his judgment on his friend as "approaching as nearly to the idea of a perfectly wise and virtuous man as perhaps the nature of human frailty will permit." His distinguished friends, with ministers of religion among them, certainly admired and loved him, and there were younger men indebted either to his influence or to his pocket. The mob had heard only that he was an atheist and simply wondered how such an ogre would manage his dying. Yet Boswell has recounted, in a passage in his *Private Papers*, that, when he visited Hume in his last illness, the philosopher put up a lively, cheerful defense of his disbelief in immortality.

Significance and influence. That Hume was one of the major figures of his century can hardly be doubted. So his contemporaries thought, and his achievement, as seen in historical perspective, confirms that judgment, though with a shift of emphasis. Some hints of the reasons for the assessment may be given under four heads:

As a writer. Hume's style was praised in his lifetime and has often been praised since; but not so much in the 20th century. It seems in retrospect to be little, if any-

thing, more than the style of his day at a high level of competence. It lacks individuality and the pulse of variety and power. Hume's quill never becomes a wing, nor does it ever plunge into depths; and the pathetic was beyond him, for he was always proudly on guard against his emotions. The touch is light, except on slight subjects, where it is rather heavy. Yet in his philosophical works—those later than the *Treatise*—he gives an unsought pleasure. Here his detachment, levelness (all on one plane), smoothness, and daylight clearness are proper merits. It is as one of the best writers of scientific prose in English that he stands in the history of style.

As a historian. Library catalogs still list Hume as "Hume, David, the Historian." Between his death and 1894, there were at least 50 editions of his *History;* and an abridgment, *The Student's Hume* (1859; often reprinted), remained in common use for 50 years. Though now outdated, its generations of readers entitle the *History* to be regarded as an event of cultural importance. Moreover, in its own day it was an innovation, soaring high above its very few predecessors. It was fuller and set a higher standard of impartiality. It saw in the nation the mental interests of the educated citizens as well as the deeds of kings and statesmen, as may be seen, for instance, in the pages on literature and science at the end of chapter 3 under the Commonwealth and at the end of chapter 2 under James II. It was unprecedentedly readable, in structure as well as in phrasing, persons and events being woven into causal patterns that furnished a narrative with the goals and resting points of recurrent climaxes. That was to be the plan and march of future history books for the general reader. No one, however, would place Hume near to the English historian Edward Gibbon.

As an economist. Hume steps forward as an economist in the *Political Discourses* incorporated in *Essays and Treatises* as part 2 of *Essays Moral, Political and Literary*. How far he influenced his friend Adam Smith, 12 years his junior, remains uncertain: they have broadly similar principles, and both have the excellent habit of illustrating and supporting these from history. How far he was original is also uncertain: *The Querist* (1735–37) by the Irish Empiricist philosopher George Berkeley and some of the private correspondence of the period leave the impression that there was more economic sense alive than the histories of economic thought suggest. Still, in the eliciting of general principles, in sense of evidence, and in lucidity of exposition, Hume wrote ahead of his day, though, unlike Smith, whose *Wealth of Nations* he just lived to see, he did not work out a system. His level of insight can be gathered from his main contentions: that wealth consists not of money but of commodities; that the amount of money in circulation should be kept related to the amount of goods in the market (two points made by Berkeley); that a low rate of interest is a symptom not of superabundance of money but of booming trade; that no nation can go on exporting only for bullion; that each nation has special advantages of raw materials, climate, and skill, so that a free interchange of products (with some exceptions) is mutually beneficial; and that poor nations impoverish the rest just because they do not produce enough to be able to take much part in that exchange. He welcomed advance beyond an agricultural to an industrial economy as a precondition of any but the barer forms of civilization.

Main contentions in economics

As a philosopher. This is the essential Hume, for even outside his philosophy all his serious work is either the seed or the fruit of his conception of philosophy as the inductive science of human nature, as well as of his philosophical conclusion that man is more a creature of sensitive and practical sentiment than of reason. It is also the Hume that has been most vigorously alive since his physical death. On the continent of Europe he is seen as one of the few classical philosophers of Great Britain, with only Francis Bacon, an early-16th-century philosopher and essayist, and John Locke, a 17th-century Empiricist, for company. For some Germans his title to fame is that, by putting the problem of causal inference devastatingly, he stung the German thinker Immanuel Kant into creating the "critical" philosophy. He was one of the factors that led Auguste Comte, a 19th-century French mathematician and philosopher, to Positivism. In Great Britain he has been and remains the symbol of a cause, of a type of philosophy that must be either exalted or disgraced. The Scottish school of common-sense philosophy, from its beginning in his lifetime under Thomas Reid until Thomas Brown in the early 19th century, recoiled from his skeptical treatment of belief in causality and in a material world, and the British Hegelian Idealists in the latter part of the 19th century attacked his Empiricism and naturalism generally. His positive influence is instanced in Jeremy Bentham, an early-19th-century British jurist and philosopher, who was confessedly moved to Utilitarianism (the moral theory that right conduct should be determined by the usefulness of its consequences) by book III of the *Treatise*, and more extensively in John Stuart Mill, a British philosopher and economist who lived later in the 19th century. By the middle of the 20th century, he was again on his pedestal, regarded by the antimetaphysicians in England as one of the few philosophers of the past still worth reading.

MAJOR WORKS

PHILOSOPHY AND RELIGION: *A Treatise of Human Nature*, 3 books in 2 vol. (1739–40); *Philosophical Essays Concerning Human Understanding* (1748; from 1758 entitled *An Enquiry Concerning Human Understanding*), mainly a rewriting of book I of the *Treatise of Human Nature; Four Dissertations* (1757), including a rewriting of book II of the *Treatise; Dialogues Concerning Natural Religion* (1779).

POLITICS AND MORALS: *Essays, Moral and Political*, 2 vol. (1741–42); a further *Three Essays, Moral and Political* (1748); *An Enquiry Concerning the Principles of Morals* (1751), a rewriting of book III of the *Treatise; Political Discourses* (1752).

HISTORY: *The History of England*, 6 vol. (1754–62; vol. 1 reprinted 1970).

OTHER WORKS: *A Concise and Genuine Account of the Dispute Between Mr. Hume and Mr. Rousseau* (1766); *The Life of David Hume, Written by Himself* (1777).

BIBLIOGRAPHY. T.E. JESSOP, *A Bibliography of David Hume and of Scottish Philosophy from Francis Hutcheson to Lord Balfour* (1938, reprinted 1966).

Life and letters: Hume's own autobiography is scarcely a narrative of his life. The definitive biography is E.C. MOSSNER, *The Life of David Hume* (1954, reprinted 1970); a competent shorter one is J.Y.T. GREIG, *David Hume* (1931). *The Letters of David Hume*, ed. by GREIG, 2 vol. (1932); and *New Letters of David Hume*, ed. by R. KLIBANSKY and E.C. MOSSNER (1954), supersede older collections.

Texts: Hume's final text is in the 1777 edition of his own collection, *Essays and Treatises*, which has been often reprinted—most recently as *Essays Moral, Political and Literary* (1872, 1963). *The Philosophical Works of David Hume*, ed. by T.H. GREEN and T.H. GROSE, 4 vol. (1874–75), has also gone through several editions and reprints. A complete critical edition is yet to be produced. The most carefully edited text of the *Treatise* and of the two *Enquiries* is by L.A. SELBY-BIGGE (1888 and 1894, both still in print). An *Abstract* of the *Treatise* (1740, really an introduction to it) was found, identified as by Hume, and reprinted by J.M. KEYNES and P. SRAFFA (1938). The best edition of the *Dialogues Concerning Natural Religion* is by N. KEMP SMITH, 2nd ed. (1947, reprinted 1964). Collections of Hume's texts on special fields are: *Hume: Theory of Knowledge*, ed. by D.C. YALDEN-THOMSON, and *Hume: Theory of Politics*, ed. by F. WATKINS (both 1951); *Hume's Writings on Economics*, ed. by E. ROTWEIN (1955); *Hume on Religion*, ed. by R. WOLLHEIM (1963); and *David Hume, Philosophical Historian*, ed. by D.F. NORTON and R.H. POPKIN (1965).

On his philosophy: General introductions are D.G.C. MACNABB, *David Hume: His Theory of Knowledge and Morality*, 2nd ed. (1966); A.P. CAVENDISH (A.H. BASSON), *David Hume* (1958, reprinted 1968); and J.V. PRICE, *David Hume* (1968). In the considerable technical literature, outstanding books are: C.W. HENDEL, *Studies in the Philosophy of David Hume* (1925, reprinted with additions, 1963); CONSTANCE MAUND, *Hume's Theory of Knowledge* (1937); H.H. PRICE, *Hume's Theory of the External World* (1940); N. KEMP SMITH, *The Philosophy of David Hume* (1941, reprinted 1966); and J.A. PASSMORE, *Hume's Intentions*, rev. ed. (1968).

(T.E.Je.)